Near-term climate change: High-resolution decadal climate predictions in the Iberian Peninsula

Cambio climático a corto plazo: Predicciones climáticas decenales de alta resolución en la península ibérica

Memoria presentada para optar al título de Doctor en Física y Ciencias del Espacio por:

Juan José Rosa Cánovas

Dirigida por:

Dra. María Jesús Esteban Parra
Dra. Sonia Raquel Gámiz Fortis

Programa de Doctorado en Física y Ciencias del Espacio
Universidad de Granada

Editor: Universidad de Granada. Tesis Doctorales
Autor: Juan José Rosa Canovas
ISBN: 978-84-1195-757-1
URI: https://hdl.handle.net/10481/103217

UNIVERSIDAD
DE GRANADA

El doctorando / *the doctoral candidate* Juan José Rosa Cánovas y las directoras de la tesis / *and the thesis supervisors* María Jesús Esteban Parra y Sonia Raquel Gámiz Fortis:

Garantizamos, al firmar esta tesis doctoral, que el trabajo ha sido realizado por el doctorando bajo la dirección de las directoras de la tesis y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

Guarantee, by signing this doctoral thesis, that the work has been done by the doctoral candidate under the direction of the thesis supervisors and, as far as our knowledge reaches, in the performance of the work, the rights of other authors to be cited (when their results or publications have been used) have been respected.

En Granada, a 10 de diciembre de 2024:

Dra. María Jesús Esteban Parra
Directora

Dra. Sonia Raquel Gámiz Fortis
Directora

Juan José Rosa Cánovas
Doctorando

Agradecimientos

Si bien se me considera el autor de este trabajo a efectos administrativos, no deja de parecerme un poco injusto el intento de atribución de la autoría individual por mi parte. No es una forma de hablar. Tampoco quiero desdeñar el esfuerzo y el tiempo que le he dedicado durante estos cinco años. Es algo que pienso desde la más firme convicción de que no existe una manera de construir y trabajar el conocimiento que no parta desde lo estrictamente colectivo.

La tarea de reunir en estas líneas los nombres de todas las personas que han contribuido de un modo u otro a la elaboración de esta tesis doctoral me resulta inasumible. Sería necesario realizar un ejercicio exhaustivo de investigación solamente para reconocer la labor de quienes han participado en el desarrollo de las herramientas informáticas que han hecho posible el trabajo[1]. Sería aún mayor el número de páginas que habría que dedicar para mencionar individualmente a quienes han influido directa o indirectamente en mi formación, en el sentido más amplio de la palabra. Asumo con resignación el hecho inevitable de que este apartado sea mucho más corto de lo que debería.

En primer lugar, quiero agradecer a María Jesús, Sonia y Yolanda su guía, su crítica, siempre constructiva, y su paciencia. Ellas son responsables en gran medida de que este trabajo vea finalmente la luz.

Durante mi labor investigadora he disfrutado de un contrato predoctoral, con referencia PRE2018-083921, asociado al proyecto de I+D+i CGL2017-89836-R, financiado por MINECO/AEI/10.13039/501100011033/FEDER, seguido de un contrato ligado al proyecto PID2021-126401OB-I00, financiado por MCIN/AEI/10.13039/501100011033/ FEDER Una manera de hacer Europa.

Una parte de las simulaciones climáticas que constituyen el núcleo de este estudio fueron realizadas en los centros de supercomputación Tirant (Universidad de Valencia) y Picasso (Universidad de Málaga), ambos pertenecientes a la Red Española de Supercomputación. La otra parte de las simulaciones se llevó a cabo en Albaicín, el

[1] Hay una tentativa de hacerlo, a todas luces insuficiente, al final del Capítulo 3, dedicado a describir la metodología utilizada en el trabajo.

centro de supercomputación de la Universidad de Granada. Muchas gracias a los equipos técnicos de cada centro por ayudarme a solucionar los problemas que no fui capaz de resolver por mí mismo.

Muchas gracias a todas las personas del Centro Euro-Mediterraneo sui Cambiamenti Climatici que me acogieron durante mi estancia en Bolonia a finales de 2022. En especial a Panos Athanasiadis, Carmen Álvarez, Dario Nicolì, Loredana Amato y Silvio Gualdi. Me llevo una buena experiencia y un bonito recuerdo de la ciudad. Muchas gracias además a Alessandro por ayudarme a gestionar el alojamiento desde España.

Quiero agradecer también al Grupo de Modelización Atmosférica Regional de la Universidad de Murcia el haber facilitado mi entrada al mundo de la investigación. Muchas gracias especialmente a Marco por su generosidad, sus consejos y su ayuda.

Mi eterna gratitud a mi gente de la SF1, tanto a quienes están actualmente como a quienes pasaron por ahí. Gracias por hacer el camino más fácil y llevadero. La lista es larga: Mati, Guadalupe, César, Alá, Emilio, Patricio, Enrique (Pérez), Clément, Sergio, Marcela, Enrique (Echeverría), Félix, Luna, David y Nico. Y gracias a Mati otra vez, por enseñarme de qué va todo esto y echarme una mano siempre que lo he necesitado. Os espero en la granja, hay sitio suficiente para todo el mundo.

Gracias a Marcos, Juan y Pedro, porque pudo salir mal pero estamos bien.

Gracias a quienes convivieron conmigo en Granada. A Fran, Simón, Raúl y Alain. Gracias a Belchí, por la música que no cesa, y a Álvaro, que se marchó de imprevisto en mitad de la fiesta. Aquello que no se olvida permanece para siempre.

Por supuesto, muchas gracias a mi familia. Desde los más pequeños hasta los más viejos. Porque han estado, porque están y porque estarán.

Y por último, gracias a todas las personas y colectivos que, desinteresadamente y hasta las últimas consecuencias, hacen posible el libre acceso al conocimiento.

Omnia sunt communia.

Granada, diciembre de 2024.

Abstract

The time scales of climate assessments play a crucial role in the development of climate services, which aim at effectively transforming the results of the scientific research into solutions to real-world problems. While the seasonal-to-interannual climate predictions take advance of an accurate description of the initial climate state to carry out estimations of the actual climate evolution in the near future, the long-term climate projections represent the potential climate evolution as a response to anthropogenic external forcings, in scales ranging from several decades to centuries, along different future scenarios of climate change. The decadal climate predictions (DCPs), the subject of study in this Thesis, bridge the gap between seasonal-to-interannual predictions and climate projections. At the decadal time scale, both initial conditions and external factors jointly contribute to the estimation of the climate signal, making DCPs valuable sources of climate information for a wide range of users and decision-makers in the social, environmental and economic spheres.

The main purpose of this Thesis has been to generate a collection of high-resolution DCPs over the Iberian Peninsula (IP) and evaluate their accuracy and reliability along with their added value over a global decadal prediction system (DPS) and a set of high-resolution uninitialized experiments. To address this task, a set of dynamical downscaling (DD) simulations was conducted with the Weather Research and Forecasting model (WRF), with data from the global CESM Decadal Prediction Large Ensemble (CESM-DPLE) as input information for the decadal experiments, from the global CESM Large Ensemble (CESM-LE) for the uninitialized simulations and from the ERA-Interim reanalysis for other additional experiments. The DD simulations were carried out in two nested domains. A coarse-grid domain was defined to cover the EURO-CORDEX region with an horizontal resolution about 50 km, whereas a fine-grid domain, with an approximate resolution of 10 km, was centered in the IP. Approximately 4.94 million CPU hours were dedicated to conduct the DD simulations required to produce a total of 1470 simulated years. To the best of my knowledge, the research presented here constitutes the first study which comprehensively assesses the performance of a dynamically downscaled DPS at an horizontal resolution of 10 km, becoming the maximum resolution attained in this branch of the climate prediction.

In spite of the huge development achieved in climate modeling during the last three decades, models are intrinsically based on approximations and, consequently, contain biases which arise from different sources. Thus, the CESM-DPLE and CESM-LE datasets were bias corrected before conducting the DD simulations to reduce the potentially negative impact that those biases may have on the downscaled product. The simulations were conducted in a context of limited access to computing resources, so a representative subset of 4 members for each global ensemble was selected to provide the input information for the simulations, since the task of downscaling the whole ensembles was not addressable. For each member of the CESM-DPLE subset, the experiments initialized every year from 1970 to 1999 were downscaled, as well as the whole 10-member ensemble available for DD along the decade initialized in 2015. For CESM-LE, however, the simulated period is shorter because the data were only available from 1990 to 2005.

The evaluation of the downscaled product has been focused on four primary climate variables: precipitation and maximum, minimum and mean near-surface air temperature. Before the analysis, the dynamically downscaled decadal experiments were recalibrated to reduce the unconditional and conditional biases and adjusting the ensemble spread of the WRF output fields. Significant improvements over the global CESM-DPLE and the downscaled uninitialized experiments have been found at both annual and seasonal scales for temperature variables, whereas the added value of the downscaled DCPs to the predictive skill for precipitation is more limited. The signal-to-noise paradox is strong in the predictions for precipitation and, to a lesser extent, also for temperature. The results suggest that high improvements in the predictive skill may be achieved by adding new members to the downscaled ensemble to compute larger ensemble averages and thus reduce the unpredictable background noise in predictions, especially for precipitation.

The sensitivity of WRF simulations to extreme initial conditions of soil moisture has also been examined. A set of simulations was conducted with ERA-Interim providing the input information for all variables with the exception of soil moisture. Three different types of soil moisture initial conditions were considered to represent a wet, a dry and a very dry soil. These initial conditions were calculated by combining the soil moisture index with some physical soil properties which depend on the soil textures. To account for the impact of the initialization date of the simulations, they started in two different dates, 1990-01-01 and 1990-07-01, covering in both cases the

10-year periods up to 1999-12-31 and 2000-06-30, respectively. The analysis of these simulations has been focused on the influence of the initial conditions on the spin-up requirements for soil moisture, precipitation and the three temperature variables. A maximum spin-up time of 8 years is needed for soil moisture in some cases, decreasing down to values generally lower than 3 years and 2 years for maximum and mean temperature, respectively. The spin-up time is commonly lower than 1 year for minimum temperature and mostly below 10 months for precipitation.

Since no spin-up time has been considered in the analyses of the downscaled decadal experiments initialized from 1970 to 1999 (it would have implied the loss of the first simulated years), the predictive skill might have experienced some deterioration because of the spin-up-related biases, at least during the first years of the simulations. Therefore, a dynamically equilibrated soil state, taken from a control WRF simulation, was used to initialize the simulations conducted for the decadal experiments in the decade 2015–2025 with the aim of improving as much as possible the predictive skill of the downscaled predictions. In regions with reliable predictions for precipitation at annual scale, the predicted anomalies for this variable are generally positive at the beginning of the decade and turn into negative during the second half, with the Pyrenees and the Central System among the areas with the strongest negative anomalies. The predictions for the temperature variables show positive anomalies throughout the entire decade over the whole domain at annual scale. The highest anomalies have been found in summer, with values up to 2 K at the end of the decade in some southeastern and northeastern locations of the IP.

Finally, a set of alternative correction methods has been examined to improve the bias correction in CESM-DPLE experiments and thus produce a more skilful downscaled product in potential future experiments. The correction of the trend performs well and contributes to producing robust predictions for the North Atlantic Oscillation, becoming a suitable method to correct the input data for DD simulations focused on Europe or North America. On the other hand, a method based on considering reference initial conditions in the correction algorithm generally get overall sligthly better results than the other analyzed methods for the prediction of the El Niño/Southern Oscillation. This method may be preferable for DD simulations targeting South America. The dependence of the predictive skill on the ensemble size has also been analyzed with data corrected with these methods. A modest 3-member ensemble has shown to be a good alternative to larger ensembles in a context of limited computing resources for some specific applications.

The research presented in this Thesis evidences the valuable role that WRF can play for the generation of high-resolution decadal experiments over the IP, demonstrating the ability of produce skilful predictions despite the limitations imposed by the restricted access to computing resources. The multiple applications of DD in the branch of the DCP and their potential ramifications open a vast field of research which could be explored in future works by taking the study presented here as a solid starting point.

Resumen

Las escalas de tiempo de los estudios climáticos cumplen un papel crucial en el desarrollo de servicios climáticos, que tienen como objetivo transformar los resultados de las investigaciones científicas en soluciones para problemas del mundo real. Mientras que las predicciones climáticas estacionales–interanuales parten de una descripción precisa del estado inicial del clima para llevar a cabo estimaciones de su evolución real en un futuro próximo, las proyecciones climáticas a largo plazo representan la potencial evolución del clima como respuesta a forzamientos externos de carácter antropogénico, en escalas que abarcan desde varias décadas hasta siglos, a lo largo de diferentes escenarios futuros de cambio climático. Las predicciones climáticas decenales (DCPs), el tema de estudio en esta Tesis, hacen de puente entre las predicciones estacionales–interanuales y las proyecciones climáticas. En la escala decenal, tanto las condiciones iniciales como los factores externos contribuyen de manera conjunta a la estimación de la señal climática, convirtiendo a las DCPs en fuentes de información climática de valor tanto para la ciudadanía como para las autoridades en las esferas social, medioambiental y económica.

El propósito principal de esta Tesis ha sido generar una colección de DCPs de alta resolución en la península ibérica (IP) y evaluar su precisión y fiabilidad junto a su valor añadido sobre un sistema de predicción decenal (DPS) global y un conjunto de experimentos no inicializados de alta resolución. Para realizar esta tarea, se llevó a cabo una serie de simulaciones de reducción dinámica de escala (DD) con el modelo Weather Research and Forecasting (WRF), tomando datos globales del CESM Decadal Prediction Large Ensemble (CESM-DPLE) como información de entrada para los experimentos decenales, del CESM Large Ensemble (CESM-LE) para las simulaciones no inicializadas y del reanálisis ERA-Interim para otros experimentos adicionales. Las simulaciones DD se llevaron a cabo en dos dominios anidados. Un dominio grande fue definido para cubrir la región de EURO-CORDEX con una resolución aproximada de 50 km, mientras que un dominio más pequeño, con una mayor resolución en torno a los 10 km, se centró en la IP. Se dedicaron alrededor de 4.94 millones de horas CPU a la realización de las simulaciones DD para producir un total de 1470 años de simulación. Hasta donde alcanza mi conocimiento, la investigación aquí

presentada constituye el primer estudio que evalúa en profundidad un DPS sujeto a una reducción dinámica de escala con una resolución de 10 km, convirtiéndose en la mayor resolución espacial jamás lograda en esta rama de la predicción del clima.

Pese al enorme desarrollo logrado en la modelización de clima durante las últimas tres décadas, los modelos están intrínsecamente basados en aproximaciones y, en consecuencia, contienen sesgos que surgen de diferentes fuentes. Por tanto, los datos del CESM-DPLE y el CESM-LE fueron sometidos a una corrección de sesgo antes de realizar las simulaciones DD para reducir el impacto potencialmente negativo que estos sesgos pudieran tener en el producto final. Las simulaciones se realizaron en un contexto de acceso limitado a recursos computacionales, así que se seleccionó un subconjunto de 4 miembros representativo de cada conjunto global para proporcionar la información de entrada para las simulaciones, ya que la tarea de emplear el total de los miembros disponibles era inabordable. Para cada miembro del subconjunto del CESM-DPLE se aplicó la reducción dinámica de escala a los experimentos inicializados cada año desde 1970 hasta 1999, así como al conjunto completo de 10 miembros disponibles para las simulaciones DD a lo largo de la década inicializada en 2015. Para el CESM-LE, sin embargo, el periodo de simulación fue más corto, ya que solo había datos disponibles desde 1990 hasta 2005.

La evaluación del producto de la reducción de escala se ha centrado en cuatro variables climáticas primarias: la precipitación y las temperaturas máxima, mínima y media del aire en superficie. Antes del análisis, los experimentos decenales de alta resolución fueron recalibrados con el objetivo de reducir los sesgos incondicionales y condiciones y de ajustar la dispersión de los miembros del subconjunto procedente de las simulaciones con WRF. Se han encontrado mejoras significativas frente al CESM-DPLE global y a los experimentos no inicializados de alta resolución en las escalas anual y estacional para las variables de temperatura, mientras que el valor añadido de los nuevos experimentos decenales es más limitado en el caso de la precipitación. La presencia de la paradoja señal-ruido es fuerte en las predicciones para la precipitación y, en menor medida, también para la temperatura. Los resultados sugieren que se podrían alcanzar grandes mejoras en la habilidad predictora de los experimentos decenales de alta resolución añadiendo nuevos miembros para tomar promedios de conjunto mayores y así reducir el ruido de fondo impredecible en las prediciones, especialmente en el caso de la precipitación.

También se ha examinado la sensibilidad de las simulaciones realizadas con WRF a condiciones iniciales extremas de humedad del suelo. Se llevó a cabo una serie de

simulaciones con ERA-Interim proporcionando la información de entrada para todas las variables excepto para la humedad del suelo. En su caso, se consideraron tres tipos diferentes de condiciones iniciales para representar un suelo húmedo, uno seco y uno muy seco. Estas condiciones iniciales se obtuvieron de la combinación del índice de humedad del suelo y algunas propiedades físicas dependientes del tipo de textura del suelo. Para tener en cuenta el impacto de la fecha de inicialización de las simulaciones, éstas comenzaron en 1990-01-01 y en 1990-07-01, cubriendo en ambos casos un periodo de 10 años hasta 1999-12-31 y 2000-06-30, respectivamente. El análisis de estas simulaciones se ha centrado en la influencia de las condiciones iniciales en la duración del periodo de spin-up para la humedad del suelo, la precipitación y las tres variables de temperatura. Un tiempo máximo de spin-up de 8 años es necesario en algunos casos para la humedad del suelo, disminuyendo hasta valores generalmente menores a 3 y 2 años para las temperaturas máxima y media, respectivamente. El tiempo de spin-up es comúnmente menor a 1 año para la temperatura mínima y está mayoritariamente por debajo de los 10 meses para la precipitación.

Dado que no se ha considerado ningún tiempo de spin-up en los análisis de los experimentos decenales de alta resolución inicializados desde 1970 hasta 1999 (de haberlo hecho, habría implicado la pérdida de los primeros años simulados), la habilidad predictora puede haber sufrido cierto deterioro a causa de sesgos relacionados con el spin-up, al menos durante los primeros años de simulación. Por tanto, se utilizó un suelo dinámicamente equilibrado, tomado de una simulación de control con WRF, para inicializar las simulaciones realizadas para los experimentos de la década 2015–2025 con el objetivo de mejorar en la medida de lo posible la habilidad predictora de las predicciones de alta resolución. En regiones con predicciones de precipitación fiables en escala anual, las anomalías pronosticadas para esta variable son generalmente positivas al comienzo de la década y negativas durante la segunda mitad, con los Pirineos y el Sistema Central entre las áreas con mayores anomalías negativas. Las predicciones para las variables de temperatura en escala anual muestran anomalías positivas durante toda la década en todo el dominio. Las anomalías más elevadas se han encontrado en verano, con valores de hasta 2 K al final de la década en algunas localizaciones al sureste y noreste de la IP.

Finalmente, una serie de métodos de corrección alternativos han sido examinados para mejorar la reducción del sesgo en los experimentos del CESM-DPLE y así conseguir una mayor habilidad predictora en nuevos experimentos decenales de alta resolución que puedan ser realizados en un futuro. La corrección de las tendencias ha demostrado funcionar bien para lograr predicciones robustas de la Oscilación del Atlántico Norte, convirtiéndose en un método apropiado para ser aplicado sobre datos de entrada de simulaciones DD en Europa o Norteamérica. Por otro lado, un método basado en la utilización de condiciones iniciales de referencia en el algoritmo de corrección ha conseguido, en general, mejores resultados que otros métodos analizados para la predicción de El Niño/Oscilación del Sur. Este método podría ser preferible para simulaciones DD centradas en Sudamérica. El grado en el que la habilidad predictora depende del número de miembros del conjunto ha sido también analizado utilizando datos corregidos con estos dos métodos. Un modesto conjunto compuesto por 3 miembros ha demostrado ser una buena alternativa a conjuntos de tamaños superiores para algunas aplicaciones específicas en un contexto de recursos computacionales limitados.

La investigación presentada en esta Tesis evidencia el valor del papel que puede cumplir el modelo regional WRF en la generación de experimentos decenales a alta resolución en la IP, demostrando su capacidad para realizar predicciones precisas y fiables a pesar de las limitaciones impuestas por el acceso a recursos computacionales. Las múltiples aplicaciones del DD en el ámbito de la DCP y sus potenciales ramificaciones abren un amplio campo de investigación que podría ser explorado en trabajos futuros tomando este estudio como un sólido punto de partida.

Contents

List of figures

List of tables

Acronyms

ACC	anomaly correlation coefficient
ACM	Asymmetrical Convective Model
AEMET	State Meteorological Agency (Agencia Estatal de Meterología, in Spanish)
AMV	Atlantic Mutidecadal Variability
ARW	Advanced Research WRF
BMJ	Betts-Miller-Janjic
CAM	Community Atmosphere Model
CB	conditional bias
CBA	conditional bias adjustment
CDF	cumulative distribution function
CDO	Climate Data Operators
CESM	Community Earth System Model
CESM-DPLE	CESM Decadal Prediction Large Ensemble
CESM-LE	CESM Large Ensemble
CESM-WACCM	CESM Whole Atmosphere Community Climate Model
CICE	Community Ice Code
CLM	Community Land Model
CMIP{5, 6}	Coupled Model Intercomparison Project {Phase 5, Phase 6}
CORDEX	Coordinated Regional climate Downscaling Experiment
CRPS	continuous ranked probability score
CRPSS	continuous ranked probability skill score
DCP	decadal climate prediction
DCPP	Decadal Climate Prediction Project
DD	dynamical downscaling
DeFoReSt	Decadal Climate Forecast Recalibration Strategy
DJF	boreal winter (December, January and February)
DPS	decadal prediction system
ECMWF	European Centre for Medium-Range Weather Forecasts
ENS{4, 10, 40}	{4, 10, 40}-member CESM-DPLE ensemble
ENSO	El Niño/Southern Oscillation

EOF	empirical orthogonal function
ERSST	Extended Reconstructed Sea Surface Temperature
EURO-CORDEX	European Coordinated Regional Climate Downscaling Experiment
FIT	polynomial fitting
GCM	global climate model
GHG	greenhouse gas
GISTEMP	Goddard Institute for Space Studies Surface Temperature Analysis
GSAT	global-mean surface air temperature
HadSLP2r	Hadley Centre's near-real-time mean sea level pressure
HIRLAM	HIgh Resolution Limited Area Model
IC	initial condition
ICDC	initial condition-based drift correction
IFS	Integrated Forecast System
IP	Iberian Peninsula
JJA	boreal summer (June, July and August)
JRA-55	Japanese 55-year Reanalysis
kNN	k-nearest neighbours
LBC	lateral boundary condition
LESS	logarithmic ensemble spread score
LESSS	logarithmic ensemble spread skill score
LSM	land surface model
MAM	boreal spring (March, April and May)
MDC	mean drift correction
MERRA	Modern-Era Retrospective Analysis for Research and Applications
MiKlip	Medium-term Climate Forecasts (Mittelfristige Klimaprognosen, in German)
MM5	fifth-generation Penn State/NCAR Mesoscale Model
MPI-ESM	Max Planck Institute Earth System Model
MSE	mean squared error
MSSS	mean squared skill score
NAO	North Atlantic Oscillation
NCAR	National Center for Atmospheric Research
NCO	NetCDF Operators

NSAT	near-surface air temperature
PBL	planetary boundary layer
PC	principal component
$\mathbf{PC_R}$	rotated principal component
PCA	principal component analysis
PDF	probability distribution function
POP	Parallel Ocean Program
PR	precipitation
RCM	regional climate model
RCP	representative concentration pathway
RMSE	root mean square error
RPC	ratio of predictable components
SIP	seasonal-to-interannual prediction
SLP	sea level pressure
SMI	soil moisture index
SON	boreal autumn (September, November and December)
SPARC	Stratosphere-troposphere Processes And their Role in Climate
SSP	shared socioeconomic pathway
SST	sea surface temperature
$\mathbf{T_{\{max,\ min,\ mean\}}}$	daily {maximum, minimum, mean} NSAT
TNI	Trans-Niño Index
TrDC	trend-based drift correction
WPS	WRF Preprocessing System
WRCP	World Research Climate Programme
WRF	Weather Research and Forecasting model
WRF-DPLE	dynamically downscaled WRF ensemble with CESM-DPLE providing the input information in DD simulations
WRF-LE	dynamically downscaled WRF ensemble with CESM-LE providing the input information in DD simulations
WSM3	WRF Single-Moment 3-class

1

Introduction

1.1. Understanding climate prediction: concepts and tools

1.1.1. *Weather, climate and the human footprint*

Since Aristotle's Meteorologica (Aristotle, 1952) was written around 340 B.C., becoming the first compilation of studies on meteorology ever published, our knowledge on the atmosphere and processes occurring within it has incessantly evolved over centuries of scientific development. The Aristotle's inaccurate, although ingenious explanations of atmospheric phenomena have been substituted by a comprehensive, theoretical-practical, scientific discipline which heavily relies on state-of-the-art technology. In atmospheric sciences, meteorology is the branch which studies *weather*, i.e., the instantaneous state of the atmosphere at a given time and place. By contrast, climatology studies *climate*, which refers to weather conditions, averaged over time, at a place (Rohli and Vega, 2018). While weather consists of very short-term variations in the atmosphere which range from minutes to days, the evolution of climate from months to thousands or millions of years is the result of the interplay between three factors (IPCC, 2021b):

a) the natural internal variability of the atmosphere and its interaction with the other components of the climate system (the hydrosphere, the lithosphere, the cryosphere and the biosphere);
b) the natural external forcing[2] mechanisms, such as solar radiation variability, volcanic eruptions or changes in the Earth's orbit;
c) the anthropogenic external forcing[2] mechanisms, such as changes in the concentration of greenhouse gases (GHGs) and aerosols or in land use.

[2] *Radiative forcing* is the change in the net, downward minus upward, radiative flux (expressed in watts per squared metre) due to a change in an external driver of climate change (IPCC, 2021b).

The anthropogenic factor is unequivocally the major responsible of the warming observed in atmosphere, land and ocean from the last century to present. During this time, the climate system has experienced certain changes, such as the increase of the global-mean surface air temperature (GSAT) by 1.07 °C in the period 2011–2020 relative to 1850–1900, which cannot be explained only by natural causes; on the contrary, they are mainly consequences of the emission of GHGs from human activity (Eyring et al., 2021). The human-induced climate change is affecting people, ecosystems and infrastructures through the observed increment in the frequency and intensity of climate and weather extremes, such as extreme temperatures, drought or heavy precipitation events, and it is expected to continue in the future, even increasing its associated risks if no adaptation and mitigation actions are implemented (IPCC, 2022). The evolution of GSAT from 1950 to 2100 in observationally constrained historical simulations and future projections along several *shared socioeconomic pathways*[3] (SSPs; O'Neill et al., 2016) is depicted in Figure 1.1.

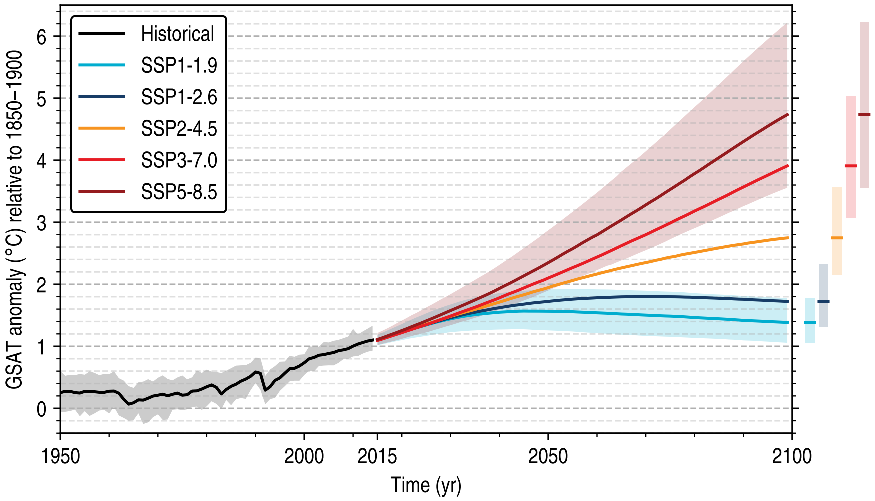

Figure 1.1: Global-mean surface air temperature (GSAT) anomaly (°C) relative to 1850–1900 for observationally constrained historical simulations (black) and future projections (coloured) along several shared socioeconomic pathways (SSPs; O'Neill et al., 2016). Solid lines denote 20-year moving averages, whereas shaded areas show 95%-level confidence intervals for historical, SSP1-1.9 and SSP5-8.5. Confidence intervals at the end of the 21th century for all SSPs are depicted on the right margin. Data provided by Fyfe et al. (2021).

[3] *Shared socioeconomic pathways* (SSPs) are scenarios of future emissions and land use changes integrated in the model simulations of the sixth phase of the Coupled Model Intercomparison Project (CMIP6). The paths followed by these scenarios were established with assumptions about how key climate drivers (demography, economics, technology, governance, etc.) will evolve in future. More information in O'Neill et al. (2016) and Chen et al. (2021).

1.1.2. *Near-term climate prediction: the decadal scale*

The time scales of climate assessments play a fundamental role in the development of climate services, which aim at effectively transferring the climate information from academy to users and decision-makers and transforming the research outcomes into solutions to real-world problems (Goddard, 2016). It is well known that the accuracy of weather prediction rapidly decays from approximately seven days onwards because of the chaotic nature of the atmosphere (Ahrens, 2009). However, when the target is climate, the prediction horizon can be extended in time (Meehl et al., 2021).

The term *climate prediction* or *climate forecast* is used to refer to an estimation of the actual evolution of the climate in the future (IPCC, 2021b). Climate predictions are sensitive to initial conditions, that is, the climate state which acts as the starting point in the forecast. The accuracy in representing that initial state directly influences on the skill of the prediction, so they are considered as *initial-condition* problems (Figure 1.2; Meehl et al., 2009). This sensitivity is higher in seasonal predictions and progressively decreases across the interannual and decadal scales (Meehl et al., 2021). The chaotic behaviour of natural variability makes climate predictions probabilistic in essence (Meehl et al., 2014). On the other hand, the term *climate projection* denotes an estimation of the climate evolution as a response to anthropogenic external

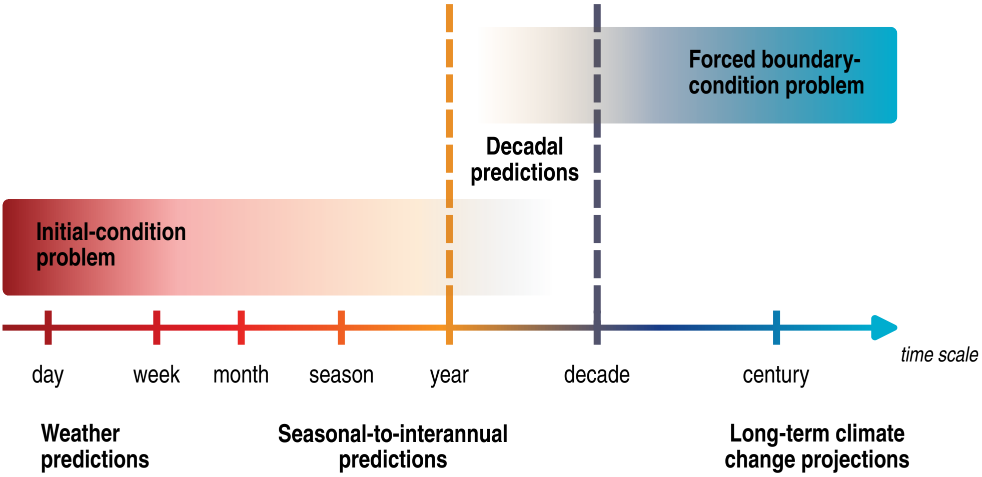
Forced boundary-condition problem
Decadal predictions
Initial-condition problem
day
week
month
season
year
decade
century
time scale
Weather predictions
Seasonal-to-interannual predictions
Long-term climate change projections

Figure 1.2: Schematic progression from initial-condition problems, with weather prediction on one end, to boundary-condition problems, with long-term climate change projections at the other end. Seasonal-to-interannual and decadal predictions are placed in between. Decadal predictions are considered both initial- and boundary-condition problems. Adapted with permission from Kirtman et al. (2013, Box 11.1, Figure 2).

forcings along a future scenario of climate change. Climate projections encompass time scales from several decades to centuries (Lee et al., 2021). At these time scales, the estimations fundamentally depend on external information and not on initial conditions (Hawkins and Sutton, 2009), so they are interpreted as *boundary-condition* problems (Figure 1.2; Meehl et al., 2009). Predictions are often referred to as *initialized* experiments because they start from an observed climate state, as opposed to *uninitialized* experiments, such as historical simulations and future projections (Meehl et al., 2021).

The *decadal climate predictions* (DCPs), the subject of study in this Thesis, bridge the gap between the seasonal-to-interannual predictions (SIPs) and long-term climate change projections, where both initial and boundary conditions can contribute to the extraction of the climate signal (Meehl et al., 2009, 2014, 2021). The DCPs have progressively been receiving much more attention by the scientific community dedicated to climate studies during the last decade, a consequence of the need to provide users and stakeholders with climate information at this time scale (Brasseur and Gallardo, 2016; Goddard, 2016; Graham et al., 2011). One of the first worldwide-coordinated research programs which included DCPs in its investigations was the fifth phase of the Coupled Model Intercomparison Project (CMIP5; Kirtman et al., 2013; Taylor et al., 2012). A few years later, the Decadal Climate Prediction Project (DCPP; Boer et al., 2016) established the experimental protocol for DCPs which was followed in the sixth phase of CMIP (CMIP6; Lee et al., 2021), built upon the knowledge and experiences gained from the previous phase.

The main tools used to study and understand the Earth's climate system are *climate models*. Climate models are computational models which, varying in their specific scope and complexity, help us to understand the physical, chemical and biological processes and interactions between the different elements which compose the climate system (Kotamarthi et al., 2021). Climate modeling has experienced an important development in the last three decades in terms of model accuracy, reliability and complexity (Randall et al., 2019). Nevertheless, climate models are intrinsically based on approximations and, therefore, contain biases which emerge from different sources, such as the model spatial and temporal resolution, the parametrization schemes or uncertainties in initial and boundary data, among others (Chen et al., 2021). In the field of DCP, very complex climate models which numerically represent the coupled system atmosphere-ocean-sea ice are among the most powerful tools used to conduct the experiments. These models are commonly referred to as *atmosphere-ocean general circulation models*, or *earth system models* if they also incorporate the representation of various biochemical cycles and/or vegetation (Kirtman et al., 2013; Lee et al., 2021). For the sake of simplicity, this sort of models will be generically named as *global climate models* (GCMs) hereinafter. GCMs divide the atmosphere, the ocean, the cryosphere and the land surface into grid cells within the equations which describe the processes governing the evolving climate are resolved. *Parametrizations* are used to represent those processes which occur on scales too small to be directly resolved by the model or which are computationally too expensive (McFarlane, 2011).

In DCPs, the initialization is focused on the GCM components which contain the largest sources of decadal climate predictability (Meehl et al., 2014). These components are characterized by having a large *climate memory* or *persistence*, meaning that the climate state at a given time influences on the physical mechanisms which determine the states in future (Yuan et al., 2013). The initialization of these components allows the production of climate predictions for time scales beyond months and years (Meehl et al., 2021). The ocean is known to be one of the most important drivers of decadal climate variability. Its large dynamical and thermal inertia is a consequence of the existence of low-frequency variability mechanisms which govern its evolution and operate worldwide (Cassou et al., 2018). Although the ocean is the main reservoir of memory in the climate system, it is not the only source of predictability. Sea ice or soil moisture, among other variables, are represented by processes which could potentially contribute to enhancing the predictability in DCPs (Bellucci et al., 2015). There are two main strategies to initialize the experiments: the *full-field* and *anomaly* initialization approaches (Meehl et al., 2009, 2014). The full-field approach (e.g., Yeager et al., 2012, 2018) consists of using observational data to derive the initial conditions for the model simulation. As the simulation evolves, the modeled climate drifts away from the observed climate towards the state systematically preferred by the model, the model climatology, which depends on its configuration (Meehl et al., 2009, 2014). The anomaly approach (Matei et al., 2012) may be used as an attempt to avoid this drift by only considering observed anomalies added to the model climatology as initial conditions. Afterwards, the model climatology is subtracted from the output to obtain the predicted anomalies. However, this method does not account for the inconsistencies which could exist between the model climatology and the observed anomalies (Meehl et al., 2014) or because of defining a model climatology under the influence of forced climate trends (Yeager et al., 2012), leading to biases in the forecasts again. Since none of these two strategies performs consistently better than the other in terms of predictive skill (Hazeleger et al., 2013), both are considered in the DCPP contribution to CMIP6 (Boer et al., 2016). In both cases, a correction of the model drift is firmly recommended by Boer et al. (2016) to reduce the lead time-dependent bias prior to carrying out any analysis.

The level of usefulness of climate predictions is linked to the amplitude of climate signal which can be predicted versus the amplitude of unpredictable background noise related to the chaotic nature of weather and climate, the uncertainties in initial conditions and the model formulation. This relationship is known as the signal-to-noise ratio (Scaife and Smith, 2018). Likewise systematic model errors can be reduced to some extent through drift correction, an ensemble of experiments generated by perturbing initial conditions is used to deal with the sensitivity of predictions to uncertainties in observational initial state (Meehl et al., 2014). A considerable amount of ensemble members may be needed to reveal the predictable signal masked by the unpredictable noise (Smith et al., 2019, 2020). By using a conceptual model to analyze the effect of the ensemble size on the predictive skill, Sienz et al. (2016) concluded that an ensemble of at least 10 members is required to conduct a robust assessment of the skill. This is also endorsed by Boer et al. (2016) under the CMIP6 protocol, who recommend considering an ensemble of 10 members (even more if possible) for every experiment initialized every year from 1960 onwards. However, there is a clear trade-off in incrementing the ensemble size: the increase of the computing resources needed by the simulations. This trade-off may not be worthwhile in some circumstances where the gain in predictive skill is low compared to the increase of the required computing time (Rosa-Cánovas et al., 2023).

1.1.3. *Regional climate models and dynamical downscaling*

Although the mechanisms of climate change operate worldwide, their effects are not experienced in the same way, for example, in a region characterized by having dry and hot summers and placed in the Mediterranean basin, and another one in south-eastern Asia, better described by a tropical rainforest climate and not having dry seasons (Beck et al., 2023). At local scale, the consequences of climate change are strongly region-dependent. The use of the so-called *regional climate models* (RCMs) is very well-established to describe the evolution of climate change and assess its impacts at a finer scale than that addressed by GCMs (Doblas-Reyes et al., 2021). RCMs are dynamical models very similar to their global counterparts in their fundamental principles. They numerically resolve the equations which describe the processes occurring in the atmosphere, land surface, ocean and ice, as GCMs do, but they are only run over a limited area. Since RCMs cover a smaller domain, the simulations conducted with them are less computationally expensive than those conducted with

GCMs, so RCMs can be used to obtain climate information at a higher resolution over a specific region of interest without a large trade-off in terms of computing time (Doblas-Reyes et al., 2021; Kotamarthi et al., 2021).

The technique used to produce finer-scale climate features with a RCM is commonly referred to as *dynamical downscaling* (DD). While GCMs can describe the global circulation, which is influenced by large-scale forcings such as changes in GHGs concentrations or solar radiation, RCMs enhance this large-scale information by taking into account the effects of forcings and processes at lower scale, such as complex topography, coastlines, land cover distribution or dynamical processes occurring at a GCM subgrid scale (Giorgi, 2019). A schematic representation of the DD approach is depicted in Figure 1.3. In DD simulations, RCMs take the large-scale information from a GCM (or a reanalysis of observations) through the meteorological fields provided as initial and time-dependent lateral boundary conditions (ICs and LBCs, respectively). In this sense, the domain of a RCM can be interpreted as a domain *nested* within the coarser domain of a GCM (or reanalysis), which subsequently transfers information about the state of the climate system in the surroundings of RCM through its boundaries (Giorgi, 2019). The LBCs are supplied for the *prognostic* model fields, i.e., those fields which are directly calculated by the model. This set of variables typically includes temperature, the components of wind velocity, hu-

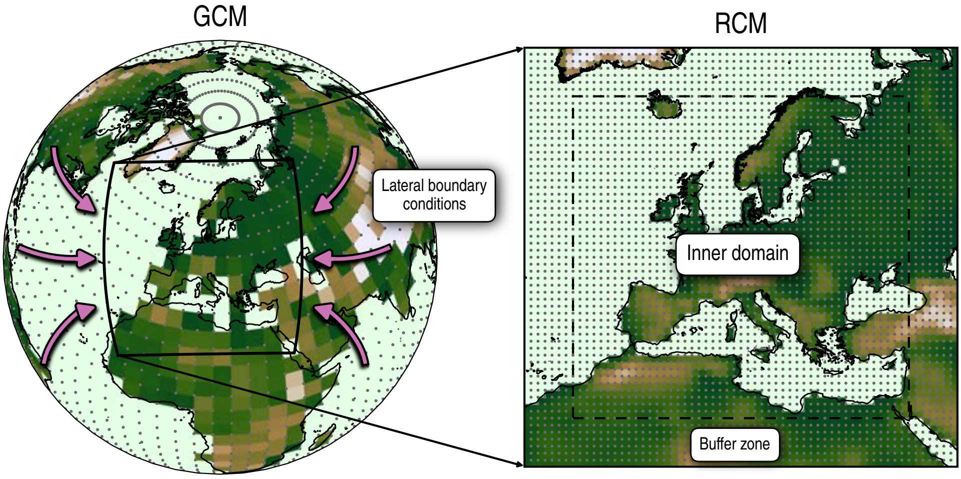

Figure 1.3: Schematic description of the DD technique. In this example, the GCM runs with a resolution of 5° (~560 km at equator), whereas the RCM does with a resolution of 0.88° (~100 km at equator). The RCM is provided with large scale information by the GCM through LBCs. The dashed line denotes the end of the buffer zone. Grey dots show the positions of the model cell corners.

midity, pressure or geopotential. For ICs, more fields may be required depending on the RCM used, such as soil temperature, soil moisture, surface skin temperature or snow-related variables. RCMs also need sea surface temperature (SST) as lower boundary condition over ocean areas, which is recommended to be time-dependent for simulations spanning from a week onwards (e.g., Giorgi et al., 2023; Skamarock et al., 2008). If the climate information follows an unidirectional path from GCM to RCM, i.e., there are no feedbacks between the models which influence on the GCM climate, an *one-way nesting* is being applied (e.g., Beck et al., 2004; Denis et al., 2002). In contrast, if these feedbacks are allowed, the simulation runs with a *two-way nesting* (e.g., Beck et al., 2004; Lorenz and Jacob, 2005).

The difference between GCM and RCM grid resolutions leads to mismatches between the nesting data and RCM results at the domain boundary. This issue is commonly addressed by defining a *lateral boundary* or *buffer zone* in the vicinities of the boundary, where the RCM results are continuously pushed towards the LBCs by a relaxation technique (Marbaix et al., 2003). In addition, the RCM results can also be relaxed to the large-scale fields not only in the boundaries, but also across the whole domain through a *spectral nudging* (Storch et al., 2000). This relaxation is applied in the spectral space, focusing on the large-scale components and leaving the small-scale components of the RCM fields unaffected, which are calculated by the RCM following the standard relaxation approach. The spectral nudging might prevent the inconsistencies between the small-scale processes generated by RCM and the large-scale circulation from the nesting data. Although studies show that the use of this technique may be beneficial —for example, eliminating the distortion of large-scale circulation by the lateral boundaries and the sensitivity of RCM to the domain geometry and location (e.g. Miguez-Macho et al., 2004; Radu et al., 2008)— the generation of mesoscale processes by RCM might be hindered if spectral nudging is too strong (Omrani et al., 2012).

Regardless of the relaxation technique and the nesting approach, the LBCs interact with the RCM physics and dynamics to generate the climatology in DD simulations. This climatology is the result of the dynamical equilibrium between both LBCs and RCM, which is reached after a certain amount of time since the start of the simulation. This time is known as the *spin-up time* (Giorgi, 2019; Giorgi and Mearns, 1999). At the beginning of the simulation, when the information provided by the LBCs start to spread throughout the domain, the bias the of RCM tends to change and eventually oscillate around an asymptotic value. When the bias achieves this asymptotic stage, the RCM reaches the dynamical equilibrium and is able to simulate the climatology with a relatively constant skill level. The analysis of the RCM simulations needs to account for the spin-up time to adequately evaluate their performances. While the spin-up time is from around several days to weeks for atmospheric variables such as temperature or precipitation, it may range from months to even several years for soil variables such as soil moisture (Giorgi and Mearns, 1999).

The regional climate modeling plays a very important role in vulnerability, impact and adaptation studies. The flexibility of RCMs provides groups or individuals with tools to investigate a wide variety of situations at regional scale with different model configurations and approaches. However, this versatility may be also a double-edge sword, since the lack of common protocols in the use of RCMs could difficult the transfer of knowledge between different projects. With this background, the Coordinated Regional climate Downscaling Experiment (CORDEX) was born in 2009, backed by the World Research Climate Programme (WRCP), as the first international program establishing a common protocol for downscaling experiments (Giorgi and Gutowski, 2015; Giorgi et al., 2009). The main goals of CORDEX, listed in Giorgi and Gutowski (2015), include improving the understanding the relevant climate processes occurring at regional and local scale, evaluating and improving the downscaling models and techniques, coordinating the production of downscaled sets of climate projections worldwide and promoting the exchange of knowledge with the end users of regional climate information.

The experiments done under the CORDEX framework just encompass long-term climate change simulations, but do not cover near-term climate predictions, such as SIPs or DCPs. Only the German research program Medium-term Climate Forecasts (MiKlip, in its German acronym; Marotzke et al., 2016) has carried out a comprehensive evaluation of the performance of DCPs at regional scale through a DD approach (e.g. Feldmann et al., 2019; Reyers et al., 2019). In MiKlip, the DD simulations run on the EURO-CORDEX domain with a resolution of 0.22° ($\sim$ 25 km) and use the Max Planck Institute Earth System Model (MPI-ESM; Stevens et al., 2013) as the GCM providing the ICs and LBCs. At the time of writing this Thesis, there are no similar assessments in published works which have not been carried out in the framework of MiKlip. To the best of my knowledge, the only study in literature not belonging to MiKlip which involves DD and DCPs can be found in Strobach and Bel (2019). However, this study is limited to use only one 30-year decadal experiment to both evaluate the predictive skill and forecast the future climate with a RCM making use of eight different parametrization schemes.

1.2. The Iberian Peninsula

The region of study in this Thesis is the Iberian Peninsula (IP; Figure 1.4), which encompasses the continental regions of Spain and Portugal. To be precise, most part of the analyses has been focused on the peninsular territory of Spain, excluding Portugal, because the observational datasets used in the evaluation of the DD experiments only contain information for Spain (see Section 2.3 for further details). The Balearic Islands, which belong to Spain and are located in the Mediterranean Sea, have also been included in the analyses because of their proximity to the IP. The peninsular territory of Portugal has been taken into account in the study carried out in Chapter 6, as reference data are available for this region in that specific case. For the sake of brevity, the term IP is used to refer to this geographical area hereinafter, regardless of whether Portugal is included in the analysis or not.

The IP, contained in the longitudes 10° W – 5° E and latitudes 36° N – 44° N, is placed in the southwesternmost part of Eurasia (Figure 1.4a). It is situated between the Atlantic Ocean and the Mediterranean Sea, bordering France to the north and separated from Africa just by the Strait of Gibraltar. The main topographic feature of the IP is its high terrain elevation (Instituto Geográfico Nacional, 2019), due to the existence of a large interior plateau: the Central Plateau. The Central Plateau is divided into two subplateaus (Northern and Southern Subplateaus) by the Central System and is surrounded by the Cantabrian Range, the Iberian System, Sierra Morena and the Baetic System. The average elevation in the IP is around 650 m, whereas the most frequent heights are between 700 and 800 m. The lowest elevation, below 200 m, is mainly located in narrow regions situated in the Mediterranean coast and wider areas in the south-western Atlantic coast. On the other hand, the highest elevation is located in Mulhacén (3,479 m, Baetic System), closely followed by Pico Aneto (3,404 m, the Pyrenees).

The climate of the IP almost entirely fits into the arid and temperate categories of the Köppen–Geiger climate classification (Figure 1.4b, Appendix A.1; Beck et al., 2023). The regions in the west, southwest and south mainly show a temperate climate characterized by hot and dry summers (Csa), with lower temperatures in the northwest (Csb). On the other hand, the central and eastern regions have a cold arid steppe climate in general, defined by low mean annual temperatures and precipitation (BSk). Some locations close to the Mediterranean coast have a hot arid steppe climate (BSh), where temperatures are higher than for the previous cold variant. In the southeast, there are some regions with an arid desert climate (BWh and BWk), where

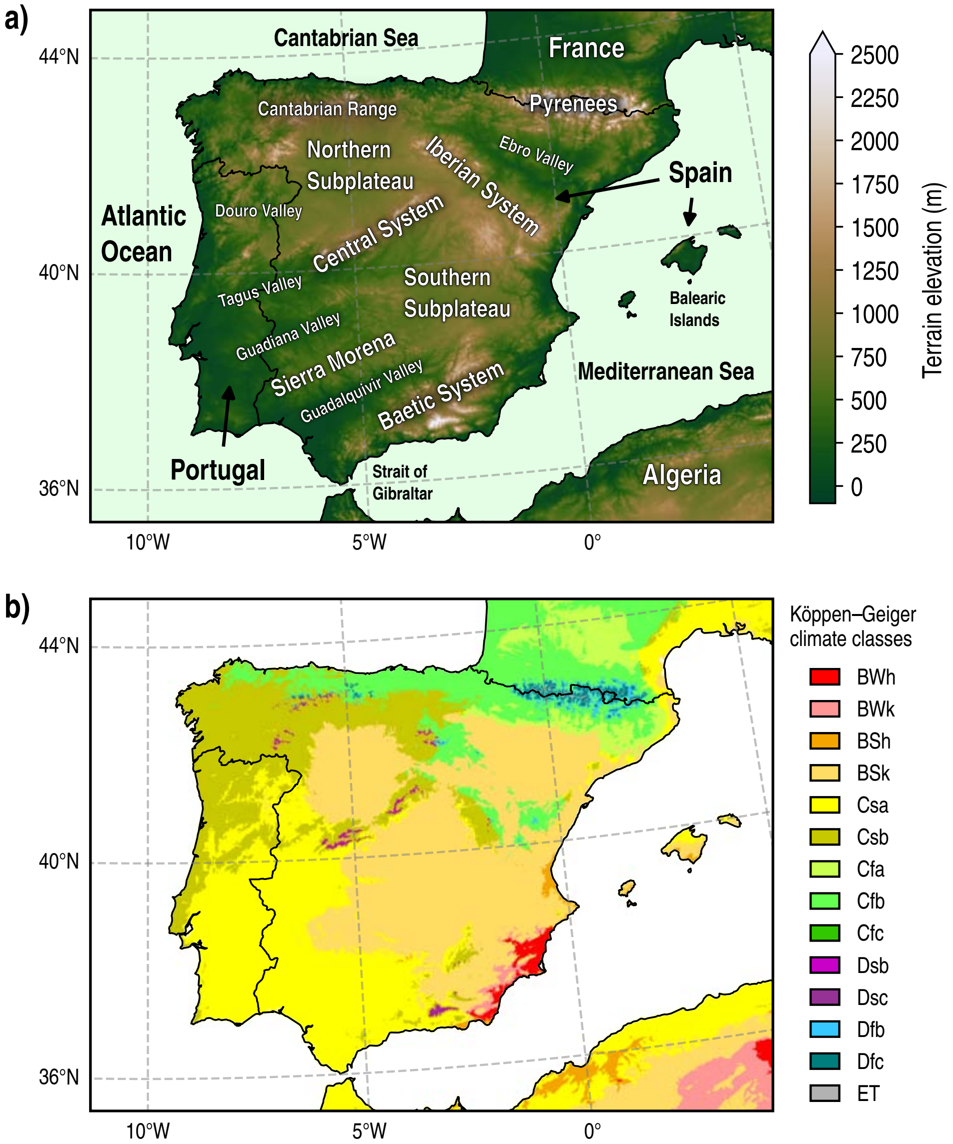

Figure 1.4: **a)** Schematic representation of the main geographical features in the Iberian Peninsula (IP). The terrain elevation data have been retrieved from Earth Resources Observation And Science Center (2017). **b)** Köppen–Geiger climate classification of IP. The full classification is composed of 30 climate classes, but only those corresponding to IP (and surroundings) are shown in the legend. Data retrieved from Beck et al. (2023).

mean annual precipitation is lower than for the steppe class. The northern regions are characterized by a temperate climate with no dry seasons and warm/hot summers (Cfa and Cfb). With respect to the high mountain regions, the Pyrenees and some locations in the Cantabrian Range are characterized by a cold climate with no dry seasons (Dfb and Dfc). In the Central or Baetic Systems, climate is also cold, but with dry summers (Dsb and Dsc). The highest altitudes present a polar climate with very low temperatures (ET).

1.3. Objectives and structure of the Thesis

The analyses and subsequent conclusions presented here are the result of the evaluation of a collection of experiments generated by approximately 4.94 million CPU hours of DD simulations, which produced a total of 1470 simulated years. At the moment of writing this Thesis, it represents the first and the only study which comprehensively assesses the predictive skill of a dynamically downscaled decadal prediction system (DPS) at a 10 km resolution, becoming the highest spatial resolution attained with a DD approach in this branch of the climate prediction.

The main objectives of this study are:

1) To generate a collection of high-resolution DCPs over the IP. This collection is the product of a set of DD simulations conducted with the Weather Research and Forecasting model (WRF) and provided with ICs and LBCs by a global DPS.
2) To evaluate the predictive skill for some of the most relevant downscaled climate variables in terms of their accuracy and reliability in reproducing the observed climate in the period 1970–2009, as well as the added value over the global initialized and downscaled uninitialized counterparts in the IP.
3) To examine the dynamically downscaled DCPs generated for the period 2015–2025 and evaluate their predictive skill up to 2020 in the IP.

Following this introduction, Chapter 2 presents the datasets used in this Thesis. Afterwards, Chapter 3 contains a description of the RCM and the methodologies followed to conduct the simulations and analyze the output product. Chapter 4 is devoted to evaluate the predictive skill of the downscaled DCPs for precipitation, as well as the added value over global DCPs and uninitialized downscaled experiments. In the same line, the results for daily maximum, minimum and mean near-surface air temperatures are evaluated in Chapter 5. The impact of the soil initialization with extreme soil moisture conditions on WRF simulations and its potential influence on the decadal predictive skill are examined in Chapter 6. The analysis of the downscaled DCPs for the decade 2015–2025 is done in Chapter 7. Additionally, an exploration of alternative drift correction techniques for global DCPs is made in Chapter 8. Finally, Chapter 9 contains the main conclusions of this Thesis and suggests some potential future works to continue the research presented here.

2

Data

This Chapter is devoted to describe the datasets used in this Thesis. Firstly, the models which have been used to provide the ICs and LBCs in DD simulations have been introduced. Secondly, the reanalysis products considered in drift correction, DD simulations and the evaluation of DCPs have been presented. Finally, the observational datasets also used in the analysis of the DCP experiments have been described.

2.1. Global climate model datasets

2.1.1. *CESM Decadal Prediction Large Ensemble*

The initial and boundary information required to conduct the decadal DD simulations has been provided by the collection of decadal experiments carried out with the Community Earth System Model (CESM), version 1.1, by the United States National Center for Atmospheric Research (NCAR) in the framework of the CESM Decadal Prediction Large Ensemble (CESM-DPLE; Yeager et al., 2018). The CESM-DPLE is one of the multiple DPSs which participate in the DCPP contribution to CMIP6 (Boer et al., 2016; IPCC, 2021a). It encompasses a set of simulations full-field initialized (see Section 1.1) every year on November 1st with start dates ranging from 1954 to 2015 (62 start dates). For each start date, an ensemble composed of 40 members was generated by randomly perturbing the initial atmospheric conditions at the round-off level.

CESM is a coupled model which assembles several submodels dedicated to solve the physics equations which govern the evolution of each component of the climate system (Yeager et al., 2018). The atmosphere component is the version 5 of the Community Atmosphere Model (CAM5; Hurrell et al., 2013), which runs with a finite-volume dynamical core at 1° resolution and 30 vertical levels. CESM uses the version 2 of the Parallel Ocean Program (POP; Danabasoglu et al., 2012) for the ocean component, with an horizontal resolution of 1° and 60 vertical levels. For the sea ice component, the version 4 of the Community Ice Code (CICE; Hunke and Lipscomb, 2008) runs at the same horizontal resolution of POP. Finally, the version 4 of the Community Land Model (CLM; Lawrence et al., 2011) is used as the land component. The observational information in the start dates is introduced through the ocean and ice components, whereas the uninitialized atmosphere and land components take their ICs from the restart files of a single member of the CESM Large Ensemble (CESM-LE; Kay et al., 2015) on November 1st (see Section 2.1.2). A more detailed description of the CESM-DPLE experimental design can be found in Yeager et al. (2018).

For the radiative forcings (which include well-mixed greenhouse gases, short-lived gases and aerosols), CESM-DPLE considers historical information until 2005 (Lamarque et al., 2010) and the representative concentration pathway (RCP) 8.5 used in CMIP5 from 2006 onwards (Lamarque et al., 2011; Meinshausen et al., 2011). In the case of ozone concentrations, a coupled chemistry-climate model, the CESM Whole Atmosphere Community Climate Model (CESM-WACCM; Marsh et al., 2013), is used to provide the information instead of CMIP5 CESM (Kay et al., 2015; Yeager et al., 2018). While CMIP5 CESM ozone forcing is known to underestimate ozone depletion at stratospheric level under Antarctica, CESM-WACCM provides a more consistent and realistic representation of the ozone hole (Eyring et al., 2013).

CESM-DPLE was chosen to provide the initial and boundary information in DD simulations since it is the only DPS which, at the time of writing this dissertation, publicly supplies all the mandatory fields needed by WRF. WRF requires information for multiple variables at several height and soil levels with a 6-hourly time aggregation (see Section 3.1 for further information). The huge computing cost of the simulations used to produce the DCPs and the large storage requirements needed to save all mandatory fields to drive a RCM directly impact on the availability of these variables. Not even CESM-DPLE provides all these fields for the whole 40-member ensemble, but only for 10 members from the experiments yearly initialized from 1954 to 1999, along with 2014 and 2015 (48 start dates). A list of the fields available to download can consulted in "Decadal Prediction Large Ensemble Project output fields list" (n.d.), whereas the data can be downloaded from "CESM1-CAM5-DP" (n.d.).

The dynamically downscaled decadal experiments conducted in the context of this Thesis cover two periods: the control period and the decade 2015–2025. The decadal experiments in the control period, also named retrospective DCPs or *hindcasts*, have been used to assess the predictive skill of the downscaled product and evaluate its reliability. The computing cost of the production of DCPs does not only affect simulations at global scale, but also at regional scale when a DD approach is used. Thus, the number of downscaled decadal experiments conducted here is constrained by the availability of computing resources. Among the decadal experiments available for DD, the hindcasts initialized every year from 1970 to 1999 (30 start dates) for 4 members of the CESM-DPLE have been selected to conduct the simulations (see Section 3.3.2 for more information about the member selection). Hereinafter, this dynamically downscaled WRF ensemble will be referred to as WRF-DPLE.

In the DCPP contribution to CMIP6 (Boer et al., 2016), the experimental design required by the Tier 2 phase of the production of decadal hindcasts consisted in decadal experiments initialized every year from 1960 to present for an ensemble size of at least 10 members (the Tier 1 phase required the same start dates but for 5-year experiments) to guarantee robust estimates of the predictive skill. The reduction of the number of start dates or the *sample size* in the downscaled experiments may affect to some extent the predictive skill assessment, since the sample size has been shown to influence on the magnitude and significance of the skill scores used to evaluate the performance of the DCPs, improving the consistency of the results with the increase of the number of start dates (Reyers et al., 2019; Sienz et al., 2016). On the other hand, decreasing the number of ensemble members may also negatively affect the predictive skill through the reduction of the signal-to-noise ratio, since the extraction of the climate signal benefits from taking larger ensemble averages to remove the unpredictable background noise present in decadal experiments (Reyers et al., 2019; Scaife and Smith, 2018; Sienz et al., 2016; Smith et al., 2019), as mentioned in Section 1.1.2. In an effort to improve as much as possible the robustness of the predictions for the decade 2015–2025, the 10 available members have been considered to conduct the DD simulations for this period. All the decadal experiments used as ICs and LBCs have been drift-corrected prior to carrying out any simulation (more information in Section 3.3).

2.1.2. *CESM Large Ensemble*

The CESM-LE (Kay et al., 2015) is a set of historial simulations and long-term projections of climate change carried out at NCAR with CESM, version 1. The CESM-LE uses the same model configuration which was applied in CESM-DPLE . While the historical simulations span the period 1850-2005, the climate change projections start in 2006 and cover the 21th century up to 2100, following the RCP 8.5 for the radiative forcings. CESM-LE is composed of 40 members which, as for CESM-DPLE, are randomly generated by perturbing the initial atmospheric conditions. The main difference between CESM-LE and CESM-DPLE is that observational information is not used in the initialization stage of the former. The common experimental designs of both CESM projects allow the analysis of the added value of initialization to predictive skill, avoiding the potential biases which might arise from considering different model configurations. Further details about the model setup and experimental design can be found in Kay et al. (2015).

The fields required as input in DD simulations are not available for the whole CESM-LE simulated period. These fields only cover the period 1990–2005 for historical simulations and 2026–2100 for long-term climate projections. In this case, the whole 40-member ensemble is available. Only the experiments for the period 1990–2005 have been dynamically downscaled and, therefore, only the hindcasts for the last 10 start dates (those initialized every year from 1990 to 1999) have been used to assess the added value of WRF-DPLE to predictive skill over the downscaled uninitialized product. This constraint introduces a large sampling bias in the evaluation which must be taken into account in the discussion of the results. A selection of 4 ensemble members has been done to generate this downscaled product (see Section 3.4.2 for more information about the member selection), which will be referred to as WRF-LE hereinafter. A list of the available fields and the instructions to download them can be accessed through "Data Sets Available to the Community" (n.d.). As done for the decadal experiments, the CESM-LE experiments have been bias-corrected before using them to run WRF (see Section 3.4 for more details).

2.2. Reanalysis datasets

A reanalysis is the output of a model which has been constrained by observations through data assimilation techniques (Chen et al., 2021). If the model is a numerical weather prediction model, the output product is called atmospheric reanalysis. The model, observations and the assimilation scheme are used in combination to produce the best gridded estimates, known as analyses, of past atmospheric states. Reanalysis is the short term to refer to a retrospective analysis. Firstly, the model conducts a short-term forecast starting from previous analysis of the atmospheric state. Then, the data assimilation merges the outputs from this forecast, commonly referred to as first guesses, with observations to produce new analyses of the climate state, which are used to initialize the next short-term forecast (Fujiwara et al., 2017). Since a model is involved in the generation of reanalysis data, these products often provide variables which are not supplied by observational datasets in a gridded format for the whole globe, such as hourly instantaneous temperature or soil moisture at several pressure or soil levels, respectively. This fact makes reanalyses specially suitable to conduct DD simulations. A comprehensive review of the global atmospheric reanalyses currently available has been recently published in the context of the Stratosphere-troposphere Processes And their Role in Climate project (SPARC) of the WRCP. It can be found in SPARC (2022).

The reanalysis products used in this Thesis, ERA-Interim (Dee et al., 2011) and ERA5 (Hersbach et al., 2020), are described in the following.

2.2.1. *ERA-Interim reanalysis*

ERA-Interim (Dee et al., 2011) is one of the full-input atmospheric reanalyses[4] produced by the European Centre for Medium-Range Weather Forecasts (ECMWF). The forecast model of ERA-Interim is the version Cy31r2 of the ECMWF Integrated Forecast System (IFS). ERA-Interim has a spectral T255 horizontal resolution, which is equivalent to a 79 km resolution, approximately, on a reduced Gaussian grid, and 60 vertical levels up to 0.1 hPa. Its archive contains 6-hourly gridded estimates of three-dimensional meteorological variables and 3-hourly estimates of a large amount of surface fields and other two-dimensional variables from January 1979 to August 2019.

In this Thesis, ERA-Interim has been used to address three main tasks:

1) To correct the biases in the decadal prediction and uninitialized information used as input in DD simulations with WRF (see Sections 3.3 and 3.4).
2) To conduct a control DD simulation with WRF for reference purposes in the analysis and production of other DD experiments (see Section 3.8).
3) To conduct a series of sensitivity experiments and analyze the extent to which DD simulations are affected by changes in initial soil moisture conditions (see Section 3.8 and Chapter 6).

ERA-Interim is currently discontinued (“Decommissioning of ECMWF Public Datasets Service”, n.d.) as it has been superseded by ERA5 (Hersbach et al., 2020), a new reanalysis which incorporates several improvements over the former, such as an upgraded forecast model, an improved data assimilation scheme and higher vertical and horizontal resolutions, among other updates. Although ERA5 is considered one

[4] A full input reanalysis uses observational surface, conventional upper-air and satellite data in the assimilation process (SPARC, 2022).

of the most reliable reanalyses available to the community (Chen et al., 2021), it has not been used for the aforementioned tasks because they were started before the ERA5 release.

Other reanalyses could provide an overall performance similar to ERA-Interim (SPARC, 2022), such as the Japanese 55-year Reanalysis (JRA-55; Kobayashi et al., 2015) and the version 2 of the Modern-Era Retrospective Analysis for Research and Applications (MERRA; Gelaro et al., 2017). Nevertheless, ERA-Interim was chosen instead because it has been widely used and tested in DD simulations in Europe and, particularly, in the IP (see, e.g., García-Valdecasas et al., 2020a; Jach et al., 2020; Katragkou et al., 2015).

2.2.2. *ERA5 reanalysis*

ERA5 (Hersbach et al., 2020) is the latest state-of-the-art full-input atmospheric reanalysis produced by ECMWF. It incorporates several updates compared to its predecessor ERA-Interim. Firstly, the forecast model of ERA5 is the version Cy412r of the ECMWF IFS. With respect to the version Cy31r2 used by ERA-Interim, it contains new improvements in all of its components (including atmosphere, land, ocean waves and observations) as well as a new data assimilation methodology. ERA5 also provides a higher horizontal and vertical resolution. Its model runs with a spectral TL639 horizontal resolution, which approximately corresponds to a 31 km resolution, for 137 vertical levels up to 0.01 hPa. In contrast to the 6- and 3-hourly output frequency of ERA-Interim, ERA5 supplies hourly outputs for an even higher set of variables. Additionally, these outputs are available from 1950 to present, and not only from 1979 as those of ERA-Interim. All these updates contribute to improving the reanalysis product compared to what was offered by the predecessor.

ERA5 has been used for two main purposes:

1) To evaluate some variables from CESM-DPLE used as ICs and LBCs in WRF simulations (see Sections 4.5 and 5.4). ERA5 is used instead of observational datasets because it allows working on the CESM-DPLE native resolution, i.e., the resolution of the input data used to drive the DD simulations.
2) To conduct the drift correction of some CESM-DPLE fields in an intercomparison of different correction approaches (see Chapter 8).

2.3. Observational datasets

2.3.1. *Precipitation and near-surface air temperature in Spain*

The Spanish State Meteorological Agency (AEMET, in its Spanish acronym), through the ROCIO_IBEB product, has provided the observational information for daily precipitation (AEMET, 2019), maximum near-surface air temperature (NSAT; AEMET, 2020a) and minimum NSAT (AEMET, 2020b). These variables will be denoted as PR, T_{max} and T_{min}, respectively, hereinafter. The three datasets cover the peninsular territory of Spain and the Balearic Islands with a grid of 5 km resolution. The information is presented at daily scale and has been produced from AEMET station-based measurements which span the period 1951-2022.

The version 2.0 of the PR dataset used in this Thesis incorporates measurements from 3236 stations, as opposed to the 2213 stations considered in version 1.0. The methodology followed to represent the in-situ observed data on the 5 km grid is detailed in Peral-García et al. (2017). An optimal interpolation (Daley, 1991), a statistical interpolation method appropriate for irregular observation distributions, was used to produce analyses of PR from observational data and first guesses of zero value. Since these information sources add biases in the calculation process, a constraint consisting of minimizing the variance of the analysis error was imposed in the algorithm to filter the noise and reveal the signal in the estimations. This procedure is mainly based on the surface analysis system integrated in the HIgh Resolution Limited Area Model (HIRLAM; Navascués et al., 2003; Rodríguez et al., 2003; Undén et al., 2002), a numerical weather prediction system which runs at AEMET and other European national meteorological services (Navascués et al., 2013). The method was extended for PR by Quintana-Seguí et al. (2016) and adjusted by Peral-García et al. (2017) for this dataset to produce daily PR from 07:00 to 07:00 am (local time).

In T_{max} and T_{min} datasets, version 1.0, observations from 1800 stations were used to generate the 5 km-resolution gridded product. The methodology followed is very similar to that for the PR dataset, but including the HIRLAM-AEMET operational analyses for temperature as first guesses after being corrected with observations, as done in Amblar-Francés et al. (2020). Since no information about the daily mean NSAT (T_{mean}) is provided by this AEMET product, it has been calculated as the arithmetic mean of the daily maximum and minimum datasets.

The AEMET datasets have been used to address the following tasks:

1) To recalibrate the WRF-DPLE outputs with the aim of improving as much as pos-

sible their ability to reproduce the observed climate in the IP (see Section 3.5).

2) To divide the IP into homogeneous regions enclosing the locations with similar climate variability for PR and NSAT (see Section 3.6).
3) To evaluate the predictive skill of the recalibrated WRF-DPLE experiments in the IP (see Chapters 4, 5 and 7).

2.3.2. *Global sea level pressure, near-surface air temperature and sea surface temperature*

A few global observational datasets have been used in this Thesis for evaluation purposes. These datasets provide gridded information for sea level pressure (SLP), SST and NSAT. The observational SLP has been supplied by the near-real-time update of Hadley Centre's monthly historical mean sea level pressure (HadSLP2r; Allan and Ansell, 2006). This dataset is presented with a 5° spatial resolution, covering the period from 1850 to 2019. It combines terrestrial and marine observational data from 2228 stations. This information is blended and gridded by using a reduced-space optimal interpolation (Kaplan et al., 2000) to generate the final product. The GISTEMP version 4 (GISTEMP4; Hansen et al., 2010; Lenssen et al., 2019) has been chosen for the evaluation of NSAT. It has been generated from in-situ observational data which have been processed to generate a monthly dataset in a 2°-resolution grid along the period from 1880 to present. The NSAT data are provided as anomaly series with respect to the average over the period 1951–1980. There are gaps in time series which affect some locations, especially in the Southern Hemisphere, so the grid points considered in the evaluation have at least the 70 % of time steps needed to calculate the annual and the multiyear means of the lead time windows (see Section 3.2 for a definition of these lead time windows). Finally, the version 5 of the Extended Reconstructed Sea Surface Temperature (ERSST5; Huang et al., 2017) has been used for the SST evaluation. This dataset considers multiple data sources for in-situ and satellite measurements along with other gridded products of SST and sea ice to reconstruct the monthly series of the global SST. The information is provided from 1854 to present with a 2° spatial resolution.

These datasets have been used in the following tasks:

1) In the case of ERSST5, for the evaluation of the individual member performances in the selection of CESM-DPLE and CESM-LE subensembles which are used in DD simulations (see Sections 3.3.2 and 3.4.2).
2) The three datasets have been considered for the intercomparison of several drift correction methods for DCPs (see Chapter 8).

3

Methodology

This Chapter is dedicated to describe the fundamental methodology applied in this Thesis. Some methodological aspects related to the analyses conducted in Chapters 6 and 8 are only briefly mentioned here, but they are fully detailed in those chapters.

The following sections contain a description of the configuration of WRF and the workflow of the DD simulations, the methodology applied to evaluate the DD experiments, the techniques used to correct the biases in the ICs and LBCs of WRF, the approach followed to recalibrate the WRF-DPLE experiments, a description of the soil initialization, a complete list of the DD experiments and a mention of the most relevant software used in the course of this Thesis.

3.1. The Weather Research and Forecasting model (WRF)

The DD simulations have been conducted with WRF by using the version 3.9.1.1 of the Advanced Research WRF (ARW) dynamics solver (Skamarock et al., 2008; Wang et al., 2008). The version 3.9.1 of the WRF Preprocessing System (WPS; Wang et al., 2008) has been used to prepare and write the WRF input data in the appropriate format to run the simulations. The source code of WRF and WPS is open and freely available in WRF Developers (2017) and WPS Developers (2017), respectively.

3.1.1. *Domain configuration and nesting approach*

The DD simulations were conducted over two nested domains (Figure 3.1). The coarse-grid domain, d01, is based on the region defined for the European Coordinated Regional Climate Downscaling Experiment (EURO-CORDEX; e.g., Jacob et al., 2014; Kotlarski et al., 2014), which is widely used for simulations carried out over Europe in the framework of CORDEX (CORDEX, 2015; Giorgi and Gutowski, 2015; Giorgi et al., 2009). This domain has a resolution of 0.44° (~ 50 km) and a size of 126×123

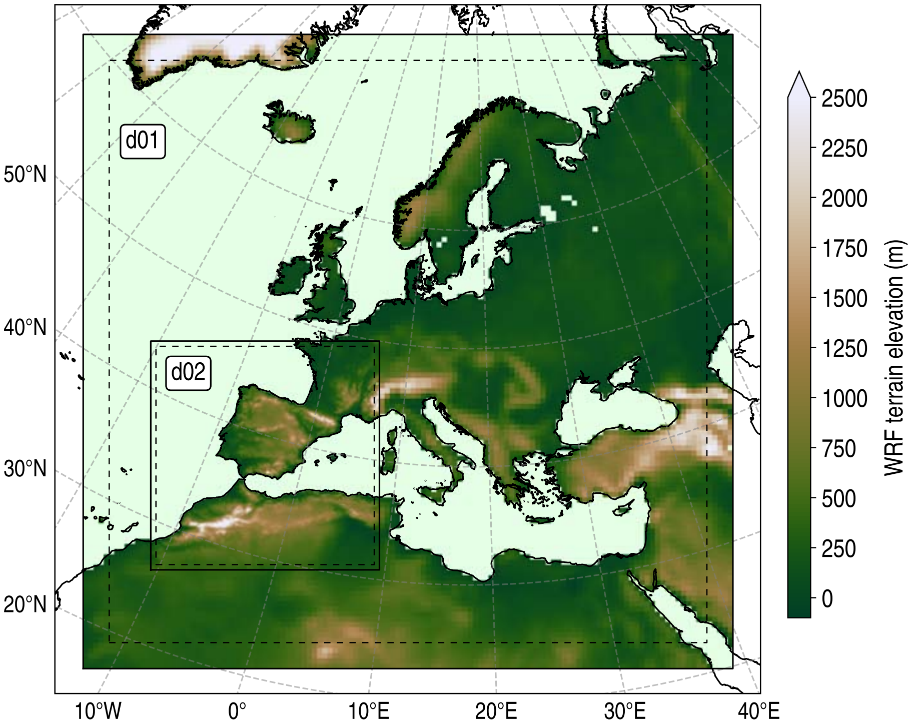

Figure 3.1: Domains for the DD simulations conducted with WRF. The domain d01, with a ∼ 50 km resolution, is based on the EURO-CORDEX region, whereas the domain d02 is centered in the IP and spans Spain, Portugal, part of France and the north of Africa, with a ∼ 10 km resolution. While the solid lines denote the boundaries of each domain, the dashed lines identify the inner boundaries of the buffer zones. The terrain elevation data have been retrieved from Earth Resources Observation And Science Center (2017).

grid points (∼ 6300 × 6150 km). It is defined with a latitude-longitude projection and a rotated pole placed at coordinates 162° W and 39.25° N to maintain a quasi-uniform resolution in length units over the whole domain. It covers the northeastern Atlantic Ocean, spans the European continent, including the westernmost regions of Russia, and extends over the north of Africa. The fine-grid domain, d02, nested in the previous one, is centered on the IP, extending southwards over northern Africa, northwards over France and also covering the Balearic Islands. This domain has a resolution of 0.088° (∼ 10 km) and a size of 221 × 221 grid points (∼ 2210 × 2210 km). Previous studies have shown that resolutions around 10 km are appropriate to reproduce the observed climate with RCMs, consistently providing more realistic results than coarser resolutions, especially in the case of precipitation in regions with complex orography or delimited by the coast (Demory et al., 2020; Prein et al., 2016). The resolution jump between both WRF domains is 5, keeping a value below 12, as suggested by Denis et al. (2003) to guarantee the generation of a reliable regional climate.

The DD simulations were conducted with a one-way nesting approach (e.g., Beck et al., 2004; Denis et al., 2002). It means that WRF was forced by LBCs through an unidirectional communication channel which did not allow feedbacks between the WRF results and the driving data, as opposed to the two-way nesting (Lorenz and Jacob, 2005). Due to the weak feedback provided to the GCM by the RCM in Europe and the complexity of running in a two-way mode, the one-way nesting is the most common approach in DD simulations (Giorgi, 2019). The LBCs supplied to WRF were updated every 6 hours. According to Denis et al. (2003), this frequency is appropriate for DD experiments which use the one-way technique to drive their 45 km-resolution RCM, without appreciable differences with experiments for which the LBCs are updated every 3 hours. Therefore, a similar behaviour is expected for the 50-km resolution domain defined here. In any case, the frequency of WRF input information is restricted by data availability. Here, the maximum output frequency available in ERA-Interim and CESM-DPLE/-LE datasets for the mandatory fields is 6-hourly, so there is not a higher-frequency alternative.

3.1.2. *Relaxation procedures toward driving fields*

As mentioned in Section 1.1.3, the difference in grid resolution between the LBCs and the RCM produces mismatches between the solutions and driving data on the domain lateral boundaries, which must be addressed by applying relaxation techniques (Marbaix et al., 2003). A linear-exponential relaxation approach, whose mathematical formulation is detailed in Skamarock et al. (2008), was followed across the buffer zone. In this procedure to relax the RCM solutions towards the driving fields, the LBCs are updated by a function with two linear ramping weight coefficients, both multiplied by a exponential factor to get a smoother result (Wang et al., 2008). The relaxation is applied for horizontal wind components, potential temperature, water vapor, and the perturbed fields of geopotential and pressure (Skamarock et al., 2008). The buffer zones in both domains have sizes of 5 grid points, the WRF default value (Wang et al., 2008), which are equivalent to lengths about 250 km and 50 km for the coarse- and fine-grid domains, respectively.

In addition to the relaxation in the buffer zone, a spectral nudging (see Section 1.1.3; Storch et al., 2000) was also applied for some fields across the whole coarse-grid domain (but not for the fine one) to prevent the generation of small-scale processes inconsistent with the large-scale circulation reproduced by the driving data during the simulations. The spectral nudging contributes to eliminating the distortion of the large-scale circulation along the lateral boundaries and the sensitivity of RCM results to the geometry and location of the domain (Miguez-Macho et al., 2004; Radu et al., 2008). It has been applied on waves roughly longer than 600 km, only for the coarse-grid domain, to force the RCM to reproduce the large-scale circulation during the generation of the fields which provide the fine-grid domain with the lateral boundary information (Messmer et al., 2017). The nudging has not been conducted in the fine-grid domain so that the RCM can freely develop the small-scale climate features over the IP. Temperature, horizontal wind components and geopotential height have been nudged with a 6-hourly frequency, the same frequency considered to update the LBCs. Following Miguez-Macho et al. (2004), humidity has not been nudged because it can present very pronounced horizontal and vertical gradients which may be missed by coarse GCM or reanalysis data. A typical nudging coefficient of $3 \cdot 10^{-4}\,\mathrm{s}^{-1}$ has been considered for the spectral nudging, being applied only above the planetary boundary layer (PBL) to avoid inconsistencies in the simulated climate as a result of the differences between the surface characteristics of the GCM and RCM (Gómez and Miguez-Macho, 2017).

3.1.3. *Description of the model core and workflow*

The dynamical core of WRF, known as the ARW dynamics solver, is the part of the WRF modeling system which integrates the equations governing the exchanges of mass, energy and momentum in the simulated system, i.e., the compressible and non-hydrostatic Euler equations (Skamarock et al., 2008). The solutions of these equations provide the spatio-temporal fields of the prognostic variables (wind speed components, potential temperature, humidity, geopotential perturbation and pressure perturbation). These equations are formulated by using a terrain-following mass coordinate system proposed by Laprise (1992). They are integrated in time by using a third-order Runge-Kutta scheme with a time-split approach (Wicker and Skamarock, 2002). The integration time steps were set to 240 s and 48 s for the coarse- and fine-grid domains, respectively, as considered in "Model" (n.d.) for domains with similar resolutions to those used in this Thesis. The spatial discretization is done by applying an Arakawa-C grid staggering, which consists of locating most scalar variables in the center of the domain grid cells and horizontal wind velocities and geopotential in horizontal and vertical cell boudaries, respectively. Here, the Euler equations were solved in 40 vertical levels with the pressure top located at 10 hPa. Further information about the ARW dynamics solver is available in Skamarock et al. (2008).

The workflow applied to produce the final output experiments which have been analyzed in the next chapters can be divided into three phases (Figure 3.2):

A) Input data pre-processing with WPS.

B) DD simulations with ARW.

C) Output data post-processing and storage.

Two types of datasets are used as input information in the phase A. The first type is the static geographical data which provide the RCM with information about constant terrestrial information (terrain height, land use categories, soil types, etc.). It can be downloaded from the WRF Users' Page ("WRF V3 Geographical Static Data Downloads Page", n.d.). The second type is the information supplied by the bias-corrected GCM (CESM-DPLE and CESM-LE) or reanalysis (ERA-Interim) data, which contain the time-varying fields which define the ICs and LBCs in DD simulations (Table 3.1). While ERA-Interim provides input fields with a 6-hourly aggregation, the time frequency of the GCMs variables is lower for some slow-variant fields, such as soil moisture or sea surface temperature. In those cases, 6-hourly time series were constructed by assigning to each time step the corresponding daily or monthly field value before correcting the bias (see Sections 3.3 and 3.4 for information about the bias correction). Both datasets are supplied to WPS, which prepares the data in the appropriate format to create the ICs and LBCs which are used to conduct the DD simulations. The WPS environment is composed of three modules:

A.1) **The Geogrid module.** The role of this module is to create the RCM domains and interpolate the static geographical information onto them. The file `geo_em.d0x` contains the information generated for the domain `d0x`.

A.2) **The Ungrib module.** While CESM-DPLE and CESM-LE data are provided in the WRF intermediate format, the data from ERA-Interim are originally written in the GRIB format. This module extracts the time-varying meteorological information from GRIB files and write it in the `FILE:<date>` files with the WRF intermediate format, where `<date>` is the date corresponding to that information.

A.3) **The Metgrid module.** The `geo_em.d0x` and `FILE:<date>` files are the inputs of the Metgrid submodule, which horizontally interpolates the meteorological data onto the domains defined by Geogrid to create the `met_em.d0x.<date>`

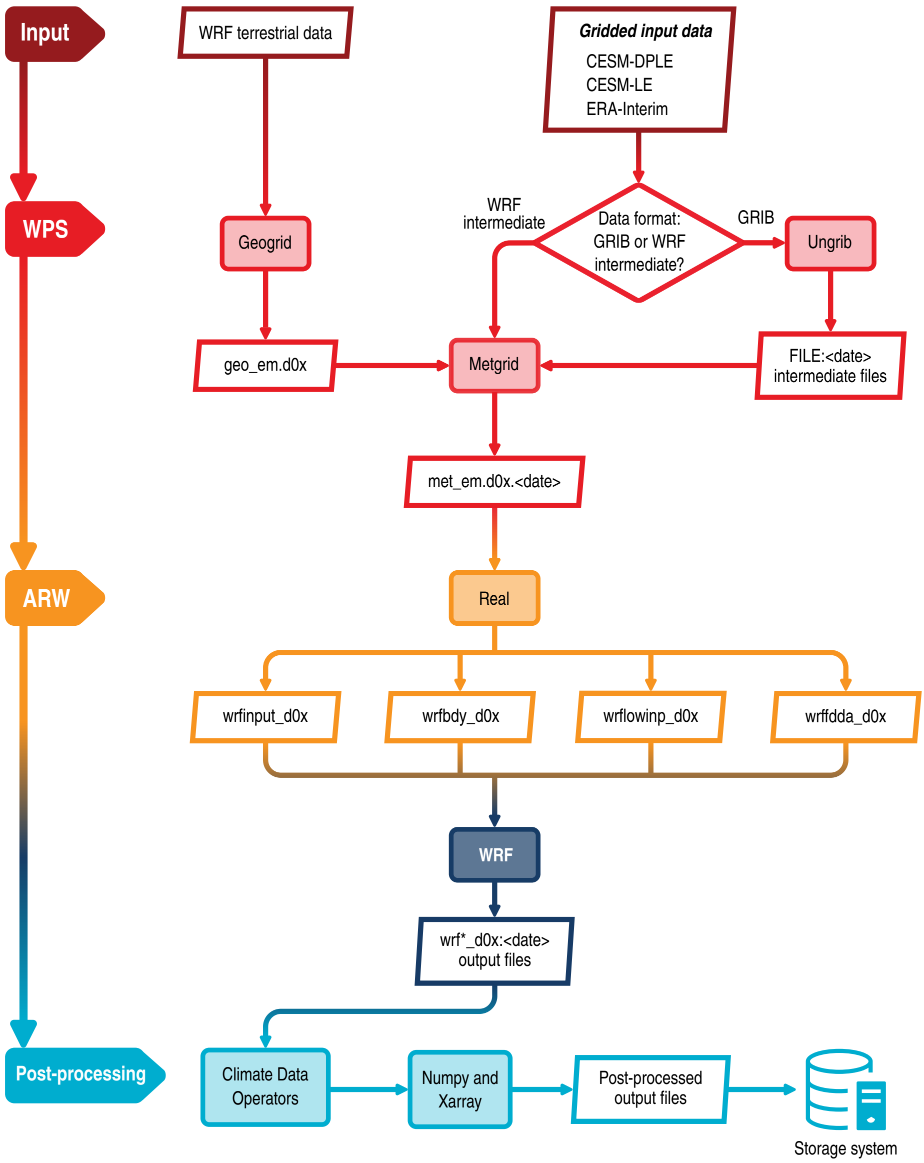

Figure 3.2: Flow chart for the DD simulations conducted with WRF.

files. These files contain both static geographical and time-varying meteorological information.

The phase B is carried out with ARW, which constitutes the fundamental core of WRF. It creates the ICs and LBCs and conducts the DD simulations. As WPS, the

Table 3.1: Variables used as input information in the DD simulations conducted with WRF.

Variables	Component	Levels	Original time frequency	
			CESM-DPLE/-LE	ERA-Interim
air temperature	atmosphere	pressure	6-hourly	6-hourly
zonal wind speed	atmosphere	pressure	6-hourly	6-hourly
meridional wind speed	atmosphere	pressure	6-hourly	6-hourly
relative humidity	atmosphere	pressure	6-hourly	6-hourly
geopotential height	atmosphere	pressure	6-hourly	6-hourly
sea level pressure	atmosphere	surface	6-hourly	6-hourly
surface pressure	atmosphere	surface	6-hourly	6-hourly
2-m air temperature	atmosphere	surface	6-hourly	6-hourly
10-m zonal wind speed	atmosphere	surface	6-hourly	6-hourly
10-m meridional wind speed	atmosphere	surface	6-hourly	6-hourly
2-m relative humidity	atmosphere	surface	6-hourly	6-hourly
2-m relative humidity	atmosphere	surface	6-hourly	6-hourly
skin temperature	atmosphere	surface	6-hourly	6-hourly
snow depth water equivalent	atmosphere	surface	monthly	6-hourly
sea ice fraction	atmosphere	surface	daily	6-hourly
sea surface temperature	ocean	surface	daily	6-hourly
soil temperature	land	soil	monthly	6-hourly
soil moisture	land	soil	monthly	6-hourly
land sea mask	–	surface	constant	constant
terrain elevation	–	surface	constant	constant

ARW environment is composed of several modules:

B.1) **The Real module.** This module reads the static and meteorological information from the `met_em.d0x.<date>` files at first. Then, it prepares the soil and atmospheric fields which are used in DD simulations by vertically interpolating them on the RCM soil and height levels, respectively. This module also verifies that all mandatory variables are available and that there are not inconsistencies in the information provided by `met_em.d0x.<date>` files. The spectral nudging described in Section 3.1.2 is started at this stage too. The ICs and LBCs are saved in the `wrfinput_d0x` and `wrfbdy_d0x`, respectively. Additional information for the low boundary of the RCM, referred to time-varying variables which are not generated by the model but are required to run the simulations (sea surface temperature, albedo, vegetation fraction, etc.), are saved in `wrflowinp_d0x` files. Finally, the fields needed to apply the spectral nudging are saved in `wrffdda_d0x` files. Since it has been applied only over the coarser domain, a single file is generated.

B.2) **The WRF module.** It contains the ARW dynamics solver, the dynamical core of the WRF modeling system. Its role is to solve the equations that govern the physical processes which determine the evolution of the simulated system and provide the solutions in form of output spatio-temporal fields through the `wrf*_d0x:<date>` files.

In the phase C, the fields contained in the `wrf*_d0x:<date>` files are post-processed to organize and rewrite the output information with the aim of making it easy to read and handle. The post-processing is done by combining tools from Climate Data Operators (CDO; Schulzweida, 2023), Numpy (Harris et al., 2020) and Xarray (Hoyer and Hamman, 2017). Once this task finishes, the data are sent to a private storage system from which they can be downloaded by authorized users.

3.1.4. *The physics parametrization schemes*

In the context of climate modeling, a parametrization is the representation of a physical process which cannot be directly solved by the climate model but that is still important for the behaviour of the climate system (Kotamarthi et al., 2021). These processes often involve mechanisms which need a higher spatial or temporal resolution to be explicitly determined or which are computationally too expensive. Parametrizations represent these unsolved processes as functions of quantities which can be solved in models, considering conceptual approximations or empirical relationships derived from observations and process studies (McFarlane, 2011).

WRF gives the possibility of choosing among a wide range of parametrization options. Some combinations of parametrizations may be more suitable than others depending on the scope of the study, the region of interest or the domain configuration, not only regarding the ability to accurately reproduce the evolving climate but also in terms the computing cost of the simulations. This RCM has been object of multiple sensitivity analyses to examine its response to different combinations of parametrizations schemes in the IP (see, e.g., Argüeso et al., 2011; Borge et al., 2008; García-Valdecasas, 2018; Santos-Alamillos et al., 2013). The parametrization schemes chosen in this Thesis are the same that were selected by García-Valdecasas (2018) in a study carried out for the same domains considered here. Among the options tested by García-Valdecasas (2018), this selection is generally the most appropriate configuration of parametrizations to simulate the current climate in the IP, providing a good compromise between the computing cost of the simulations and their ability to reproduce the observed climate. This configuration includes parametrizations for microphysics, cumulus, the land surface, the PBL, radiation and the surface layer:

- **Microphysics.** Microphysics parametrizations resolve water vapour, cloud and precipitation-related processes (Skamarock et al., 2008). Here, the WRF Single-Moment 3-class scheme (WSM3; Hong et al., 2004) has been used. WSM3 accounts for three types of hydrometers (vapour, cloud water/ice and rain/snow). It provides a realistic representation of ice physics at a 10–30-km model resolution while being computationally efficient.

- **Cumulus.** Cumulus schemes represent the effects of convective and shallow clouds at subgrid scale (Skamarock et al., 2008). The Betts-Miller-Janjic scheme (BMJ; Janjić, 1994, 2000) has been chosen in this case. It is built upon the Betts-Miller scheme (Betts, 1986; Betts and Miller, 1986) and includes some modifications to improve the representation of deep and shallow convection (Janjić, 1994).

- **Land surface.** Land surface models (LSMs) take information from surface layer scheme, radiative forcing from the radiation scheme, precipitation from the microphysics and cumulus schemes, together with internal land fields and land surface properties to generate heat and moisture fluxes at a grid cell level over land and sea ice. Among all available options, the Noah LSM (Chen and Dudhia, 2001; Ek et al., 2003; Wang et al., 2010) is used here. Noah LSM uses four soil layers with depths 0–10 cm, 10–40 cm, 40–100 cm and 100–200 cm, and only a canopy layer. More details about this parametrization are provided in Section 6.2.

- **Planetary boundary layer (PBL).** The parametrization of the PBL resolves the subgrid vertical fluxes associated to eddy transport not only in the PBL but also in the whole atmospheric column. The flux profiles are determined in the boundary and stable layers, whereas the atmospheric tendencies of temperature, humidity and horizontal momentum are calculated for the entire column. It takes the surface fluxes from the surface layer and LSM (Skamarock et al., 2008). The version 2 of the Asymmetrical Convective Model (ACM2; Pleim, 2007) parametrizes these processes in the WRF simulations conducted here. While the original version of the model represents the large-scale convective transport but omit subgrid turbulent mixing, ACM2 incorporates an eddy diffusion scheme to also account for the small-scale turbulent transport processes.

- **Radiation.** Radiation schemes represent the atmospheric heating due to the net radiative flux and the downward long- and short-wave radiation in surface (Skamarock et al., 2008). The spectral-band schemes used in the version 3 of the Community Atmosphere Model (CAM3; Collins et al., 2004) have been chosen for the representation of long- and short-wave radiation. They interact with cloud fraction and take into account trace gases and aerosols to solve the radiative processes.
- **Surface layer.** Friction velocities and exchange coefficients are resolved by surface layer schemes. They allow the calculation of heat and moisture fluxes by land surface models and surface stress by the PBL parametrization (Skamarock et al., 2008). These processes are represented by a revision of the surface layer scheme for the fifth-generation Penn State/NCAR Mesoscale Model (MM5; Grell et al., 1994), originally based on the Monin-Obukhov similarity theory (Monin and Obukhov, 1954). The revision was formulated by Jiménez et al. (2012).

3.2. Evaluation of the predictive skill of decadal climate predictions

The methodology applied in the evaluation of the decadal experiments is fundamentally described in Goddard et al. (2013), where the authors establish the coordinated framework for the verification of interannual-to-decadal predictions which is commonly followed in this research field.

Let $Y'_{kj\tau}$ be a DCP, where k is the member of the ensemble, j stands for the start date and τ denotes the lead time. On the other hand, let $X'_{j\tau}$ be the verification data used to evaluate the predictive skill of the DCPs. This verification dataset can be constituted by observational or reanalysis information (see Chapter 2). Note that this dataset encompasses continuous time series which span the whole period considered in the evaluation. Thus, it cannot be associated to specific start dates j or lead times τ

in the same way as DCPs are. Here, $X'_{j\tau}$ does not denote the verification information at lead time τ in an experiment initialized on the date j, but the verification information in that continuous series at the time which corresponds to the start date j and lead time τ in $Y'_{kj\tau}$.

Unless otherwise indicated, the evaluation of the DCPs is focused on field anomalies. The use of anomalies reduces the biases associated to the lead time-dependent climatological mean of predictions, also known as the mean drift or the unconditional bias, which may potentially impact on the metrics considered for the evaluation. For predictions, the lead time-dependent climatology is calculated as

$$\overline{Y'}_{k\tau} = \frac{1}{N_\mathrm{d}} \sum_{j=1}^{N_\mathrm{d}} Y'_{kj\tau} \,, \qquad [3.1]$$

where N_d is the sample size, i.e., the number of start dates. For the verification data, it is given by

$$\overline{X'}_{\tau} = \frac{1}{N_\mathrm{d}} \sum_{j=1}^{N_\mathrm{d}} X'_{j\tau} \qquad [3.2]$$

Therefore, the anomaly of $Y'_{kj\tau}$ is

$$Y_{kj\tau} = Y'_{kj\tau} - \overline{Y'}_{k\tau} \,, \qquad [3.3]$$

whereas the anomaly of $X'_{j\tau}$ is obtained from

$$X_{j\tau} = X'_{j\tau} - \overline{X'}_{\tau} \qquad [3.4]$$

The ensemble mean $\{Y\}_{j\tau}$ is calculated as follows:

$$\{Y\}_{j\tau} = \frac{1}{N_\mathrm{ens}} \sum_{k=1}^{N_\mathrm{ens}} Y_{kj\tau} \,, \qquad [3.5]$$

where N_ens is the number of ensemble members, i.e., the ensemble size.

Considering the ensemble mean $\{Y\}_{j\tau}$ and the verification data $X_{j\tau}$, the averages along a given lead time period are

$$\{\hat{Y}\}_j = \frac{1}{\theta_2 - \theta_1 + 1} \sum_{\tau=\theta_1}^{\theta_2} \{Y\}_{j\tau} \qquad [3.6]$$

$$\hat{X}_j = \frac{1}{\theta_2 - \theta_1 + 1} \sum_{\tau=\theta_1}^{\theta_2} X_{j\tau} \ , \qquad [3.7]$$

where θ_1 and θ_2 are the first and last time steps, respectively, of the lead time period.

In this Thesis, the evaluation of the DCPs has been done by considering several lead time periods with lengths of 1, 4 and 8 years. Most part of the analysis of the DCPs has been focused on lead years 1, 2–5, 6–9 and 2–9, as suggested in Goddard et al. (2013). The length of the lead time window influences on the frequency noise which is reduced by averaging. Lead year 1 contains seasonal-to-interannual variability and presents the largest imprint of initialization. The highest frequencies are reduced in lead years 2–5, 6–9 and 2–9. Lead years 2–5 and 6–9 provide information of the dependence of the performance of DCPs on lead time. While DCPs may be still more influenced by year-to-year variability in lead years 2–5, the contribution of the climate change signal to the predictions is higher in lead years 6–9. The lead years 2–9 average represents the decadal scale, excluding the contribution of the initialization contained in the first year and the interannual variability frequencies.

The predictive skill of the WRF-DPLE hindcasts has been assessed for the control period, which comprises the experiments initialized every year from 1970 to 1999. However, when comparing with WRF-LE, only the experiments initialized from 1990 to 1999 has been used to match the period with available WRF-LE data. On the other hand, the set of WRF-DPLE DCPs initialized in 2015 spans the decade 2015–2025. The evaluation has been carried out at annual scale as well as for each boreal season. Winter represents the 3-month average of the monthly field in December, January and February (DJF); the average of March, April and May is calculated for spring (MAM); for summer, June, July and August (JJA); finally, autumn is obtained from the average of September, October and November (SON).

In general, the verification and other reference datasets used in the evaluation conducted in Chapters 4 to 7 have been interpolated to the WRF-DPLE spatial grid (described in Section 3.1.1) by using a bilinear approach. In the evaluation conducted in Sections 4.5 and 5.4, the verification datasets have been interpolated to the coarser CESM-DPLE grid (described in Section 2.1.1). In Chapter 8, however, CESM-DPLE data have been interpolated to the coarser grids of the verification datasets.

3.2.1. *Evaluation of the accuracy with deterministic metrics*

The term *accuracy* is usually defined as the degree of correspondence between predictions and observations (Murphy, 1988). The mean squared error (MSE) is used in this Thesis as the basic measure of the accuracy of DCPs. Following Murphy (1988), the MSE for the DCPs along the period corresponding to the start dates $j = 1, ..., N_\mathrm{d}$ is defined from Eqs. [3.6] and [3.7] as

$$\mathrm{MSE}(\{Y\}, X) = \frac{1}{N_\mathrm{d}} \sum_{j=1}^{N_\mathrm{d}} (\{\hat{Y}\}_j - \hat{X}_j)^2 \,, \qquad [3.8]$$

where the difference between DCPs and verification data for the start date j is the anomaly error:

$$E_j = \{\hat{Y}\}_j - \hat{X}_j \qquad [3.9]$$

Lower values of MSE indicate a better prediction accuracy. The root mean square error (RMSE) can be calculated from MSE by taking the squared root in Eq. [3.8]. An advantage which RMSE has over MSE is that the former is expressed in the same units as the evaluated field, so it is more easily interpretable as a measure of the error. It is defined as

$$\mathrm{RMSE}(\{Y\}, X) = \sqrt{\frac{1}{N_\mathrm{d}} \sum_{j=1}^{N_\mathrm{d}} (\{\hat{Y}\}_j - \hat{X}_j)^2} \qquad [3.10]$$

The relative RMSE, or RMSE$_R$, is calculated by dividing the anomaly error E_j by X'_j in Eq. [3.10]:

$$\mathrm{RMSE_R}(\{Y\}, X) = \sqrt{\frac{1}{N_\mathrm{d}} \sum_{j=1}^{N_\mathrm{d}} \left(\frac{\{\hat{Y}\}_j - \hat{X}_j}{\hat{X}'_j}\right)^2} \,, \qquad [3.11]$$

with the relative anomaly error as

$$E_{\mathrm{R},j} = \frac{\{\hat{Y}\}_j - \hat{X}_j}{\left|\hat{X}'_j\right|} \qquad [3.12]$$

While the metrics presented above evaluate the performance of DCPs in terms of their ability to predict the magnitude of the verification anomalies, the anomaly correlation coefficient (ACC) can be used to evaluated the ability to reproduce their variability. According to Wilks (2006, in Section 7.6.4), the ACC is designed to detect similarities in the patterns of departures (i.e., anomalies) from the climatological mean. It is defined as

$$\mathrm{ACC}(\{Y\}, X) = \frac{\sum_{j=1}^{N_\mathrm{d}} \{\hat{Y}\}_j \cdot \hat{X}_j}{\sqrt{\sum_{j=1}^{N_\mathrm{d}} (\{\hat{Y}\}_j)^2} \sqrt{\sum_{j=1}^{N_\mathrm{d}} (\hat{X}_j)^2}}, \qquad [3.13]$$

The ACC is computed as the Pearson correlation coefficient. It has two important properties. Firstly, ACC ranges from -1 to 1, with $\mathrm{ACC} = -1$ and $\mathrm{ACC} = 1$ indicating a perfect negative and positive linear association, respectively, between $\{\hat{Y}\}_j$ and $\hat{X}_j$. Secondly, ACC^2 denotes the proportion of the variance of one variable explained by the other (Wilks, 2006, in Section 3.5.2).

In the evaluation of the accuracy of a DPS, sometimes it is useful to conduct the assessment in terms of its predictive *skill*, which is the accuracy of the predictions of interest relative to other reference predictions or experiments. Following Murphy (1988), a generic skill score K can be defined in terms of a generic measure of accuracy A as follows:

$$K = \frac{A - A_\mathrm{r}}{A_\mathrm{p} - A_\mathrm{r}}, \qquad [3.14]$$

where A, A_p and A_r denote the measures of the accuracy of the predictions of interest, the perfect predictions and the reference predictions or experiments, respectively. The skill score K represents the improvement in terms of the accuracy of the predictions over the reference experiments relative to the maximum potential improvement.

The skill score based on MSE is commonly denoted as the mean squared skill score (MSSS; Goddard et al., 2013). Considering Z as the reference experiments and that $\mathrm{MSE}(\{Y\}_\mathrm{p}, X) = 0$ is satisfied for a perfect prediction, this skill score is calculated as

$$\mathrm{MSSS}_Z = \mathrm{MSSS}(\{Y\}, Z, X) = \frac{\mathrm{MSE}(\{Y\}, X) - \mathrm{MSE}(Z, X)}{\mathrm{MSE}(\{Y\}_\mathrm{p}, X) - \mathrm{MSE}(Z, X)} =$$

$$= 1 - \frac{\mathrm{MSE}(\{Y\}, X)}{\mathrm{MSE}(Z, X)} \qquad [3.15]$$

The MSSS_Z is positive when the accuracy of the test predictions is higher than the accuracy of the reference experiments, i.e., $\mathrm{MSE}(\{Y\},X) < \mathrm{MSE}(Z,X)$. It is negative under the opposite conditions, with $\mathrm{MSE}(\{Y\},X) > \mathrm{MSE}(Z,X)$. If $\mathrm{MSSS}_Z = 1$, then $\mathrm{MSE}(\{Y\},X) = 0$ (perfect prediction), whereas if $\mathrm{MSSS}_Z = 0$, then $\mathrm{MSE}(\{Y\},X) = \mathrm{MSE}(Z,X)$. Although MSSS_Z is upper bounded by 1, it is not lower bounded (Murphy, 1988).

According to the decomposition that Murphy (1988) applied over MSSS, Eq. [3.15] can be written as follows:

$$\mathrm{MSSS}_Z = \frac{\mathrm{ACC}(\{Y\},X)^2 - \mathrm{CB}(\{Y\},X)^2 - \left[\mathrm{ACC}(Z,X)^2 - \mathrm{CB}(Z,X)^2\right]}{1 - [\mathrm{ACC}(Z,X)^2 - \mathrm{CB}(Z,X)^2]} \,, \qquad [3.16]$$

where CB is the conditional bias of a given dataset with respect to the verification data. For example, the CB of $\{\hat{Y}\}_j$ is expressed by

$$\mathrm{CB}(\{Y\},X) = \mathrm{ACC}(\{Y\},X) - \frac{s_{\{Y\}}}{s_X} \,, \qquad [3.17]$$

with

$$s_{\{Y\}} = \sqrt{\frac{1}{N_\mathrm{d}} \sum_{j=1}^{N_\mathrm{d}} (\{\hat{Y}\}_j)^2} \qquad [3.18]$$

and

$$s_X = \sqrt{\frac{1}{N_\mathrm{d}} \sum_{j=1}^{N_\mathrm{d}} (\hat{X}_j)^2} \qquad [3.19]$$

as the standard deviations of predictions and verification data, respectively, along the start dates $j = 1, ..., N_\mathrm{d}$ at a certain lead time average.

Following Murphy and Epstein (1989), the CB is interpreted in terms of the linear regression between $\hat{X}_j$ and $\{\hat{Y}\}_j$:

$$\hat{X}_j = m \cdot \{\hat{Y}\}_j + n \,, \qquad [3.20]$$

where m and n are the slope and intercept of the regression model, respectively. While the slope m is expressed as

$$m = \frac{s_X}{s_{\{Y\}}} \cdot \mathrm{ACC}(\{Y\},X) \,, \qquad [3.21]$$

the intercept n is written as

$$n = \overline{\hat{X}} - m \cdot \overline{\{\hat{Y}\}}\,, \qquad [3.22]$$

where $\overline{\hat{X}}$ and $\overline{\{\hat{Y}\}}$ are the averages of $\hat{X}_j$ and $\{\hat{Y}\}_j$ along the start dates j, respectively. It should be noted that $\overline{\{\hat{Y}\}} = \overline{\hat{X}} = n = 0$ is satisfied here because all calculations are done over anomalies, but this relation would not be necessarily satisfied if calculations were done with full fields. Although these quantities are equal to zero in this case, they are kept in equations in benefit of a general interpretation of CB.

By substituting Eq. [3.20] in Eq. [3.9], the anomaly error E_j can be written as

$$E_j = \{\hat{Y}\}_j - \hat{X}_j = \{\hat{Y}\}_j - m \cdot \{\hat{Y}\}_j - n = \{\hat{Y}_j\} \cdot (1 - m) - n \qquad [3.23]$$

If $m \neq 1$, E_j systematically depends on the value of $\{\hat{Y}\}_j$ in the start date j; therefore, the prediction is *conditionally* biased. On the other hand, if $m = 1$ (an optimal result for a prediction of the verification anomaly $\hat{X}_j$), then $E_j = -n$ and $\mathrm{CB}(\{Y\}, X) = 0$ in Eq. [3.17], so the conditional bias is removed. The prediction would also be *unconditionally* biased if $\overline{\{\hat{Y}\}} - \overline{\hat{X}} \neq 0$ (i.e., if the mean drift were not be removed), but this is not the case because calculations are done with anomalies. Thus, in absence of CB, $E_j = -n = 0$ is always satisfied here. Positive or negative values of CB are caused by an imbalance in Eq. [3.17] which prevents the desirable result $m = 1$.

If MSSS_Z in Eq. [3.16] is calculated with climatology as the reference experiment, it can be written as follows (Murphy, 1988):

$$\begin{aligned}\mathrm{MSSS_C} = \mathrm{MSSS}(\{Y\}, \overline{X}, X) = 1 - \frac{\mathrm{MSE}(\{Y\}, X)}{\mathrm{MSE}(\overline{X}, X)} = 1 - \frac{\mathrm{MSE}(\{Y\}, X)}{s_X} = \\ = \mathrm{ACC}(\{Y\}, X)^2 - \mathrm{CB}(\{Y\}, X)^2\,, \qquad [3.24]\end{aligned}$$

where $\mathrm{ACC}(\{Y\}, X)^2$ is a measure of the maximum potential skill which can be attained by removing the conditional bias. If the calculations were done with full fields, an additional component representing the unconditional bias would be present in Eq. [3.24].

In the assessment of the predictive skill, the difference between the performances of test predictions and reference experiments in terms of ACC is calculated as follows:

$$\Delta\mathrm{ACC}_Z = \mathrm{ACC}(\{Y\}, X) - \mathrm{ACC}(Z, X) \qquad [3.25]$$

On the other hand, the difference in terms of CB is expressed as

$$\Delta \mathrm{CB}_Z = |\mathrm{CB}(Z, X)| - |\mathrm{CB}(\{Y\}, X)| \qquad [3.26]$$

In both $\Delta\mathrm{ACC}_Z$ and $\Delta\mathrm{CB}_Z$, positive (negative) results indicate a better (worse) performance of the test predictions compared to the reference experiments.

DCPs are characterized by having a very low signal-to-noise ratio over the North Atlantic latitudes for some variables such as PR or SLP (Smith et al., 2019). In these cases, very large ensemble sizes are needed to properly extract the climate signal after removing the unpredictable background noise by calculating the ensemble mean, as mentioned in Section 1.1.2. The low signal-to-noise ratio may lead to a very curious phenomenon known as the *signal-to-noise paradox*, in which the model is better at predicting the real world than at predicting itself (Scaife and Smith, 2018; Smith et al., 2019, 2020). The degree of presence of this paradox in DCPs is measured by the ratio of predictable components (RPC; Scaife and Smith, 2018):

$$\mathrm{RPC}^2 = \frac{\mathrm{ACC}(\{Y\}, X)^2}{\{\mathrm{ACC}(\{Y\}, Y)\}^2}\,, \qquad [3.27]$$

where $\mathrm{ACC}(\{Y\}, X)$ is the correlation between the model ensemble mean and the verification data, calculated in Eq. [3.13], and $\{\mathrm{ACC}(\{Y\}, Y)\}$ is the average correlation between the model ensemble mean and a single ensemble member. The ideal result in Eq. [3.27] is $\mathrm{RPC} = 1$, which means that verification and model climates contain the same ratio of predictable variance. If $\mathrm{RPC} < 1$, $\mathrm{ACC}(\{Y\}, X)$ would be smaller than its desirable value given the ratio of predictable variance of the model. This can be a consequence of the use of too few ensemble members to remove the unpredictable noise, low spread of the prediction ensemble, systematic errors in the predicting signal, etc. The signal-to-noise paradox emerges with $\mathrm{RPC} > 1$, leading to a counterintuitive situation in which $\mathrm{ACC}(\{Y\}, X)^2 > \{\mathrm{ACC}(\{Y\}, Y)\}^2$ is satisfied, so the model is better at replicating the real world climate than its own climate (Scaife and Smith, 2018; Smith et al., 2019, 2020). The weak model signals associated to the signal-to-noise paradox negatively affect the predictive skill of DCPs in terms of both accuracy and reliability (Scaife and Smith, 2018), so it has been taken into account for the discussion of the results obtained in this Thesis, especially in Chapters 4 and 5.

Finally, the calculation of trends is done by means of the Theil-Sen's slope estimator (Sen, 1968). For a time series composed of pairs (t_j, F_j), with a field anomaly F_j (i.e., $\{\hat{Y}_j\}$ or $\hat{X}_j$) measured at time t_j with samples or start dates $j = 1, \ ..., N_\mathrm{d}$, this approach estimates the trend as the median of the slopes β^F_{ij} which are calculated as follows:

$$\beta^\mathrm{F}_{ij} = \frac{F_j - F_i}{t_j - t_i} \ , \qquad [3.28]$$

with $t_j > t_i$ and $i \neq j$. The advantage of the Theil-Sen's slope estimator over the ordinary least squares estimator is that the former is less sensitive to outliers. More detailed information about its characteristics are available in Sen (1968). The Theil-Sen's estimator has been applied by using the Python package provided by Hussain and Mahmud (2019).

3.2.2. *Evaluation of the reliability with probabilistic metrics*

Given the probabilistic nature of DCPs, the assessment of the accuracy of a DPS can be complemented by a quantification of the uncertainty of the predictions. In the framework described in Goddard et al. (2013), probabilistic metrics are used to test if the ensemble spread of the DPS is adequate to represent the uncertainty of individual predictions. In other words, the purpose of the probabilistic metrics is to answer the question of whether the actual climate can be interpreted as one realization among all possible realizations of the DPS. If the answer is affirmative, these predictions are *reliable*, and the ensemble spread can be used to quantify the true range of possibilities associated to an individual prediction.

According to Goddard et al. (2013), the probabilistic quality of the predictions, with a given verification dataset as reference, is measured by the continuous ranked probability skill score (CRPSS):

$$\mathrm{CRPSS} = 1 - \frac{\sum_{j=1}^{N_\mathrm{d}} \mathrm{CRPS}_{Y,j}}{\sum_{j=1}^{N_\mathrm{d}} \mathrm{CRPS}_{X,j}} \qquad [3.29]$$

where $\mathrm{CRPS}_{Y,j}$ and $\mathrm{CRPS}_{X,j}$ are the continuous ranked probability scores for prediction and verification distributions. The CRPS can be interpreted as the mean absolute error in the probabilistic space. It is given by

$$\mathrm{CRPS}(\{\hat{Y}_j\}, \hat{X}_j) = -\int_{-\infty}^{\infty} [\mathcal{G}_{\{\hat{Y}_j\}}(y) - \mathcal{H}(y - \hat{X}_j)]^2 dy \qquad [3.30]$$

where $\mathcal{G}_{\{\hat{Y}_j\}}$ and $\mathcal{H}$ are the cumulative distribution functions (CDFs) of the predictions and the verification datasets (as the Heaviside function), respectively. As shown by Gneiting and Raftery (2007), assuming that the distribution of predictions is Gaussian with mean $\{\hat{Y}\}_j$ (Eq. [3.6]) and variance σ^2, Eq. [3.30] turns into

$$\mathrm{CRPS}(\mathcal{N}(\{\hat{Y}\}_j, \sigma^2), \hat{X}_j) = \\ = \sigma \left[\frac{1}{\sqrt{\pi}} - 2\varphi\left(\frac{\hat{X}_j - \{\hat{Y}\}_j}{\sigma}\right) - \left(\frac{\hat{X}_j - \{\hat{Y}\}_j}{\sigma}\right)\left(2\phi\left(\frac{\hat{X}_j - \{\hat{Y}\}_j}{\sigma}\right) - 1\right)\right] \quad [3.31]$$

where φ and ϕ denote the Gaussian probability distribution function (PDF) and CDF, respectively. Following Goddard et al. (2013), the mean of the prediction and verification distributions in Eq. [3.29] is $\{\hat{Y}\}_j$, while the variance of the prediction distribution is the average ensemble variance, given by

$$\overline{\sigma_Y^2} = \frac{1}{\mathrm{N_d}} \sum_{j=1}^{N_\mathrm{d}} \sigma_{Y,j}^2 = \frac{1}{\mathrm{N_d}} \sum_{j=1}^{N_\mathrm{d}} \frac{1}{\mathrm{N_{ens}} - 1} \sum_{k=1}^{N_\mathrm{ens}} (\{\hat{Y}\}_j - \hat{Y}_{kj})^2 \quad [3.32]$$

and the variance of the verification distribution is the squared standard error, denoted as

$$\sigma_X^2 = \frac{\sum_{j=1}^{N_\mathrm{d}} (\{\hat{Y}\}_j - X_j)^2}{\mathrm{N_d} - 2} \quad [3.33]$$

The average ensemble spread is used instead of the ensemble spread of each initialized experiment because the sampling errors associated to the small ensemble size may affect to the estimation of such ensemble spread. The optimal result in Eq. [3.29] is CRPSS = 0. It is attained for $\overline{\sigma_Y^2} = \sigma_X^2$, when the prediction and verification distributions are equal and, therefore, the average ensemble spread is certainly adequate to quantify the prediction uncertainty (Goddard et al., 2013). There may be situations in which $\overline{\sigma_Y^2} \neq \sigma_X^2$, indicating that predictions are overdispersive or underdispersive, depending on the magnitude of $\overline{\sigma_Y^2}$ relative to σ_X^2. In accordance with Kadow et al. (2016), this dispersion can be measured by the logarithmic ensemble spread score (LESS):

$$\mathrm{LESS} = \ln \frac{\overline{\sigma_Y^2}}{\sigma_X^2}, \quad [3.34]$$

for which positive (negative) values denote overdispersive (underdispersive) predictions. A combination of CRPSS and LESS provides useful information about the magnitude and sign of the dispersion as well as their impacts on the probabilistic quality of the predictions.

When comparing the performance of test predictions with a reference dataset Z, the differences in terms of CRPSS are evaluated as follows:

$$\Delta\mathrm{CRPSS}_Z = |\mathrm{CRPSS}(Z, X)| - |\mathrm{CRPSS}(\{Y\}, X)| \;, \tag{3.35}$$

with positive (negative) values indicating a better (worse) performance for the test predictions.

In the case of LESS, the logarithmic ensemble spread skill score (LESSS; Kadow et al., 2016) can be defined from Eq. [3.14] as

$$\mathrm{LESSS}_Z = 1 - \frac{\mathrm{LESS}(\{Y\}, X)^2}{\mathrm{LESS}(Z, X)^2} \;, \tag{3.36}$$

with a similar interpretation to that for MSSS_Z in Eq. [3.15], but in terms of LESS instead of MSE in this case.

3.2.3. *Assessment of the statistical significance*

The statistical significance of the results obtained for the metrics presented above has been generally assessed by following the same approach described in Goddard et al. (2013). This approach consists of a non-parametric block bootstrapping with replacement (Wilks, 2006, in Section 5.3.4), applied to generate 5000 random samples. For each metric, these samples are used to create a PDF and, then, calculate its associated confidence intervals. The steps to follow are:

1) A set of N_d random start dates are selected with replacement by considering blocks of 5-year consecutive start dates to account for autocorrelation.
2) A random ensemble of N_ens members is selected with replacement for each start date composing this set.
3) The metrics used to evaluate the accuracy and reliability are calculated from this randomly generated collection of experiments and their pairs in the verification dataset.
4) The steps 1) to 3) are repeated by 5000 times to generate a PDF per metric.
5) The confidence intervals for the 90 % confidence level are calculated for each metric from its respective PDF. The result obtained for a given metric will be significantly different from λ at the 90 % confidence level if the confidence

interval does not contain λ. In a hypothesis testing-based formulation (Wilks, 2006, in Section 5.1.4), the result will be statistically significant if, for a metric M, p-value < 0.1 is satisfied in a two-tailed hypothesis test with null (alternative) hypothesis H_0: $M = \lambda$ (H_A: $M \neq \lambda$).

This bootstrapping has been applied in most part of Chapters 4 and 5 to assess the statistical significance of the different metrics. Specifically, this approach has been used for MSSS_Z, ACC, ΔACC_Z, CB, ΔCB_Z, RPC, CRPSS, ΔCRPSS_Z, LESS and LESSS_Z. In general, the results are statistically significant at the 90 % confidence level if p-value < 0.1 is satisfied for $\lambda = 0$. The only exception is the RPC metric, for which $\lambda = 1$.

The statistical significance of trends has been assessed with a modified version of the Mann-Kendall test proposed by Hamed and Rao (1998), which takes into account the autocorrelation in time series. On the other hand, the statistical significance of the difference between trends has been calculated over the trend of the difference between time series. This approach, as opposed to check whether confidence intervals separately calculated for each trend overlap, facilitates the identification of real trend differences between variables by reducing the common variability noise in both time series (Santer et al., 2000). A trend is statistically significant at the 90 % confidence level if p-value < 0.1 is satisfied for $\lambda = 0$. As for trend calculation, the statistical significance has been evaluated by using the Python package provided by Hussain and Mahmud (2019).

3.3. Drift correction and subensemble selection for CESM-DPLE

3.3.1. *Description of the mean drift correction*

In the contribution of the DCPP to CMIP6, Boer et al. (2016) suggested removing the lead time-dependent drift in DCPs prior to carrying out any analysis. The mean drift correction method (MDC; Boer et al., 2016; CLIVAR, 2011), which accounts for the mean or climatological drift (i.e., the unconditional bias), has been used here. This method has been applied to the CESM-DPLE experiments which provide the input information in DD simulations. Since RCMs inherit the biases in GCMs through the LBCs, their correction helps to reduce the biases in RCM outputs and to achieve a better representation of the regional climate (Bruyère et al., 2014). The MDC has been applied to all variables listed in Table 3.1. A multivariate correction, as opposed to correct only a subset of variables, is expected to minimize the artificial drift in DD input data while maintaining the physical coherence between the fields involved. The calculation of the model drift has been done for each grid point and pressure/soil level[5] with monthly full fields, since the drift emerges at this time scale (Paeth et al., 2019). It has been calculated along the period spanning from 1979-11 to 2018-10 with ERA-Interim as reference (see Section 2.2). The ERA-Interim reanalysis has been used here because it provides all needed variables at all soil and height levels with the appropriate time aggregation.

At lead time τ, the mean drift is given by

$$d_\tau^{\mathrm{MDC}} = \{\overline{Y'}\}_\tau - \overline{X'}_\tau \, , \tag{3.37}$$

where $\{\overline{Y'}\}_\tau$ is the ensemble mean calculated from Eq. [3.1] with members $k = 1, \ldots, N_{\mathrm{ens}}$ and $\overline{X'}_\tau$ is given by Eq. [3.2].

Since WRF is provided with 6-hourly full fields, the monthly drift has been linearly interpolated along time to get 6-hourly time series. Additionally, for the fields listed in Table 3.1 which are not supplied with a 6-hourly aggregation, 6-hourly time series have also been constructed by assigning to each time step the daily or monthly mean value of the original field. Finally, the corrected prediction for the ensemble member k and initial date j at lead time τ has been calculated by removing the 6-hourly series of drift from $Y'_{kj\tau}$:

$$Y'^{\mathrm{MDC}}_{kj\tau} = Y'_{kj\tau} - d_\tau^{\mathrm{MDC}} \tag{3.38}$$

This correction has been applied in a cross-validated manner (CLIVAR, 2011). In other words, the information used to reduce the drift in an experiment starting at a certain date does not include the information of that specific experiment in order to avoid an artificial enhancement of the predictive skill.

3.3.2. *CESM-DPLE subensemble selection*

After applying the drift correction, the 4-member CESM-DPLE subensemble used to produce the WRF-DPLE experiments has been selected. Since the ocean is the primary source of climate memory, this selection has been focused on the results obtained for SST in terms of the spatially averaged ACC (i.e., $\langle \mathrm{ACC} \rangle$) over the EURO-CORDEX

[5] The lead time-dependent drift has been calculated for the pressure levels of 1000, 975, 950, 925, 900, 850, 800, 750, 700, 650, 600, 550, 500, 450, 400, 350, 300, 250, 200, 150, 100, 70, 50, 30, 20 and 10 hPa. For soil variables, it has been calculated for the ERA-Interim soil levels, whose depths are 0–7, 17–28, 28–100 and 100–289 cm.

domain in Figure 3.1 for lead years 2–9. It has been calculated with ERSST5 as the verification dataset along the same drift-correction period. The result for this metric has been chosen as a decision factor because it is not influenced by the magnitude of the variable, so it is not affected by neither unconditional or conditional biases (Murphy, 1988) nor a low signal-to-noise ratio (Scaife and Smith, 2018), becoming a measure of the potential skill which can be achieved after addressing those issues (see Eq. [3.24]). It has been evaluated in lead years 2–9 because this lead time window removes the seasonal-to-interannual variability to fully focus on the decadal scale. The CESM-DPLE subensemble has been constructed with the two members showing the best performance in reproducing the observed SST variability (the "best" members), the member showing the worst performance (the "worst" member) and a member with an "intermediate" behaviour. Quotation marks are used here to stress the relative nature of these denotations, since these members may not reproduce the same performances for other fields or metrics. Moreover, since each member, generated by randomly perturbing the initial atmospheric conditions at the round-off level, constitutes a sample realization of the climate variability, the differences between individual performances are expected to be small (Rosa-Cánovas et al., 2023). Paeth et al. (2017) and Rosa-Cánovas et al. (2023) followed a similar approach to build a 3-member ensemble for their respective studies also in the context of the DCP. With this strategy, a representative subensemble of the whole CESM-DPLE available to be downscaled has been constructed here. Since these members with heterogeneous skill levels cover the whole range of possible individual performances in the CESM-DPLE 10-member subensemble available for DD, this selection may help to retain part of the spread of this 10-member subensemble, needed to quantify the uncertainty of the DCPs.

The results of this evaluation have been depicted in Figure 3.3. In addition to $\langle\text{ACC}\rangle$, the outcomes obtained for the spatially averaged RMSE and CRPSS ($\langle\text{RMSE}\rangle$ and $\langle\text{CRPSS}\rangle$, respectively) have also been included, both calculated with full fields. In the three panels, the results are shown in order of decreasing average accuracy or reliability (represented by crosses) from left to right. The CESM-DPLE 4-member subensemble has been built with the members 10, 2, 1 and 8 as the two "best", the "intermediate" and the "worst" members, respectively (Figure 3.3a). As announced above, the differences between individual members are very small. For example, the gap between $\langle\text{ACC}\rangle$ results obtained by members 10 and 8 is only about 0.05. For the same metric, the performance of the 4-member ensemble (ENS4) is slightly less skilful than that of the non-corrected and corrected 10-member ensembles ($\text{ENS10}_{\text{Raw}}$

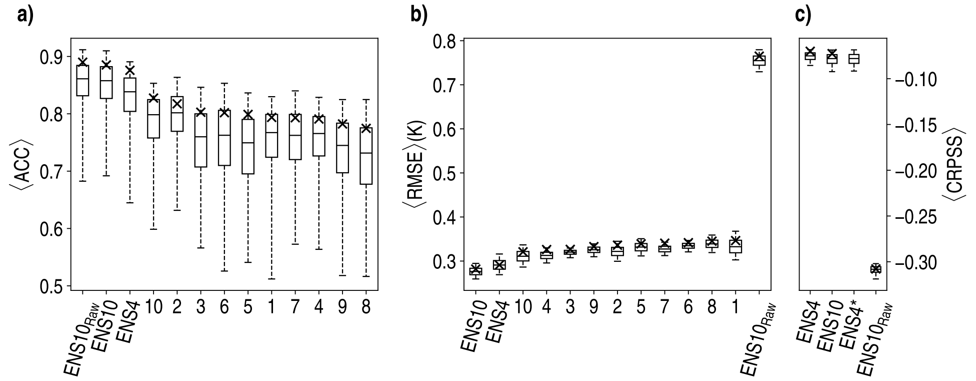

Figure 3.3: **a**) Spatially averaged ACC (i.e., ⟨ACC⟩) in the EURO-CORDEX domain. **b**) As **a**) but for ⟨RMSE⟩. **c**) As **a**) but for ⟨CRPSS⟩. The metrics have been calculated for individual members and several ensemble means ENSX, where X is the ensemble size. The label *Raw* in ENS10$_{\text{Raw}}$ indicates that the predictions have not been drift corrected, whereas the symbol "*" in ENS4* indicates that the members of ENS4 have been randomly chosen, as opposed to ENS4, whose members have been manually selected depending on their ⟨ACC⟩. Crosses denote the spatially averaged metric for a given member or ensemble mean. On the other hand, boxplots represent the results obtained for the non-parametric bootstrapping. Horizontal lines, boxes and whiskers identify the median values, 50 % and 90 % confidence intervals, respectively. The bootstrapping of CRPSS has been done without allowing member replacement. The results obtained for a given metric are ordered in terms of decreasing accuracy or reliability (crosses) from left to right.

and ENS10, respectively), as expected, although the differences are below 0.025. Since MDC only reduces the lead time-dependent mean drift, it hardly affects the results in terms of ⟨ACC⟩. The slight differences observed between ENS10 and ENS10$_{\text{Raw}}$ in Figure 3.3a are only due to the cross-validation approach considered in drift correction. They are mainly perceived in the width of the confidence intervals, which are narrower for ENS10. By contrast, the effects of drift correction are easily observed in Figure 3.3b, where ENS10$_{\text{Raw}}$ obtains the highest ⟨RMSE⟩ value by far. The differences between the result for ENS10$_{\text{Raw}}$ and the rest of ensemble and members are above 0.4 K. As for ⟨ACC⟩, the ⟨RMSE⟩ improves by increasing the ensemble size, although the difference between ENS10 and ENS4 is below 0.025 K. The results for ⟨CRPSS⟩ (Figure 3.3c) are also shown to examine how ENS4 performs relative to the 10-member ensembles and a randomly chosen 4-member subensemble (ENS4*). While ENS10$_{\text{Raw}}$ performs clearly worse than the other ensembles also in terms of reliability, the differences among the corrected experiments are very small. There is still a slight improvement of the reliability for the median spatial average of ENS4 over the median of ENS4*, showing than ENS4 performs better than at least the 50 % of the randomly chosen 4-member ensembles in terms of ⟨CRPSS⟩. However, the predictions of SST are not reliable on average in any of these cases, since ⟨CRPSS⟩ is significantly different from zero. Although Goddard et al. (2013) suggest correcting the CB together with the mean drift to properly estimate this metric because of the negative influence these biases have on the reliability, the CB has not been removed here because the purpose of this evaluation was to make the subensemble selection and to test the performance of the MDC method, which do not consider the correction of the CB.

3.4. Bias correction and subensemble selection for CESM-LE

3.4.1. *Description of the mean bias correction*

The uninitialized CESM-LE experiments have also been corrected before using them to generate the LBCs and ICs for the WRF-LE simulations. The procedure followed in this case is different from that applied to CESM-DPLE. It is based on the mean bias correction method followed by Bruyère et al. (2014) and Holland et al. (2010) to reduce the biases in the model mean annual cycle. As for MDC in Section 3.3, it has been applied over all variables listed in Table 3.1 at monthly scale for each grid point and pressure/soil level. The ERA-Interim dataset has been used as reference and the correction has been conducted along the period spanning from 1979-01 to 2005-12.

According to Bruyère et al. (2014) and Holland et al. (2010), the full field U', output from an uninitialized experiment, can be decomposed into

$$U' = U'_{\mathrm{AC}} + U_{\mathrm{P}} \,, \qquad [3.39]$$

where U'_{AC} is the average annual cycle and U_{P} is a perturbation term. In the same line, the decomposition is written as follows for the verification dataset X' used as reference:

$$X' = X'_{\mathrm{AC}} + X_{\mathrm{P}} \qquad [3.40]$$

After generating the 6-hourly time series following the same procedure described in Section 3.3, the bias-corrected full field U'_{BC} has been obtained by substituting U'_{AC} for X'_{AC} in Eq. [3.39]:

$$U'_{\mathrm{BC}} = X'_{\mathrm{AC}} + U_{\mathrm{P}} \,, \qquad [3.41]$$

3.4.2. *CESM-LE subensemble selection*

The 4-member CESM-LE subensemble used to drive WRF has been selected after applying the mean bias correction. A similar selection procedure to that described in Section 3.3 has been followed here, but with some differences. Firstly, the SST has been evaluated for ⟨RMSE⟩ instead of ⟨ACC⟩. Since the CESM-LE experiments are not initialized, they are not expected to reproduce the actual climate variability, as opposed to CESM-DPLE. Therefore, the accuracy in determining the magnitude of the verification full field has been evaluated in this case instead of the ability to capture the climate variability. Secondly, the evaluation has been focused on monthly full fields instead of the multiyear averages for lead years 2–9. Additionally, two evaluation periods have been considered here, the same used for the bias correction (from 1979-01 to 2005-12) and another, shorter, which spans only the period used in the evaluation of the experiments (from 1990-11 to 2005-10; see Section 3.2). The two "best", the "intermediate" and the "worst" members have been selected depending on the individual performances observed along both periods.

The results have been depicted in Figure 3.4 only for the selected members and several ensemble means. The results for CRPSS are not shown here because the uninitialized experiments are not of probabilistic nature, as DCP are. The members which fit the performance requirements in both periods are the members 5 and 9

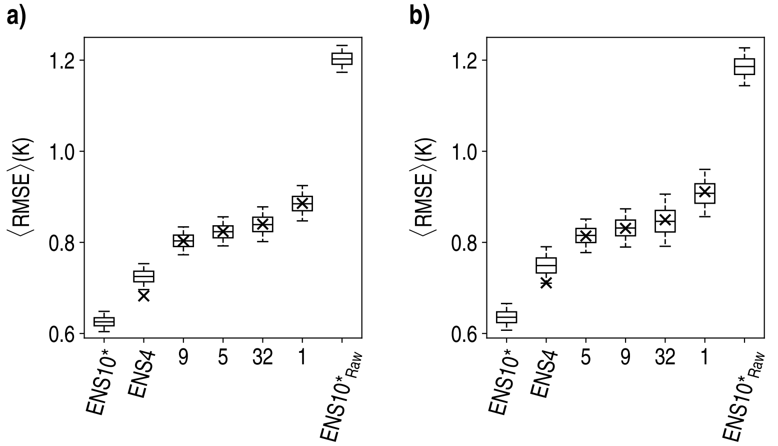

Figure 3.4: **a)** Spatially averaged RMSE (i.e., ⟨RMSE⟩) over the EURO-CORDEX domain in the period from 1979-01 to 2005-12. **b)** As **a)** but for the period from 1990-11 to 2005-10. The metrics have been calculated for individual members and several ensemble means ENS*X*, where *X* is the ensemble size. Only the results for the selected members of the 40-member CESM-LE subensemble are shown. The label *Raw* in ENS10*$_{\text{Raw}}$ indicates that the predictions have not been bias corrected, whereas the symbol "*" indicates that the members of ENS10 have been randomly chosen among the 40 available members. Crosses denote the spatially averaged metric for a given member or ensemble mean. On the other hand, boxplots represent the results obtained for the non-parametric bootstrapping. Horizontal lines, boxes and whiskers identify the median values, 50 % and 90 % confidence intervals, respectively. The results are ordered in terms of decreasing accuracy from left to right.

(the two "best" members), 32 (the "intermediate" member) and 1 (the "worst"). The "best" and "worst" members here are not those which strictly give the lowest and the highest ⟨RMSE⟩, respectively, in both periods since they do not coincide. The selection has been done by considering the members which generally adjust to these categories in both periods. In any case, the differences among the results obtained for individual members are not very pronounced, as occurred in Section 3.3. While there is a gap around 0.1 K between the values obtained for the "best" and "worst" members in both periods, with the "best" showing ⟨RMSE⟩ outcomes about 0.8 K. The results improve as the ensemble size increases, with values near 0.7 K for ENS4 and median values around 0.63 K for ENS10* (the 10-member randomly selected among the 42 available members). The highest errors have been found for the uncorrected ENS10*$_{\mathrm{Raw}}$ ensemble mean, for which values up to 1.2 K have been found.

3.5. Recalibration of the WRF-DPLE experiments

The approach followed to reduce the mean drift in the ICs and LBCs provided by CESM-DPLE for WRF simulations has been detailed in Section 3.3. However, although the MDC technique contributes to removing the unconditional bias, there are other types of bias which still hinder the predictive skill of the WRF-DPLE experiments, such as the existence of potential conditional biases and a misrepresentation of the dispersion of the ensemble members. These additional error sources in CESM-DPLE, together with the biases introduced by the RCM itself, have a negative impact on the predictive skill of the downscaled product.

Post-processing tunning is a common practice when it comes to reducing the biases in dynamically downscaled uninitialized experiments and improving the representation of the simulated climate fields (e.g., Gómez-Navarro et al., 2018; Teutschbein and Seibert, 2012). In the case of DCPs, the Decadal Climate Forecast Recalibration Strategy (DeFoReSt; Pasternack et al., 2018, 2021) was designed with the aim of reducing the unconditional, conditional and dispersion-related biases in dynamically downscaled DCPs. It has been used over decadal experiments produced in the framework of the MiKlip project, showing the ability to improve the predictive skill compared to the uncorrected product (Feldmann et al., 2019; Pasternack et al., 2018). Thus, the DeFoReSt approach has also been used here to correct the outputs from the WRF-DPLE simulations before conducting the analyses presented in Chapters 4, 5 and 7. The training dataset comprises the 4-member ensemble of downscaled experiments yearly initialized from 1970 to 1999 (the whole hindcast period). The AEMET dataset, after being linearly interpolated to the WRF-DPLE grid, has been used as reference in the recalibration process.

The fundamental principle of DeFoReSt is based on the fact that predictive skill can be improved by minimizing

$$|\overline{\mathrm{CRPS}}| = \left| \frac{1}{N_\mathrm{t}} \sum_{j=1}^{N_\mathrm{t}} \mathrm{CRPS}_j(Y', X') \right| , \qquad [3.42]$$

where $\mathrm{CRPS}_j(Y', X')$ can be decomposed as in Eq. [3.31] and N_t is the sample size of the training dataset. This metric has been chosen because its unique minimum $|\overline{\mathrm{CRPS}}| = 0$ denotes a perfect prediction in terms of probability, which is attained when the real climate is represented for the same probability distribution as the prediction.

The PDF of the recalibrated field $f^{\mathrm{Cal}}(Y', j, \tau)$ is assumed to be Gaussian with mean and variance being functions of the ensemble mean $\{Y'\}_{j\tau}$ and the ensemble variance $\sigma^2_{Y',j\tau}$ (note that full fields are used here), as well as the start date j and lead year τ:

$$f^{\mathrm{Cal}}(Y'; j, \tau) \sim \mathcal{N}\left(B(j,\tau) + C(j,\tau)\{Y'\}_{j\tau}, S(j,\tau)^2 \sigma^2_{Y',j\tau}\right) , \qquad [3.43]$$

with

$$\{Y'\}_{j\tau} = \frac{1}{N_\mathrm{ens}} \sum_{k=1}^{N_\mathrm{ens}} Y'_{kj\tau} \qquad [3.44]$$

and

$$\sigma^2_{Y',j\tau} = \frac{1}{N_\mathrm{ens} - 1} \sum_{k=1}^{N_\mathrm{ens}} (\{Y'\}_{j\tau} - Y'_{kj\tau})^2 \qquad [3.45]$$

In Eq. [3.43], the terms $B(j,\tau)$, $C(j,\tau)$ and $S(j,\tau)$ account for the unconditional bias, the conditional bias and the ensemble spread inflation, respectively. They are expressed as third- and second-order polynomials in terms of τ with the aim of addressing a potential non-linear dependence on the lead year τ. In addition, a linear dependence on j is included to also account for linear trends. These terms are written as follows:

$$\begin{aligned} B(j,\tau) &= \sum_{l=0}^{3}(b_{2l} + b_{(2l+1)}j)\tau^{l} \,, \\ C(j,\tau) &= \sum_{l=0}^{3}(c_{2l} + c_{(2l+1)}j)\tau^{l} \,, \\ S(j,\tau) &= \exp\left(\sum_{l=0}^{2}(s_{2l} + s_{(2l+1)}j)\tau^{l}\right) , \end{aligned} \qquad [3.46]$$

where the ensemble spread inflation $S(j,\tau)$ is constrained to be positive by using a exponential function.

From Eqs. [3.42], [3.43] and [3.46], the minimization function for the start date j and lead year τ can be written as

$$\begin{aligned} &\Gamma\left[\mathcal{N}\left(B(j,\tau) + C(j,\tau)\{Y'\}_{\tau}, S(j,\tau)^2\sigma^2_{Y',\tau}\right), X_{\tau}\right] = |\overline{\text{CRPS}}| = \\ &\left|\frac{1}{N_{\text{t}}}\sum_{q=1}^{N_{\text{t}}}\sqrt{S(j,\tau)^2\sigma^2_{Y',q\tau}}\left[\frac{1}{\sqrt{\pi}} - 2\varphi(\Omega_{q\tau}) - \Omega_{q\tau}[2\phi(\Omega_{q\tau}) - 1]\right]\right| , \end{aligned} \qquad [3.47]$$

with

$$\Omega_{q\tau} = \frac{X_{q\tau} - \left(B(j,\tau) + C(j,\tau)\{Y'\}_{q\tau}\right)}{\sqrt{S(j,\tau)^2\sigma^2_{Y',q\tau}}} \qquad [3.48]$$

as the standardized prediction error for the qth start date in the training dataset. The training dataset has been composed with a cross-validation approach, so the information used to correct the experiment initialized in the start date j excludes this specific start date to prevent an artificial enhancement of the predictive skill. Since the analysis of the downscaled DCPs is focused at both annual and seasonal scales, the recalibration is applied at annual scale as well as over each single season.

The minimization has been conducted with the Scipy implementation of the Nelder-Mead's algorithm (Nelder and Mead, 1965; Virtanen et al., 2020). The initial guesses of the u_l and c_l coefficients in Eq. [3.46], needed by the minimization algorithm, have been obtained from a linear regression between $X'_{q\tau}$ and $\{Y'\}_{q\tau}$, with $q = 1, ..., N_{\text{t}}$, for a given start date j and lead year τ:

$$X'_{q\tau} \sim B(j,\tau) + C(j\tau)\{Y'\}_{q\tau} \,, \qquad [3.49]$$

whereas the initial s_l coefficients have been set to zero to start the algorithm from a unity inflation, i.e., $S(j,\tau) = 1$.

The recalibrated prediction for the member k, start date j and lead year τ is expressed as

$$Y'^{\,\mathrm{Cal}}_{kj\tau} = B(j,\tau) + C(j,\tau)\{Y'\}_{j\tau} + S(j,\tau)(Y'_{kj\tau} - \{Y'\}_{j\tau}) \qquad [3.50]$$

The use of cross-validation in the adjustment of the conditional bias with a relatively small number of start dates may lead to high sampling errors, producing recalibrated predictions with lower skill than the original (raw) experiments (Goddard et al., 2013). A preliminary analysis of the recalibrated WRF-DPLE predictions confirmed that they were affected by this issue. Thus, an heuristic approach has been applied to filter the experiments for which recalibration does not deteriorate their raw predictive skill. For a given lead year τ, the collection of predictions initialized in the start dates $j = 1, ..., N_\mathrm{d}$ is recalibrated if one the following conditions are satisfied:

1) $\mathrm{MSSS}_{\mathrm{C,\,Cal}} > \mathrm{MSSS}_{\mathrm{C,\,Raw}}$ and $\Delta\mathrm{ACC}_{\mathrm{Raw}} = \mathrm{ACC}_{\mathrm{Cal}} - \mathrm{ACC}_{\mathrm{Raw}} > 0$;
2) $\mathrm{MSSS}_{\mathrm{C,\,Cal}} > \mathrm{MSSS}_{\mathrm{C,\,Raw}}$ and $\Delta\mathrm{ACC}_{\mathrm{Raw}} < 0$, but $\Delta\mathrm{CB}_{\mathrm{Raw}} = |\mathrm{CB}_{\mathrm{Raw}}| - |\mathrm{CB}_{\mathrm{Cal}}| > -\Delta\mathrm{ACC}_{\mathrm{Raw}}$.

Firstly, the predictions for the lead year τ are recalibrated as long as the adjustment provides an improvement in terms of MSSS_C and ACC (condition 1). The improvement of ACC is also considered because a gain in terms of MSSS_C may also be accompanied by a high loss in terms of ACC (note that the former depends on the quadratic form of the latter in Eq. [3.24]). However, there may also be a slight deterioration of ACC together with a high improvement in terms of CB and MSSS_C. In those cases, the recalibration is applied only if the gain for CB is higher than the loss for ACC (condition 2).

3.6. Regionalization of precipitation and temperature in the Iberian Peninsula

The main advantage of the high-resolution predictions generated in the context of this Thesis is their ability to reproduce fine-scale climate features which cannot be resolved by a GCM due to its coarser spatial resolution. The outputs of a RCM can be examined and depicted with a very high spatial detail, representing the effects that complex terrain, land cover distribution or dynamical processes occurring at low scale have on local climate. However, when it comes to graphically representing the temporal evolution of the RCM outputs, a reduction of the spatial dimensionality might be convenient as an alternative to work with the huge amount of grid points which compose the domain.

A multistep regionalization approach has been used to divide the IP into several homogeneous regions where the downscaled DCPs have been evaluated. For a given climate field, these regions group together those grid points depicting a similar climate variability. Therefore, the spatially averaged field over the points contained by each region can be interpreted as an overall representation of the evolution of that field in that specific region. The procedure followed here is based on the approach described by Argüeso et al. (2011). The regionalization has been done by concatenating three different techniques: the principal component analysis (PCA; Preisendorfer, 1988), a hierarchical or agglomerative cluster analysis and a non-hierarchical cluster analysis (Wilks, 2006, in Chapter 14).

The PCA has been used to extract the main spatio-temporal variability modes of a field. Let $F(t, x)$ be a field of anomalies or standardized anomalies over a set of locations $x = 1, ..., p$ at times $t = 1, ..., n$. The PCA decompose $F(t, x)$ to represent it in the form

$$F(t,x) = \sum_{m=1}^{p} r_m(t) e_j(x) \, , \tag{3.51}$$

where $r_m(t)$ are the projections of $F(t, x)$ onto the vector basis defined by $\vec{e}_m = [e_j(1), ..., e_j(p)]$ for the variability modes $m = 1, ..., p$. The time series constituted by $r_m(t)$ are the principal components (PCs) of $F(t, x)$, whereas the vectors $\vec{e}_m$ are its empirical orthogonal functions (EOFs). The EOFs, which are uncorrelated over space, are computed by maximizing the variance explained by the PCs, which are uncorrelated over time. If $F(t, x)$ is a field of anomalies, the EOFs are the eigenvectors of the covariance matrix of $F(t, x)$, but if $F(t, x)$ is standardized, the correlation matrix is used instead. In this Thesis, the covariance matrix has been used to conduct the PCA. The significant variability modes have been selected by using the North's rule of thumb (North et al., 1982). Then, the significant EOFs have been rotated with a varimax approach to facilitate their physical interpretation (Preisendorfer, 1988). The rotated EOFs have been obtained from this rotation and the projection of the field $F(t, x)$ onto them has given the rotated PCs.

Although the rotated EOFs provide useful information about the locations with similar climate variability, it is difficult to determine the regional boundaries from them. Thus, the cluster analysis techniques have been used with this purpose, taking the rotated EOFs as input information. Firstly, the agglomerative cluster analysis has been used to merge the points in rotated EOFs based on the squared Euclidean distance with the Ward's method (Wilks, 2006, in Section 14.2). The ideal result of the agglomerative clustering is to obtain a distribution of clusters which maximizes the similarity between points in the same cluster and the differences between clusters. A drawback of this technique is that it does not allow the exchange of points between clusters once they has been merged, leading to not optimal results if points are misplaced at first (Argüeso et al., 2011). However, in combination with the pseudo-F test (Calinski and Harabasz, 1974), it allows to compute the most optimal number of clusters. This information, along with the centroids calculated by averaging the rotated EOFs over those optimal clusters, has been used as input for a non-hierarchical k-means algorithm (Wilks, 2006, in Section 14.3.1), which relocates the points and obtains the most optimal distribution of clusters (Argüeso et al., 2011).

The multistep regionalization has been applied to the seasonal time series of the AEMET PR and NSAT variables (see Section 2.3) from 1970-12 to 2009-11 (the hindcast control period at seasonal scale). Previously, the time series have been detrended to prevent the climate change signal in the NSAT variables from accumulating most part of the explained variance in the main variability mode obtained from the PCA. For comparative reasons, the approach has been applied to T_{max}, T_{min} and T_{mean} together so that the same distribution of regions is shared by the three variables. A total of 8 regions has been obtained for PR (Figure 3.5a). These are the northwest (NW), central north (CN), northeast (NE), east (EA), eastern interior (EI), western interior (WI), southwest (SW) and central south (CS) regions. For NSAT, also 8 regions have been obtained (Figure 3.5b). They has been denoted as the north (NO), high mountain (MT), western interior (WI), northeast (NE), central interior (CI), east (EA), southwest (SW) and central south (CS). Note that some regions in PR and NSAT distributions share the same nomenclature only for descriptive reasons, but they do not necessarily cover the same area.

3.7. Spin-up time and soil initialization

The spin-up time is an important concept in the framework of DD, related to the ability of the RCM to represent the physical processes at the beginning of a simulation. It is defined as the time needed by the RCM to generate a dynamical equilibrium between its internal physics and the LBCs provided by the GCM. During this time, the RCM outputs are affected by biases which undermine the model ability to appropriately reproduce the physical processes involved in the evolution of climate (Giorgi, 2019; Giorgi and Mearns, 1999). While the atmospheric fields may often need lengths

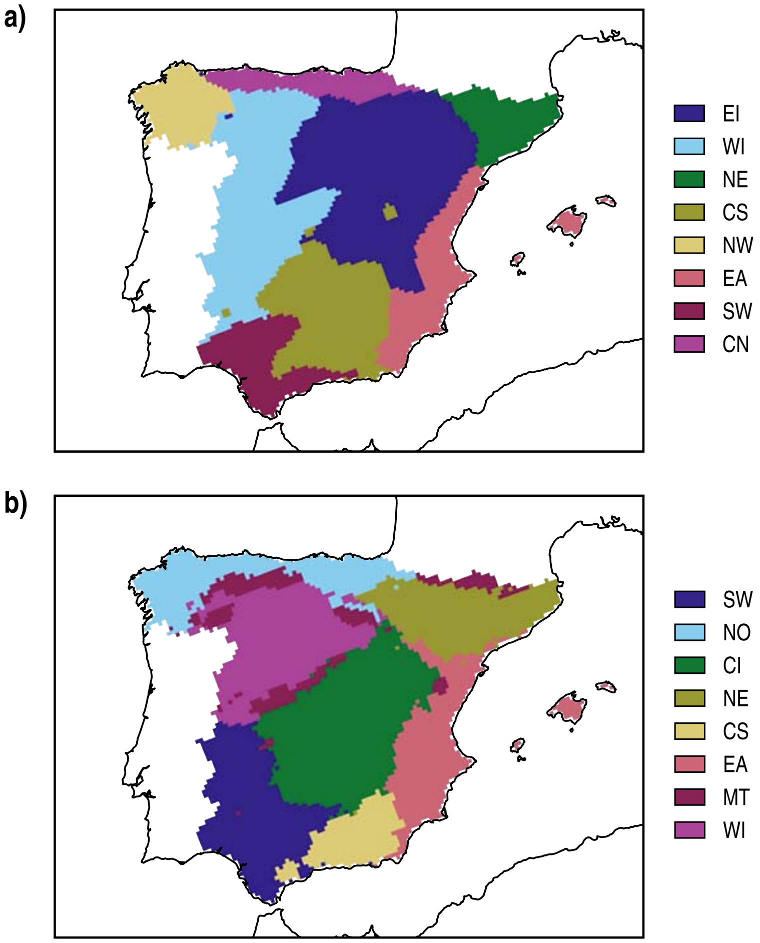

Figure 3.5: **a**) Regionalization of the Iberian Peninsula for the AEMET seasonal PR. **b**) As **a**) but for T_{max}, T_{min} and T_{mean} together. Labels indicate the nomenclature used to identify each region. The meaning of each label is detailed in the main text.

spanning few days to weeks (Gómez and Miguez-Macho, 2017; Jerez et al., 2020), the larger response of the soil variables may lead to much longer spin-up periods, spanning even several years (Khodayar et al., 2015).

Both WRF-DPLE and WRF-LE experiments are potentially affected by the spin-up issue. Since the lengths of individual simulations are very short (10 years for WRF-DPLE and 15 years for WRF-LE), the simulation period available for analysis would be severely reduced if the first simulated years were simply discarded in order to get a fully equilibrated RCM. In this line, Khodayar et al. (2015) estimate that a spin-up time of 80 months (~ 6.7 years) may be required in the IP for soil moisture on average (this topic is comprehensively addressed in Chapter 6). Therefore, no spin-up time has been considered in WRF-DPLE experiments. This may affect the predictive skill of fields at the early lead times of the simulation for each decadal experiment. Since the WRF-LE experiments are constituted by a single run per member, if no spin-up time were considered in WRF-LE either, the WRF-DPLE experiments would be compared with WRF-LE experiments without a constant level of skill during the first WRF-LE simulated years, leading to an uneven evaluation of the WRF-DPLE predictive skill compared to WRF-LE. For example, while the lead year 1 in a WRF-DPLE experiment started in 1999 suffers from the spin-up issue, its pair in the WRF-LE ensemble (which was started in 1990) does not.

To prevent this situation, the WRF-LE experiments started from an already dynamically equilibrated soil state, provided by a control simulation. Proceeding in this way, since the initial soil fields are already consistent with the RCM soil physics at the beginning of the simulation, the spin-up time is expected to be reduced. This strategy constitutes the default initialization method for DD experiments in the framework of MiKlip (Kothe et al., 2016). Although this conservative approach may favour WRF-LE over WRF-DPLE at the early lead times, it has been considered as a better alternative to the uneven evaluation derived from not taking any spin-up time.

The WRF-LE initial soil state in 1990-11 has been taken from the outputs of a WRF control simulation driven by ERA-Interim. This simulation has been started in 1982-01, so that the initial soil state is in dynamical equilibrium with the RCM physics after a 8-year and 10 months spin-up period in 1990-01. Additionally, the results obtained in Chapter 6 from the analysis of the spin-up time needed by WRF in the IP have led to also consider a dynamically equilibrated soil state in the simulations of the DCPs for the decade 2015–2025, presented in Chapter 7, which are initialized in 2015-11. The soil state is defined by the following fields in `wrfinput_d0x` files (see Section 3.1.3): soil temperature profile, soil temperature at lower boundary, skin temperature, soil moisture profile, unfrozen soil moisture profile, relative soil moisture, snow water equivalent, physical snow depth, snow coverage and canopy water content.

3.8. Description of the collection of the dynamically downscaled experiments

Part of the simulations conducted in the context of this Thesis has been carried out in the Tirant and Picasso high-performance computing nodes of the Spanish Supercomputing Network. The other part has been run in Alhambra, the supercomputer of the University of Granada. All simulations have been highly parallelized by using 112 cores. The collection of DD WRF experiments encompasses the following simulations:

- **Control simulation with WRF driven by ERA-Interim.**
  This simulation has been produced for reference purposes in the analysis and production of other experiments. It encompasses the period from 1982-01 to 2019-08. A total of 38.67 years has been simulated for this experiment.

- **Dynamically downscaled hindcasts produced with WRF driven by CESM-DPLE (WRF-DPLE) in the control period.**
  The drift-corrected CESM-DPLE 4-member subensemble, selected in Section 3.3, was dynamically downscaled. This set of experiments encompasses the hindcasts initialized every year from 1970 to 1999 (30 start dates). Each drift-corrected hindcast is composed of 121 months. The outputs of these simulations have been recalibrated by appliying the DeFoReSt approach with an additional filter, as described in Section 3.5. A total of 1210 years was simulated to generate these experiments. The results are analyzed in Chapters 4 and 5.

- **Dynamically downscaled uninitialized experiments produced with WRF driven by CESM-LE (WRF-LE) in the control period.**
  The bias-corrected CESM-LE 4-member subensemble, selected in Section 3.4, was dynamically downscaled. In this case, the available data only cover the period 1990–2005 (15 years). The DD simulations began in 1990-11 and ended in 2005-12, so each experiment spans 182 months. A total of 60.67 years has been simulated to generate the WRF-LE ensemble. The initial soil state was in dynamical equilibrium with the RCM soil physics. It was taken from the control simulation conducted with WRF and ERA-Interim, as detailed in Section 3.7. The results have been used to evaluate the added value to predictive skill of WRF-DPLE over WRF-LE in Chapters 4 and 5.

- **Experiments to test the sensitivity of WRF simulations to extreme initial conditions of soil moisture.**
  This set of experiments has been fully described in Chapter 6. It comprises a series of WRF simulations driven with ERA-Interim, initialized with different soil types in terms of soil moisture (wet, dry and very dry) for two initialization dates (January and July). Each of these 6 simulations spans 10 years along the period 1990–2000. A total of 60 years was simulated to generate this set of experiments.

- **Dynamically downscaled DCPs produced with WRF driven by CESM-DPLE (WRF-DPLE) for the decade 2015–2025.**
  The full drift-corrected CESM-DPLE 10-member subensemble was dynamically downscaled. This set of experiments encompasses the DCPs initialized in 2015, spanning the last of the decades available from CESM-DPLE. As for WRF-LE, the initial soil state was in dynamical equilibrium with the RCM soil physics. It was taken from the control simulation conducted with WRF and ERA-Interim (see Section 3.7). Each drift-corrected prediction is composed of 121 months. The outputs from these simulations have been recalibrated by appliying the DeFoReSt approach with an additional filter, as described in Section 3.5. A total of 100.83 years has been simulated to generate these experiments. The results are analyzed in Chapter 7.

A total of 1470 years, approximately, was simulated to conduct the DD simulations analyzed in this Thesis. In terms or computing time, 4.94 million CPU hours were needed to produce these experiments.

3.9. A note on the software used in this Thesis

The software tools used to conduct the DD experiments, work with climate information and write this Thesis are provided under different types of free software licenses. Sometimes, these products are developed and maintained by a small and independent group of people or even by a single person. In other cases, they are developed at universities or research institutes and/or endorsed by governmental institutions or private entities. A common characteristic shared by these tools is the non-profit work and support provided by communities and individuals who contribute to the projects in multiple forms, such as code development and revision, bug fixing, snippet and solution sharing on online forums, funding, etc. This Section is devoted to acknowledge the crucial role, often unrecognized, that free software plays in the scientific development.

As mentioned in Section 3.1, the DD simulations have been conducted with WRF (version 3.9.1.1; Skamarock et al., 2008; Wang et al., 2008). Part of the data preprocessing prior to conducting the simulations has been done with WPS (version 3.9.1; Wang et al., 2008). Multiple packages are needed to compile and run the RCM, with NetCDF (Rew and Davis, 1990), zlib (Gailly and Adler, 1995), libpng (Schalnat et al., 1995), HDF5 (The Board of Trustees of the University of Illinois and The HDF Group, 1998), JasPer (University of British Columbia et al., 1999) and Open MPI (The

Open MPI Team, 2004) among the most important ones. The RCM has been run in high-performance computing environments which use Slurm (Yoo et al., 2003) as the job scheduler.

The calculations required for the analysis and statistical treatment of the experiments have been done by directly or indirectly using different versions of Numpy (Harris et al., 2020), Scipy (Virtanen et al., 2020), Scikit-learn (Pedregosa et al., 2011) and Xarray (Hoyer and Hamman, 2017), a group of scientific computing-oriented Python packages. Climate Data Operators (CDO; Schulzweida, 2023), a set of tools for climate research developed at the Max-Plank-Institute for Meteorology, and the NetCDF Operators (NCO; Zender, 2008) have also been very useful for this purpose. Additionally, the scripts provided by Monaghan et al. (2014) have helped to write the WRF input data in the WRF intermediate format.

Most part of the graphical representations have been done by using Matplotlib (Hunter, 2007) and Cartopy (Met Office, 2010), two Python packages for data visualization and cartography, respectively. Moreover, Inkscape (The Inkscape Team, 2003) has been used for vector graphics design and manipulation. The colour palettes considered in graphical representations have been carefully chosen, to the extent possible, to be colorblind-friendly. They have been retrieved from Crameri (2023), Gomis et al. (2018) and "Qualitative colour schemes" (n.d.).

Finally, this document has been written in LaTeX(Mittelbach and Schöpf, 1989), making use of a wide variety of packages available in the CTAN repository (https://ctan.org/).

4

Retrospective decadal climate predictions for precipitation

This Chapter is devoted to analyze the WRF-DPLE predictive skill for PR. On the course of the following lines, several aspects related to the accuracy and reliability of the retrospective predictions or hindcasts are going to be addressed by means of the methodology described in Chapter 3. Firstly, the predictive skill has been examined by just comparing the results of the hindcasts with the observational datasets of AEMET. Secondly, this performance has been compared to that achieved by the global counterpart product, the CESM-DPLE. Afterwards, a similar approach has been followed to compare with the WRF-LE uninitialized simulations. Finally, the predictive skill of the regionally averaged hindcasts has been explored.

4.1. Predictive skill of the WRF-DPLE ensemble

The following results have been obtained from the analysis of the WRF-DPLE ensemble after being recalibrated by applying the DeFoReSt approach (Pasternack et al., 2018, 2021) with an additional filtering to adjust the inherent biases present in DCPs. A detailed description of the technique is available in Section 3.5.

❦ *Accuracy analysis*

The spatial distributions of RMSE_R (Eq. [3.11]) and ACC (Eq. [3.13]) for the multiannual mean anomalies of PR have been depicted in Figure 4.1. The highest values of RMSE_R are shown in lead year 1. This situation will be also observed when the analysis is focused on the seasonal scale, regardless the variable. It is expected that the RMSE_R decreases as the length of the averaging window increases. In this line, these averages can be interpreted as smoothing filters which remove the intraannual

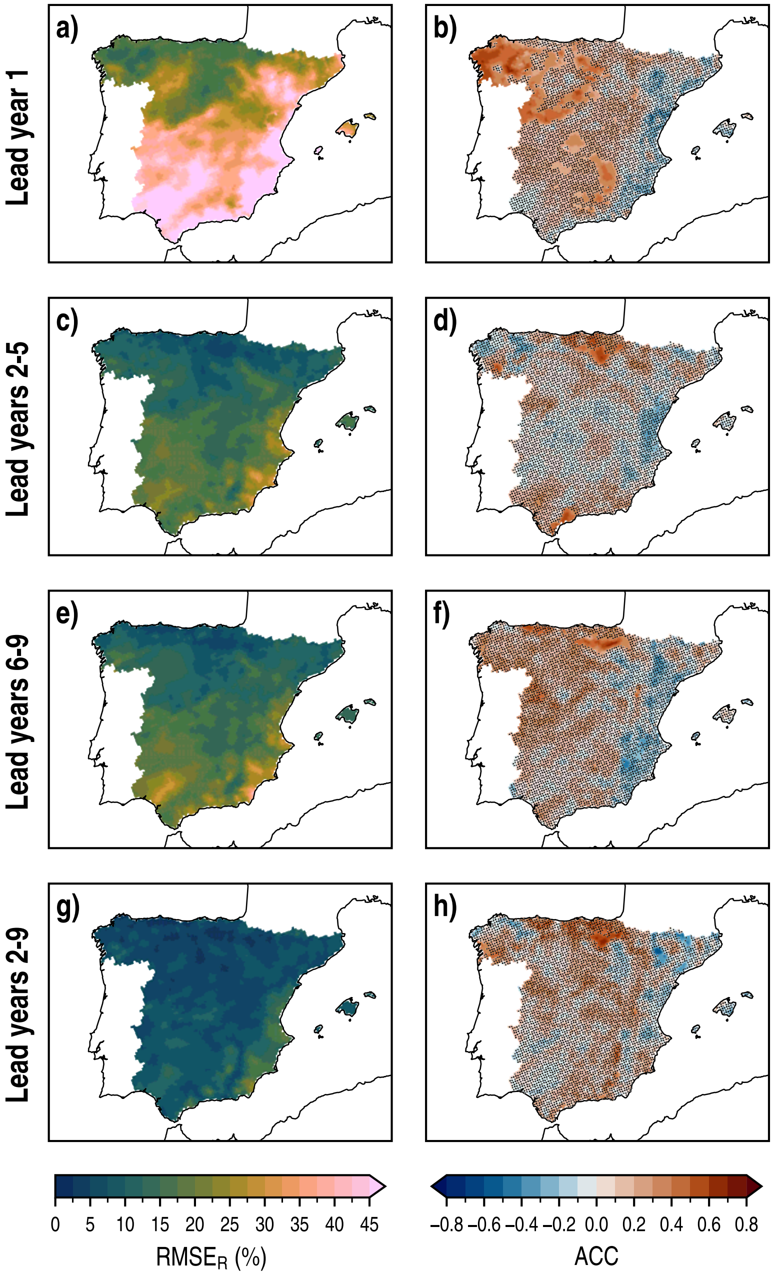

Figure 4.1: Spatial distributions of RMSE$_R$ (left column) and ACC (right column) for the WRF-DPLE multiannual mean anomalies of PR in lead years 1, 2–5, 6–9 and 2–9 (rows) at annual scale. In ACC maps, the absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

climate variability while leaving it untainted at the interannual or decadal scales.

In lead year 1, the largest $\mathrm{RMSE_R}$ values, which are above 45 %, have been found along the regions close to the Mediterranean coast and some southern locations over Sierra Morena. High $\mathrm{RMSE_R}$ values above 30 % are generally present in the southern half of the IP, the northwestern sector of the Northern Subplateau, the Ebro Valley and the Balearic Islands. On the other hand, the lowest errors have been found in the northwest and the eastern regions of the Northern Subplateau, where the minimum values are between 10 % and 15 %. Qualitatively similar spatial patterns have been obtained in lead years 2–5 and 6–9, having slightly higher errors at the latter, but with an important decrease of the errors in the southern half of the IP compared to the first year of the decade. In these cases, the largest $\mathrm{RMSE_R}$ values, around 35 % and 45 % are observed in the southeastern regions for lead years 2–5 and 6–9, respectively. Again, the lowest errors, with values below 10 % in some regions, have been found in the north. As mentioned above, the lowest $\mathrm{RMSE_R}$ values are shown in lead years 2–9, which do not surpass the maximum 25 % found in the southeast. On the other hand, the northern half of the IP shows values mainly below 7.5 %.

The most optimistic results in terms of ACC have been found in lead year 1, the period which gathers the most part of the statistically significant positive results (Figure 4.1, right column). In this case, they are found in the northwesternmost regions, the Northern Subplateau (excluding the central part) and some areas in the southern half of the domain. The largest ACC values are between 0.6 and 0.7 over the northwest of the domain. Positive but not significant values have been found over a very large fraction of the domain, excluding the negative ACC results observed mainly along the Mediterranean coast and Balearic Islands, although they are not statistically significant either. The eastern coast of the IP is not well represented in terms of ACC in any of the lead time periods considered. In lead years 2–5, the only statistically significant positive results have been found in the north, very close to the Strait of Gibraltar and in a small region in the northwest of the IP, with values around 0.5. The amount of locations with positive results has decreased compared to lead year 1, but only very small areas in the Mediterranean coast show a significant negative ACC with values around -0.3. The situation is slightly different in lead years 6–9. In this case, the significant, positive results are mainly concentrated in the north of the domain, with the maximum values between 0.6 and 0.7. Again, there are some significant and negative values about -0.5 over small regions in the east of the domain. In general, negative values are mainly confined to the eastern part of the domain, while positive (but not significant) values cover more than the half of the territory. A similar situation is observed in lead years 2–9, where the northern cluster of statistically significant correlations is maintained. In this case, the significant negative ACC values appear in the northeastern regions. Spurious significant positive results can be also found in very small areas across the IP.

At seasonal scale, the magnitude of the RMSE_R values strongly depends on the season (Figure 4.2). Note that the colormap used in these maps is in a binary logarithmic ($\log_2$) scale to account for the very extreme errors found in some cases without saturating too much the visual representations. The largest and lowest errors are observed always in lead years 1 and 2–9, respectively, because of the same smoothing effect of the multiannual averaging mentioned above. The differences between results in lead years 2–5 and 6–9 are hardly appreciable, although slightly smaller or higher errors may be observed in some regions depending on the lead time. For example, compare the central eastern regions in boreal winter (DJF) at both lead times or the southern regions in spring (MAM). In any case, there is not a robust influence of lead time on RMSE_R. From an interseasonal point of view, the highest RMSE_R values have been found in summer (JJA), followed by DJF, autumn (SON) and MAM, in decreasing order. RMSE_R values above 512 % cover large portions of the southern part of the IP in JJA for lead year 1. There is a cluster of very high errors which is maintained in the south across lead time, with maximum errors also above 512 % in lead years 2–5 and 6–9, while they do not surpass this value in lead years 2–9. The lowest errors are commonly observed in the northern part of the IP. Very high errors above 512 % have been also found in southern and northeastern locations in SON and over the south in MAM and SON for lead year 1. In contrast to JJA, these errors are largely reduced at the other lead times.

The results in terms of ACC at seasonal scale also vary depending on the season (Figure 4.3). Again, the ACC spatial distribution maps show not statistically significant results in general. The most promising results have been found in JJA, the season which accumulates more significant positive ACC values. The best model performance mainly occurs in JJA for lead years 1 and 2–5, although relatively similar outcomes are also observed at certain lead times in DJF and MAM. In JJA, strong correlations have been found in some regions mainly situated in the southern part of the IP for lead year 1, with maximum values around 0.6. Optimistic outcomes can be also observed in both the northernmost and southernmost areas of the domain in lead years 2–5. The worst results have been found in SON. Large areas of the domain are covered by negative ACC scores at all time scales, but mostly with not statistically significant values. The strongest negative correlations are shown for lead years 2–9 in

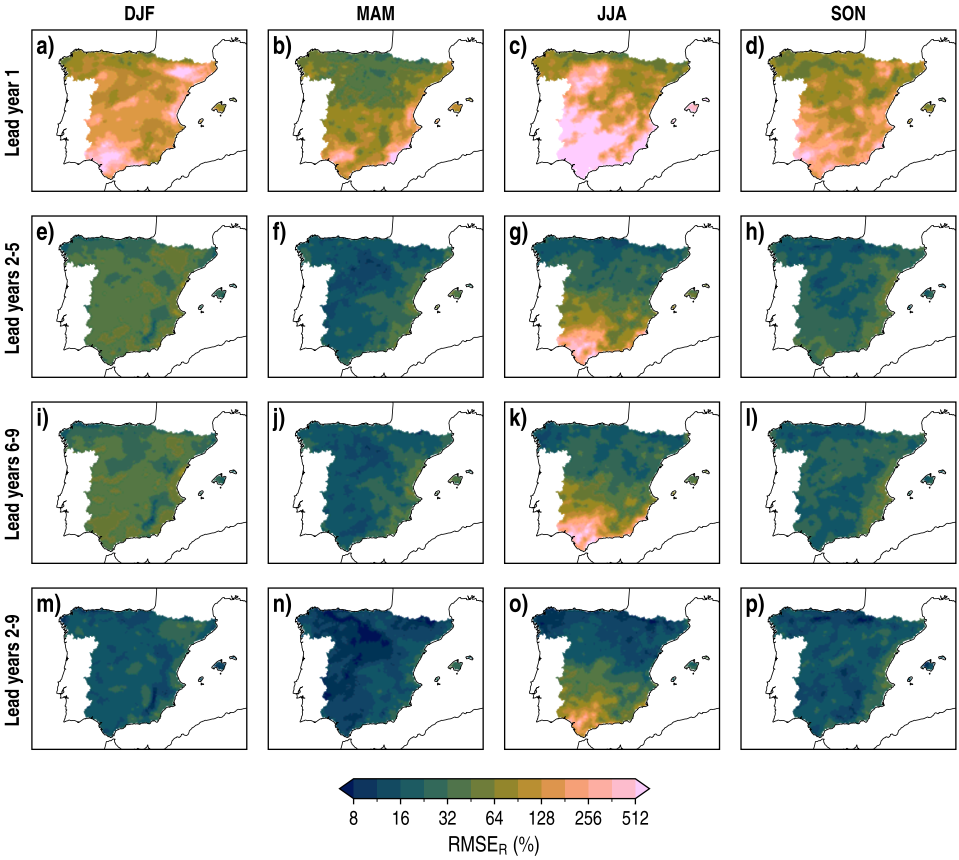

Figure 4.2: Spatial distributions of RMSE$_R$ for the WRF-DPLE multiannual mean anomalies of PR for lead years 1, 2–5, 6–9 and 2–9 (rows) in DJF, MAM, JJA and SON (columns). Note that the colormap used in these maps follows a binary logarithmic ($\log_2$) scale.

the central inner regions and northeast of the IP, with values down to -0.6.

Note that the very high errors obtained for the RMSE$_R$ do not necessarily imply a bad performance of the downscaled hindcasts. Since the anomaly error is divided by the observed full-field value (i.e., the observed value prior to subtracting the mean to compute the anomaly) in the calculation of the RMSE$_R$, the highest outcomes shown in Figures 4.1 and 4.2 may be caused by very low PR values occurring along the lead time series. Because of the smoothing effect of the lead time averages, the occurrence of very low full-field PR values is more probable in lead year 1 and in the seasonal analysis, consequently leading to higher errors. The examination of the averaged AEMET PR distributions along the period 1970-2009, depicted in Figure 4.4,

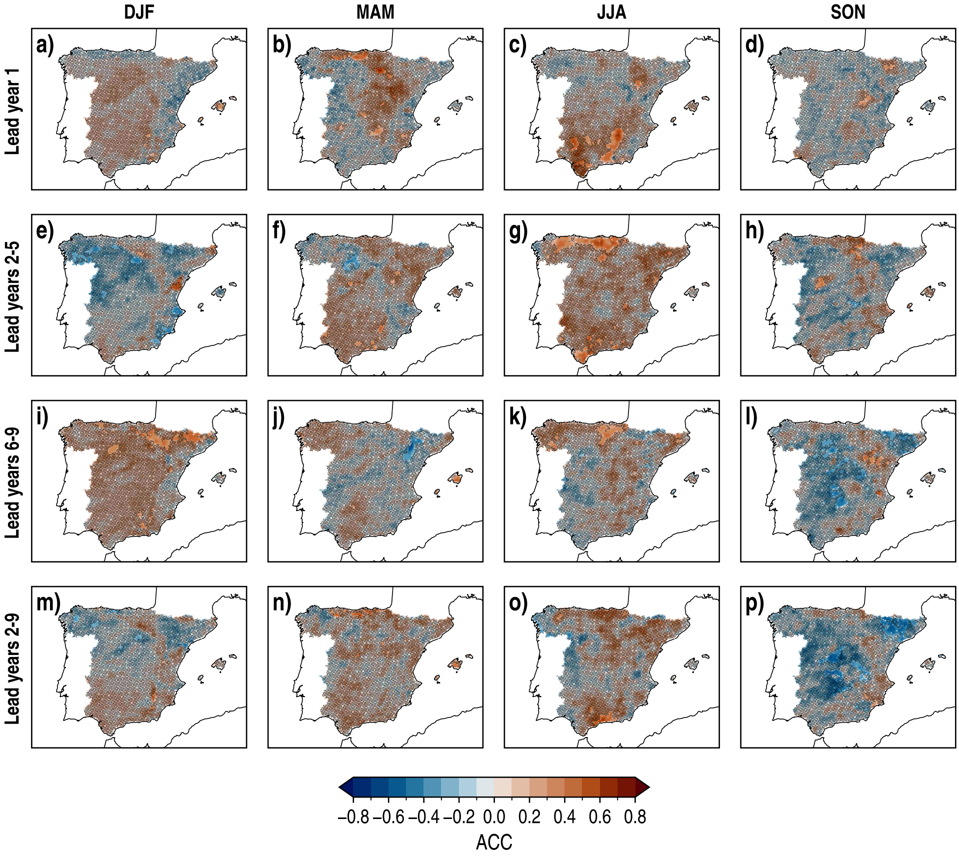

Figure 4.3: As Figure 4.2, but for ACC. The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

can shed some light on this situation. For instance, there is a direct link between the low full-field PR in the southern regions in JJA, with values below 10 mm/month, and the exacerbated errors observed there in that season. A more comprehensive view of the skill of WRF-DPLE to reproduce the magnitude of the PR anomaly will be provided later in the analysis of the MSSS_C.

On the other hand, the low ACC values could be partially explained by the small signal-to-noise ratio habitually present in DCPs, which is often revealed over the North Atlantic latitudes through the signal-to-noise paradox (see Section 3.2; Scaife and Smith, 2018; Smith et al., 2019, 2020). This paradox describes the counterintuitive phenomenon in which the model is better at predicting the real climate variability than at predicting its own modelled variability. It occurs when the ratio of predictable

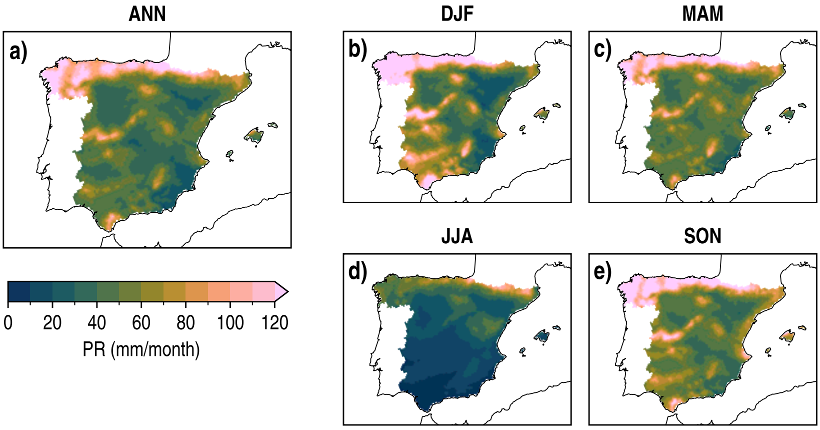

Figure 4.4: Spatial distributions of the time averages of the AEMET full-field PR at annual and seasonal scales for the period 1970-2009. While the annual series covers the period from 1970-11 to 2009-10, seasonal series span the period from 1970-12 to 2009-11.

variance in observations is higher than for the model, showing the existence of a small model signal-to-noise ratio and stressing the need of taking ensemble means of a large number of members to properly extract the climate signal. The PR is strongly affected by this phenomenon, as revealed by Smith et al. (2019) at global scale for a multimodel ensemble, as well as by Figure 4.5 for WRF-DPLE and CESM-DPLE in the IP. Figure 4.5 depicts the spatial distributions of the ratio of predictable components (RPC; see Eq. [3.27]) for PR in lead years 2–9, which shows that the ratio of predictable variance in observations can be up to 4 times higher than its value in model forecasts over large areas of the domain in both downscaled and global hindcasts. Despite the RPC of the global hindcasts shows a smoother spatial distribution, caused by its coarser resolution, it shows certain similarities with the pattern observed for the downscaled hindcasts. There are still some differences, such as a reduction of the RPC in the downscaled hindcasts along the Mediterranean coast or an increase over the Northern Subplateau. The presence of dispersed locations with statistically significant values below 1 is also more frequent in the downscaled hindcasts. These results are mainly caused by close-to-zero correlations between the ensemble mean and observations. Note that the RPC shown in Figure 4.5 is really an underestimation of the true value. An infinite ensemble size would be needed to totally suppressing the model background noise, properly extract the model signal and calculate the actual ACC$(\{Y\}, X)$ in Eq. [3.27], which is expected to be higher as the ensemble size increase (Scaife and Smith, 2018).

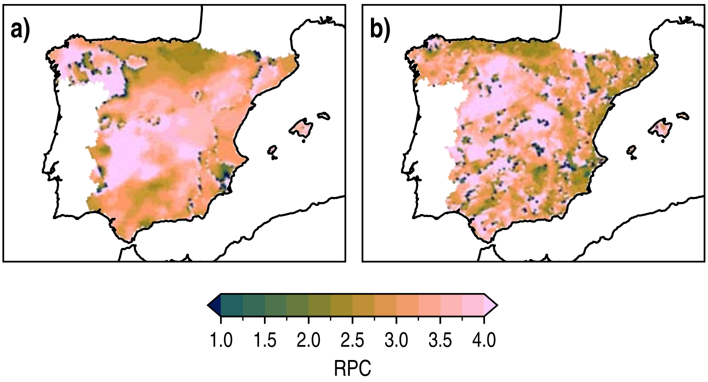

Figure 4.5: Signal-to-noise paradox in the hindcasts for PR over the IP. **a**) Spatial distribution of RPC for the multiannual mean anomalies of the CESM-DPLE PR in lead years 2–9 at annual scale. **b**) As **a**) but for WRF-DPLE. The absence (presence) of black dots indicates (not) statistically significant results different from 1 at the 90 % confidence level.

Reyers et al. (2019) addressed the impact of the ensemble size on the predictive skill for PR in their dynamically downscaled ensemble. They found that correlations could be improved in magnitudes within the range from approximately 0.1 to 0.3 when incrementing the ensemble size from 4 to 10 members for lead time series averaged over the IP. Depending on the initialization scheme of the GCM used to run the RCM, the ACC could range from 0.12 to 0.41 for a fixed ensemble size of 10 members. However, these results should be taken with caution because of the differences between their experimental design and that considered in this Thesis. Reyers et al. (2019) use different global and regional models and the resolution of their RCM is coarser than the one defined for WRF here. Also, they use 5 start dates with gaps of 10 years between consecutive dates to conduct their simulations, against the 30-year-consecutive start dates used in this Thesis. As Boer et al. (2016) suggest and Reyers et al. (2019) also support, the addition of more start dates would increase the robustness of their findings. Nevertheless, the results obtained by Reyers et al. (2019) can be used to qualitatively explain the low correlations observed in Figures 4.1 and 4.3. In addition, results in Figure 4.5 also suggest that there is a potential to improve the predictive skill in terms of ACC if a larger ensemble is used to remove the unpredictable background noise in the forecasts.

In the comparison with climatology, the accuracy provided by the WRF-DPLE hindcasts is certainly poor (Figure 4.6). The MSSS$_C$ scores are mostly negative, with many regions showing statistically significant values across all time scales. There are

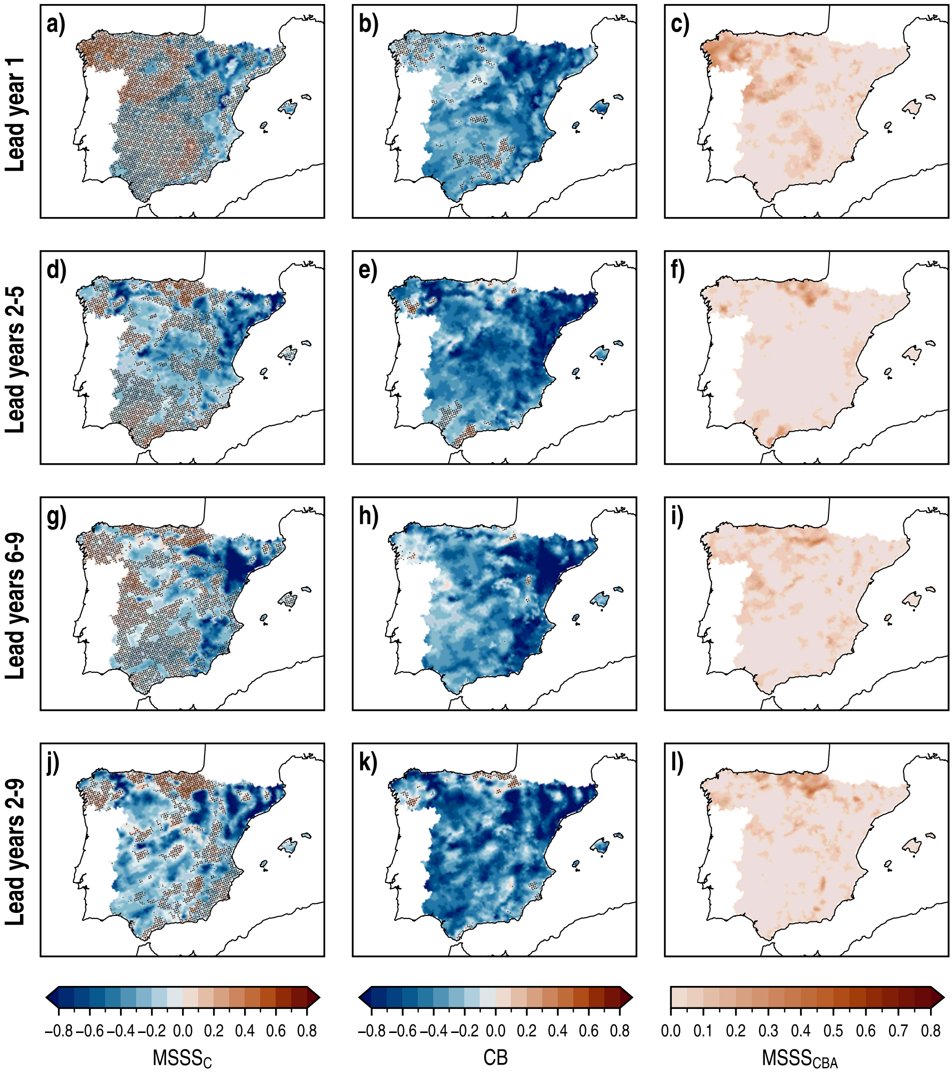

Figure 4.6 : Spatial distributions of $\mathrm{MSSS_C}$ (left column), with climatology as reference, CB (center column) and the same $\mathrm{MSSS_C}$ calculated for lead time series with an adjusted CB, i.e., equal to zero ($\mathrm{MSSS_{CBA}}$; right column), for the WRF-DPLE multiannual mean anomalies of PR in lead years 1, 2–5, 6–9 and 2–9 (rows) at annual scale. In $\mathrm{MSSS_C}$ and CB maps, the absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

some areas where $\mathrm{MSSS_C}$ is positive, matching closely the regions which showed significant positive results for ACC in Figure 4.1. This fact is not surprising because of the close relation between this two metrics (see Eq. [3.24]). However, the positive $\mathrm{MSSS_C}$ values are not significant in this case. The worst represented regions are in the northeast of the IP, which shows strong negative results at all lead times, with values even below -0.8 in some locations. These low scores are the consequence of two concurring factors: the low correlations observed in Figure 4.1, discussed above, and the large CB (in absolute value) depicted in Figure 4.6. The CB depends on ACC, as stated by Eq. [3.17], but also on the ratio $s_{\{Y\}}/s_X$ between the standard deviation of hindcast and observed lead time series. In locations where ACC is negative, CB is always negative. If ACC is positive, the ratio between standard deviations must be larger than ACC to get a negative CB, which is what generally happens. CB values are almost completely significantly different from zero, excepting over some locations which showed high-enough ACC values in Figure 4.1. These are mainly found in lead year 1.

MSSS_{CBA} represents the previous MSSS_{C} calculated for the lead time series after the conditional bias adjustment (CBA), that is, after removing it. This is equivalent to take CB = 0 in Eq. [3.24]. MSSS_{CBA} can be interpreted as the maximum MSSS_{C} which can be potentially achieved given the value of ACC. The MSSS_{CBA} is very low for almost the whole IP regardless of the lead time. The results are always statistically significant and positive because this metric only can be positive or zero. Only if ACC values in a given location are exactly equal to zero for all bootstrapping iterations, which is something almost impossible, MSSS_{CBA} would be not significant. The regions with the highest MSSS_{CBA} values match the regions with the highest ACC results in absolute value, which mostly correspond to significant positive correlations. In lead year 1, the highest scores around 0.3 are located in the northwest of the domain. For the rest of lead times, they are mainly placed in the northern regions with similar values. Even after completely removing the CB, the predictive skill would be very limited with climatology as the reference basis.

The results for the seasonal multiannual means lead to similar interpretations, with minor differences depending on the season being examined. They can be consulted in Figures B.1 to B.3 in Appendix B.1.

❦ *Reliability analysis*

In addition to the deterministic scores discussed above, which give a measure of the accuracy of the hindcasts, an analysis of the decadal predictive skill must also include an assessment of their reliability, which is examined by means of probabilistic metrics. The aim of this assessment is to address whether the WRF-DPLE average ensemble spread is adequate to represent the prediction uncertainty, and the CRPSS

(Eq. [3.29]) is the score selected to carry out this task. The spatial distributions of the CRPSS are depicted in Figure 4.7. As detailed in Section 3.2, the desired value for CRPSS is zero, which would mean that the path followed by observations could be considered as a possible path to be followed by a single member of the ensemble. In that case, the average ensemble spread $\overline{\sigma_Y^2}$ (Eq. [3.32]) could be used to determine the true range of possibilities for the predicted future climate.

The results shown in Figure 4.7 are certainly promising in this respect, since CRPSS values not significantly different from zero are widespread at all lead times, with the exception of lead year 1, when the southern half of the IP is almost completely covered by significant results. There are CRPSS values very close to zero in many locations mainly situated in the northern half of the IP, which extend southward in lead years 2–5 and 6–9, whereas almost cover the whole domain in lead years 2–9. These outcomes are strongly related to the results for LESS (Eq. [3.34]) depicted in the same Figure 4.7. Statistically significant results in LESS, both positive or negative, indicate robust discrepancies between the WRF-DPLE average ensemble variance $\overline{\sigma_Y^2}$ and the squared standard error σ_X^2 given by Eqs. [3.32] and [3.33], respectively. There is a clear case of underdispersive predictions over almost the whole domain in lead year 1, where $\overline{\sigma_Y^2} < \sigma_X^2$, with the lowest values observed in the southern regions. This underdispersion is attenuated in lead years 2–5 and 6–9, when even positive but not significant LESS values appear over some northern regions. The positive LESS results are predominant in lead years 2–9, but they are mostly not significant with the exception of a very small area in the Northern Subplateau. In these locations, the LESS is around 1.5, which means that $\overline{\sigma_Y^2} > \sigma_X^2$ by a factor of almost 4.5. The patterns shown by LESS at all lead times are very similar to those corresponding to CRPSS, as the regions with small absolute LESS values are the same that those showing small absolute CRPSS values. Not significant outcomes for LESS often lead to not significant CRPSS results, indicating that $\overline{\sigma_Y^2}$ is appropriate to quantify the uncertainty of the forecasts.

The same discussion can be extrapolated to reliability analysis with multiannual means at seasonal scale (Figures B.4 and B.5 in Appendix B.1). In this case, the most promising results have been obtained in MAM. All lead times show a domain mostly covered for not significant CRPSS values. They are a consequence of the not significant LESS also found in MAM. The LESS distributions show that the predictions are overdispersive in some regions, especially in lead years 2–9, but these results lack of statistical significance. The least optimistic outcomes have been found in DJF and SON, when the locations with hindcast reliability are scarcer than in MAM and JJA,

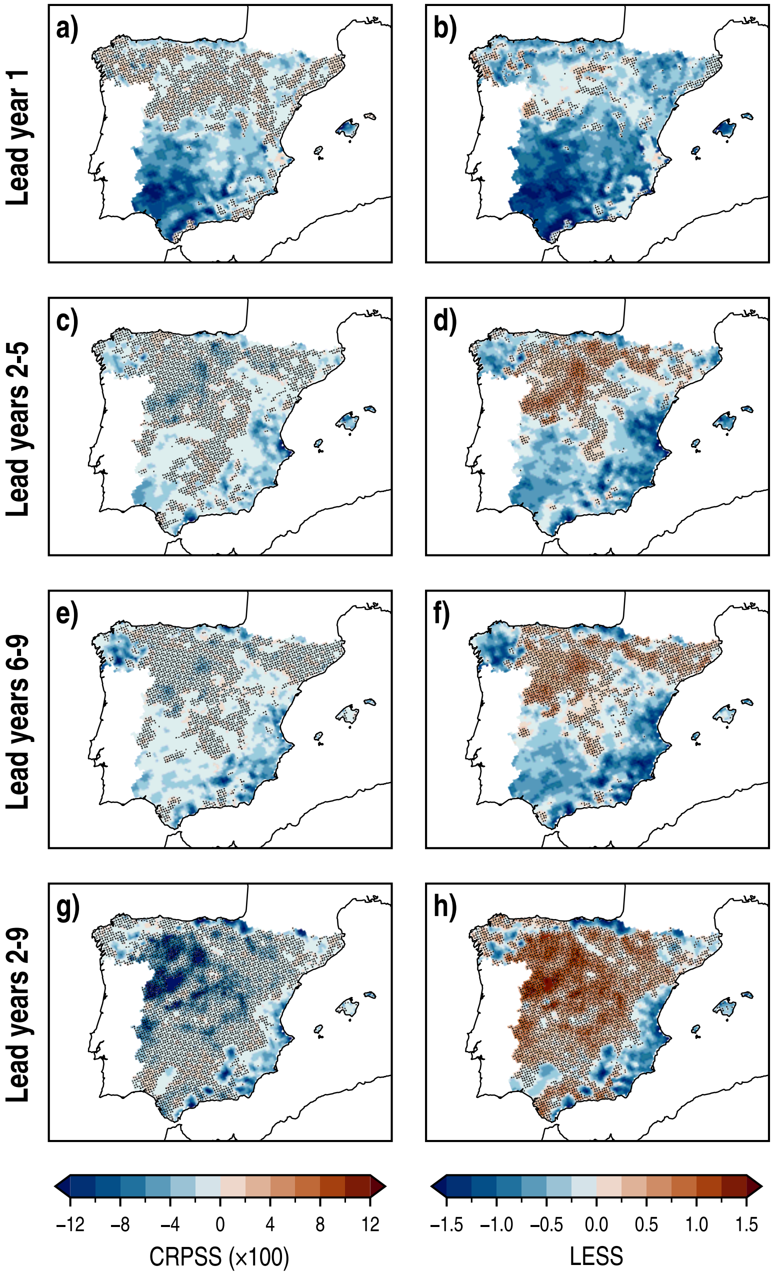

Figure 4.7: Spatial distributions of CRPSS (left column) and LESS (right column) for the WRF-DPLE multiannual mean anomalies of PR in lead years 1, 2–5, 6–9 and 2–9 (rows) at annual scale. The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

mainly because of the presence of robust underdispersion.

4.2. Comparison with the CESM-DPLE subensemble

After examining the predictive skill of the WRF-DPLE PR hindcasts, the comparison of the downscaling experiments with the CESM-DPLE subensemble, the global counterpart, is needed to quantify the value gained or lost by the downscaled ensemble. It is worth remarking that, as the downscaled hindcasts, the global hindcasts have been subjected to a recalibration process to remove their inherent biases. In this case, only the climate drift (unconditional bias) has been corrected, as done for the ICs and LBCs prior to conducting the DD simulations (see Section 3.3).

❦ *Accuracy analysis*

The results for the analysis of the WRF-DPLE multiannual mean anomalies of PR in terms of $\mathrm{MSSS_G}$ (Eq. [3.16]), $\Delta\mathrm{ACC_G}$ (Eq. [3.25]) and $\Delta\mathrm{CB_G}$ (Eq. [3.26]), with CESM-DPLE as reference are depicted in Figure 4.8. The statistically significant added value of the WRF-DPLE ensemble is restricted to very small regions regardless of the lead time. In lead year 1, some improvement over the global product is found along the Mediterranean coast and some inner and northern locations. However, there is almost no statistical significance in these outcomes, and the significant $\mathrm{MSSS_G}$ values are always below 0.3. Negative scores cover a large portion of the domain, with the lowest significant values mainly over inner and northeastern locations, getting minimum values between -0.6 and -0.5. In lead years 2–5, the general situation is slightly better than for lead year 1, in the sense that broader regions with positive and significant outcomes has been found. The largest cluster of significant positive results can be observed in the northwest of the domain, with values between 0.3 and 0.4 (Figure 4.8d). The significant negative values are mainly distributed in the northern half of the domain, with an important presence in the northeastern regions and minimum scores below -0.6. Although regions which show a negative performance of the downscaled hindcasts continue being present in lead years 6–9, the results in terms of $\mathrm{MSSS_G}$ improve compared to the previous lead times. In this case, a small area of $\mathrm{MSSS_G}$ values between 0.2 and 0.5 is found over some eastern locations. The spatial distribution of $\mathrm{MSSS_G}$ in lead years 2–9, with some significant and positive results in the north and northwest, continues being dominated by the negative scores.

Because of the relation shown in Eq. [3.16], the results for $\mathrm{MSSS_G}$ can be explained by the combination of the spatial distributions of both $\Delta\mathrm{ACC_G}$ and $\Delta\mathrm{CB_G}$, depicted

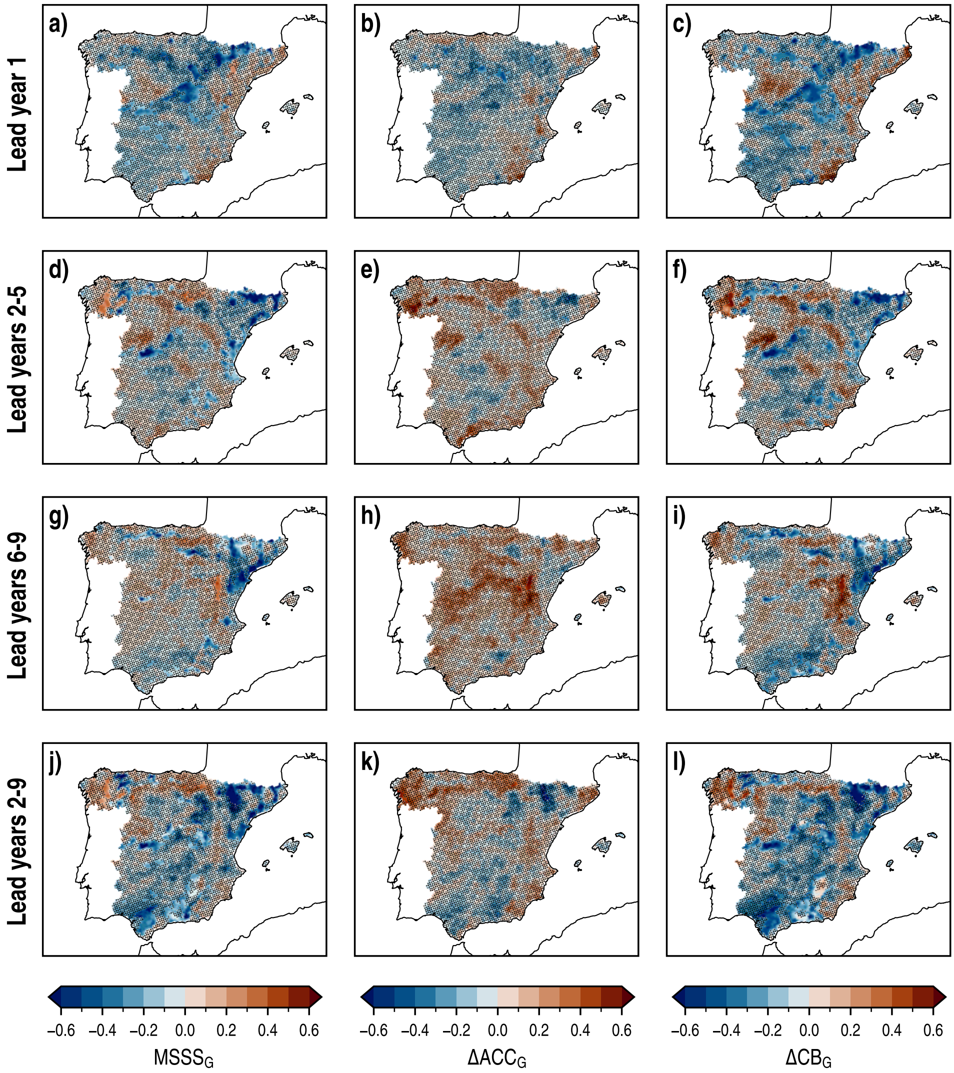

Figure 4.8 : Spatial distributions of MSSS$_G$ (left column), ΔACC$_G$ (center column) and ΔCB$_G$ (right column), with CESM-DPLE as reference, for the WRF-DPLE multiannual mean anomalies of PR in lead years 1, 2–5, 6–9 and 2–9 (rows) at annual scale. The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

in Figure 4.8. The patch of significant negative results for MSSS$_G$ in lead year 1 is the product of the confluence of a large loss in terms of CB for the downscaled experiments added to a slight decrease in correlation compared to the CESM-DPLE subensemble. In lead years 2–5, the results for ΔACC$_G$ show a modest improvement in favour of the WRF-DPLE ensemble compared to the first year of the decade, but with a consistent majority of not significant scores for both positive an negative results. In the northwest of the domain, the ΔACC_G values above 0.5, along with the important positive results for ΔCB_G, are responsible of the aforementioned significant positive $MSSS_G$ cluster spanning those regions. The same occurs with the negative performance observed for both metrics in the northeastern regions. Large areas of added value to ACC have been found in lead years 6–9, with almost the whole domain covered by positive but not significant differences. The best results in terms of both ΔACC_G and ΔCB_G, with values above 0.6, are depicted in the same area over the central east of the IP. The worst results for ΔCB_G are shown in lead years 2–9, with a large fraction of the domain covered by negative differences and patches of significant negative results scattered over the whole IP. The differences in correlation show the best results mainly concentrated in the northern and northwestern regions, with values between 0.2 and 0.6.

For their 10-member ensemble of downscaled hindcasts, Reyers et al. (2019) found moderate but positive added value to the predictive skill, in terms of $MSSS_G$, in the eastern regions of the IP (Figure 4 in Reyers et al., 2019). Added value in $MSSS_G$, ΔACC_G and ΔCB_G was also found when the lead time series spatially averaged over the IP are analyzed (Table 2 in Reyers et al., 2019). However, they also noticed the dependence the predictive skill has on the initialization scheme of the global model, which plays an important role in how well the PR is represented. As the authors showed, depending on the initialization scheme applied to the global model, this added value to the predictive skill over the global model for this region could turn into the opposite. Note that, as mentioned previously in Section 4.1, Reyers et al. (2019) only considers five start dates with gaps of 10 years between consecutive dates in their ensemble of hindcasts, also with different spatial resolution and climate models from those used in this Thesis, so these results should be taken with caution. As the authors suggest and Boer et al. (2016) affirm, the addition of more start dates would contribute to increasing the robustness of their results.

Some moderately positive outcomes can be extracted from the analysis at seasonal scale. The spatial distributions of seasonal $MSSS_G$ are shown in Figure 4.9. The best performance of the WRF-DPLE ensemble with the global product as reference can be observed in JJA. Excepting lead year 1, large areas of the domain are covered with results which show the added value of the downscaled hindcasts (although they are not always statistically significant), especially in lead years 2–9. At this lead time, a large area with significant positive $MSSS_G$ values is present in the central regions of

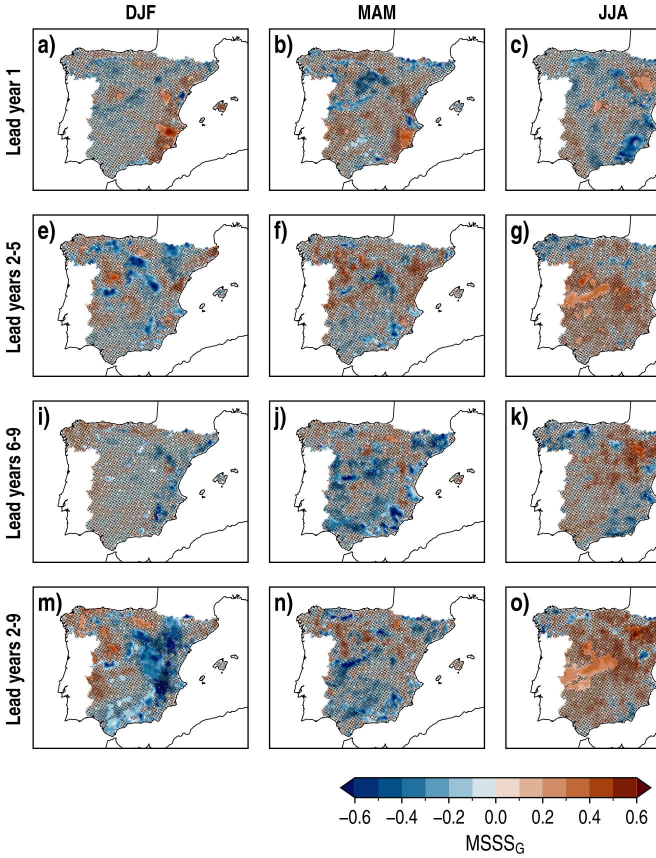

Figure 4.9: Spatial distributions of MSSS$_G$ for the WRF-DPLE multiannual mean anomalies of PR, with CESM-DPLE as reference, for lead years 1, 2–5, 6–9 and 2–9 (rows) in DJF, MAM, JJA and SON (columns). The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

the IP. These scores approximately range from 0.1 to 0.5 and are the consequence of the joint action of positive results in ΔACC$_G$ (Figure B.6, Appendix B.1), with values above 0.6 in some locations, and the significant positive results in ΔCB$_G$ (Figure B.7, Appendix B.1) spanning a wider area, with maximum values above 0.6. Very good results have also been found for MSSS$_G$ over a region located in the northeastern quarter of the IP, where some values surpass 0.6. Again, although ΔACC$_G$ shows moderately high (but not significant) positive results, the promising MSSS$_G$ outcomes are mainly motivated by the large improvement in terms of CB. Similar performances have been found, for example, for lead year 1 in DJF and MAM over some regions in the southeast or over the northeast in SON. Notwithstanding, there are still important losses in predictive skill (see, e.g., DJF in lead years 2–9 or SON in lead years 6–9).

❦ *Reliability analysis*

The comparison between the WRF-DPLE and CESM-DPLE hindcasts in terms of their reliability can be addressed by examining the Figure 4.10, which collects the results obtained for ΔCRPSS$_G$ and LESSS$_G$ (see Eqs. [3.35] and [3.36], respectively). At first glance, a notable absence of statistically significant results predominates in the ΔCRPSS$_G$ spatial distributions. Improvements for this score (but not significant) are generally observed at all lead times, excepting lead years 2–9. In lead years 1, 2–5 and 6–9, the highest positive ΔCRPSS$_G$ values are frequently observed over mountain regions, such as the Central System, Sierra Morena, the Baetic System or the Pyrenees. On the contrary, the smallest scores are mainly found in the Northern and Southern Subplateaus, the Ebro and Guadiana valleys, and some locations in the northwest and southeast. These regions are characterized by their flatness and/or low altitude (see Figure 1.4a). In lead years 2–9, the area covered by negative results is wider, with the lowest (but not significant) values situated over the Northern Subplateau.

As happened with CRPSS and LESS in Figure 4.7, the results in ΔCRPSS$_G$ are highly influenced by those obtained for LESSS$_G$. In this case, LESSS$_G$ shows the improvement or deterioration in the representation of $\overline{\sigma_Y^2}$ compared to σ_X^2. Those regions where WRF-DPLE outperforms CESM-DPLE in terms of LESS (significant positive LESSS$_G$), which are very common at all lead times with the exception of lead years 2–9, also show an added value in terms of CRPSS (although not significant). The presence of very extreme values in LESSS$_G$ maps (i.e., positive or negative values with large magnitude) may be noteworthy, but it is linked to the definition of the metric (see Eq. [3.36]). Since LESSS$_G$ depends on the quadratic form of the WRF-DPLE and CESM-DPLE LESS, differences between these metrics may lead to large values for LESSS$_G$, especially when small numbers are involved in the calculation.

The results for the seasonal multiannual time series can be consulted in Figures B.8 and B.9 (Appendix B.1). Although results vary depending on the season, the interpretation carried out at annual scale can be extrapolated to the seasonal scale. Curiously, the best improvements in terms of CRPSS are observed in DJF and SON, the seasons which generally depicted the worst CRPSS values in Figure B.4. This is caused by a large improvement in terms of LESS by the downscaled hindcasts over the GCM which is fundamentally observed in those seasons.

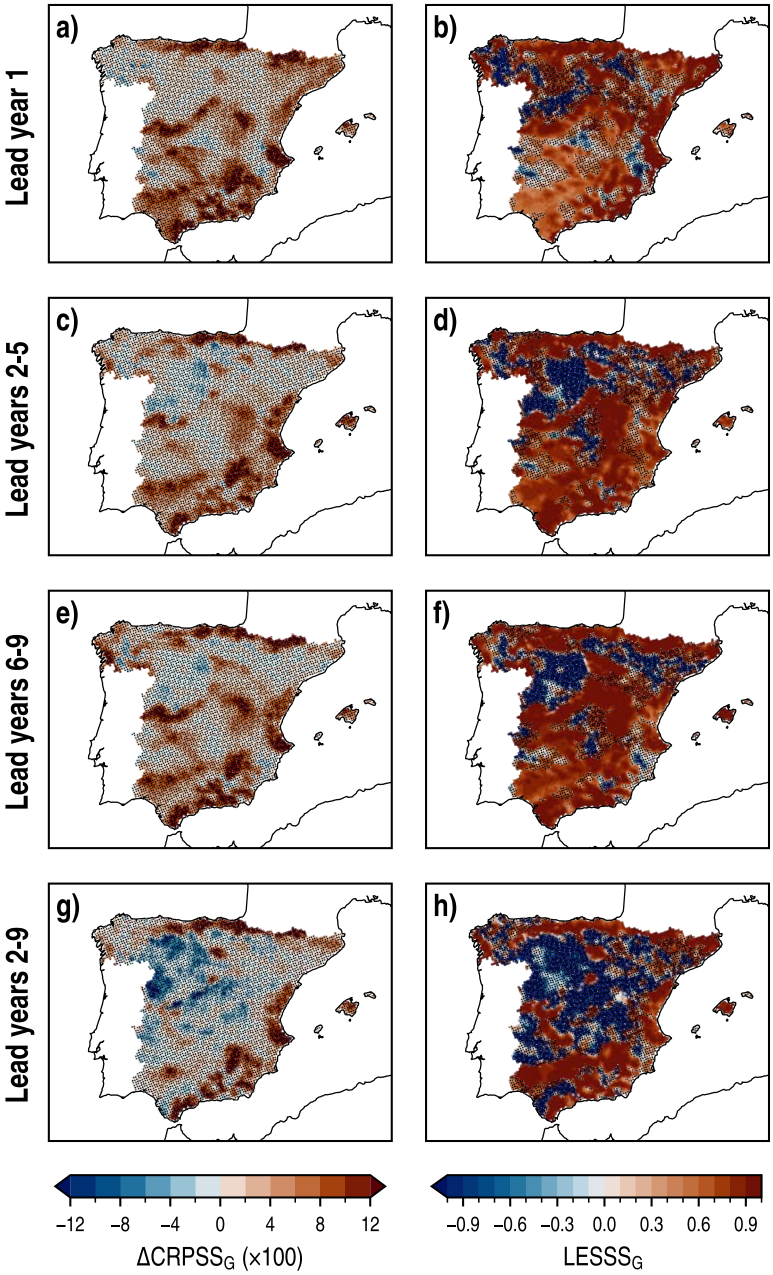

Figure 4.10: Spatial distributions of ΔCRPSS$_G$ (left column) and LESSS$_G$ (right column) for the WRF-DPLE multiannual mean PR anomalies, with CESM-DPLE as reference, in lead years 1, 2–5, 6–9 and 2–9 (rows) at annual scale. The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

4.3. Comparison with the WRF-LE ensemble

The comparison between the performances of the WRF-DPLE and WRF-LE ensembles has been constrained to the analysis of the deterministic metrics because the uninitialized experiments are not characterized by having a probabilistic nature, in contrast to the DCPs (see Section 1.1.2). Note that these results are affected by a large sampling bias because the analysis period in this case is very much shorter than that used when examining the predictive skill of the WRF-DPLE and its performance compared to CESM-DPLE. As said in Section 2.1.2, the CESM-LE data available for DD only covers the period 1990-2005. Therefore, WRF-DPLE hindcasts initialized every year from 1990 to 1999 (10 start dates) have been used to address this comparison only for lead years 1 and 2–5.

❦ *Accuracy analysis*

The best results in terms of the $\mathrm{MSSS_U}$, calculated with WRF-LE as reference, are observed in lead year 1 (Figure 4.11). Positive outcomes widely cover the northwest close the Cantabrian coast, west and south of the IP, but the statistically significant results are limited to smaller areas. The highest significant values has been found in the northwestern regions, with scores ranging from 0.3 to 0.8. On the other hand, WRF-LE clearly outperforms WRF-DPLE over the northeastern quarter of the domain, where a large area of significant negative values below -0.8 has been found, indicating that there is not a positive effect of initialization on predictive skill. The performance of WRF-DPLE, relative to that of WRF-LE, generally worsen in lead years 2–5 in terms of $\mathrm{MSSS_U}$. Some regions in the north and south maintain positive results, again with the most optimistic results placed close the Cantabrian coast around 0.7. However, part of the not significant positive results observed in the southern half of the domain for lead year 1 has turned into negative. In this case, the significant negative values also cover the northeastern regions, the Northern Subplateau and the Mediterranean coast.

As happened for the results depicted in Figure 4.8, the outcomes of $\mathrm{MSSS_U}$ can be explained by the improvement or deterioration of the predictive skill in terms of ACC and CB. The positive results of $\mathrm{MSSS_U}$ in lead year 1 are mainly a consequence of the added value to predictive skill of WRF-DPLE observed in $\Delta\mathrm{ACC_U}$. Significant results from 0.2 to 0.8 can be observed mainly in the northwest and some smaller southern locations. The confluence of statistically significant outcomes for $\Delta\mathrm{ACC_U}$ and $\Delta\mathrm{CB_U}$ can also lead to significant and negative results in $\mathrm{MSSS_U}$. The lost value

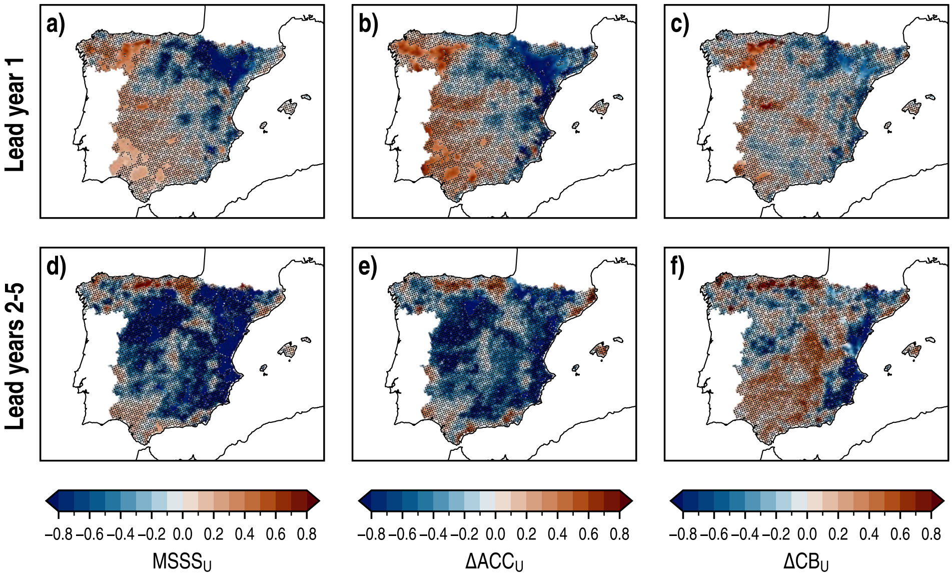

Figure 4.11: Spatial distributions of MSSS_U (left column), $\Delta\text{ACC}_\text{U}$ (center column) and ΔCB_U (right column) for the WRF-DPLE multiannual mean anomalies of PR, with WRF-LE as reference, in lead years 1, 2–5, 6–9 and 2–9 (rows) at annual scale. The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

in terms of ACC, which cannot be compensate by the gain in terms of CB, is the main responsible of the low MSSS_U scores observed in lead year 2–5. The results found by Reyers et al. (2019) on the added value of their downscaled hindcasts over their global uninitialized simulations[6] show more or less positive scores in terms of MSSS depending on the initialization scheme of the global model used to conduct the DD, highlighting the importance of that process in the representation of PR, as previously said in Sections 4.1 and 4.2. With an ensemble of global uninitialized experiments composed of 10 members as reference, the added value of their downscaled hindcasts is observed for an ensemble size of 7 or more members in terms of both MSSS and ACC. If another initialization scheme is used, the added value may be only shown in terms of ACC for an ensemble size of 10 member.

At seasonal scale, the results for MSSS_U (Figure 4.12) are more optimistic in JJA and SON, when an added value of WRF-DPLE is observed over large areas in the domain, especially in lead year 1, although the statistically significant results are shown only in small regions. These spatial distributions are consequence of the

[6] Note that *downscaled*, and not *global*, uninitialized simulations are used in this Thesis.

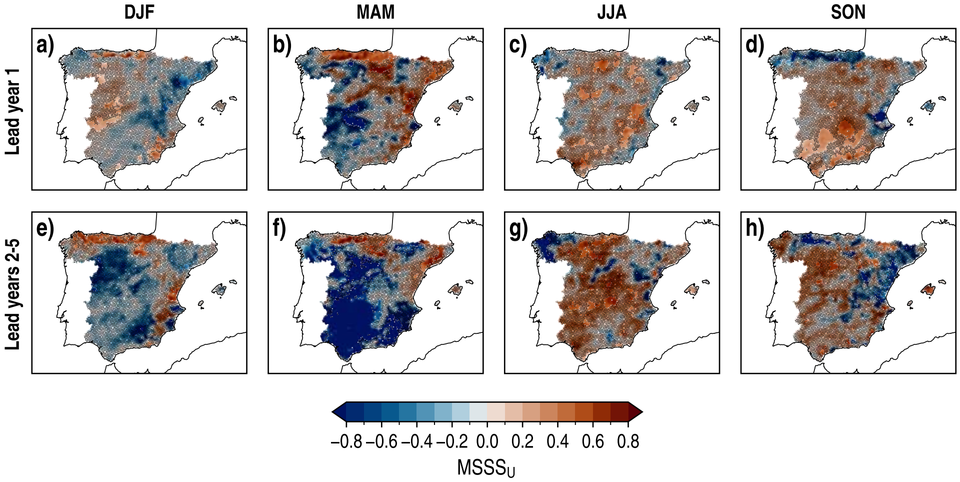

Figure 4.12: Spatial distributions of MSSS$_U$ for the WRF-DPLE multiannual mean anomalies of PR, with WRF-LE as reference, for lead years 1, 2–5, 6–9 and 2–9 (rows) in DJF, MAM, JJA and SON (columns). The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

important positive changes in ACC and CB (Figures B.10 and B.11 in Appendix B.1, respectively) observed during these seasons. In DJF and MAM, the presence of areas where there is a deterioration of the predictive skill is more frequent. Even so, there are some locations which show significant, positive results in ΔACC$_U$ and ΔCB$_U$ which lead to significant MSSS$_U$ outcomes from 0.3 to over 0.8, such as the regions along the Cantabrian coast.

4.4. Predictive skill for regional averages

The regions found after applying the regionalization scheme described in Section 3.6 to the AEMET seasonal PR means (Figure 3.5a) have been used to calculate regional averages of the lead time series. Since the regionalization groups together those locations which have similar PR regimes, these averages can be interpreted as a overall representation of the variable evolution in their respective regions. The spatially averaged lead time series have been subjected to a similar analysis to that done at grid-point scale in order to evaluate the predictive skill, whose results are summarized in Table 4.1.

The performance of the downscaled hindcasts within each region reflects what has already been discussed along the previous sections. In general, there is a poor representation of the PR across all lead times in terms of accuracy. However, there are

Table 4.1: Skill scores for the spatially averaged WRF-DPLE multiannual mean anomalies of PR in lead years 1, 2–5, 6–9 and 2–9 at annual scale. The subscripts *C*, *G* and *U* denote the reference data used to calculate the skill score: AEMET climatology, CESM-DPLE global hindcasts and WRF-LE uninitialized experiments, respectively. The bold formatting indicates results different from zero at the 90 % confidence level. Dashes denote data unavailability at that lead time.

Region	**Lead years**	**MSSS$_{\mathbf{C}}$**	**ACC**	**CB**	**CRPSS (×100)**	**MSSS$_{\mathbf{G(U)}}$**	**ΔACC$_{\mathbf{G(U)}}$**	**ΔCB$_{\mathbf{G(U)}}$**
EI	1	-0.11	0.19	**-0.39**	-0.56	-0.13 (-0.70)	-0.11 (-0.46)	-0.11 (0.05)
	2-5	-0.17	0.00	**-0.42**	-0.01	0.01 (-1.39)	-0.05 (-0.48)	0.02 (0.32)
	6-9	**-0.20**	0.00	**-0.44**	-0.11	0.06 (–)	0.09 (–)	0.08 (–)
	2-9	0.05	0.27	**-0.14**	-4.91	0.02 (–)	0.01 (–)	0.05 (–)
WI	1	-0.02	0.24	**-0.27**	**-2.75**	-0.15 (0.19)	-0.13 (0.26)	-0.15 (0.09)
	2-5	**-0.13**	-0.02	**-0.37**	-0.02	0.00 (-0.57)	-0.02 (-0.60)	0.00 (0.17)
	6-9	0.04	0.28	**-0.20**	**0.01**	0.08 (–)	0.21 (–)	0.02 (–)
	2-9	**-0.12**	0.15	**-0.37**	-7.01	0.02 (–)	0.09 (–)	0.01 (–)
NE	1	-0.14	0.03	**-0.37**	**-0.51**	0.06 (-0.30)	0.02 (-0.38)	0.09 (-0.21)
	2-5	**-0.78**	-0.02	**-0.88**	**0.08**	**-0.36** (-0.07)	-0.11 (-0.13)	**-0.32** (-0.05)
	6-9	**-0.52**	-0.10	**-0.73**	-0.36	-0.21 (–)	0.05 (–)	-0.20 (–)
	2-9	**-0.74**	**-0.20**	**-0.89**	0.02	**-0.36** (–)	-0.09 (–)	-0.34 (–)
CS	1	0.01	0.24	**-0.22**	**-5.58**	-0.09 (0.10)	-0.06 (0.11)	-0.17 (-0.06)
	2-5	-0.11	0.03	**-0.34**	0.06	-0.04 (-0.45)	-0.04 (-0.54)	-0.06 (0.56)
	6-9	-0.06	0.14	**-0.29**	**-0.93**	-0.02 (–)	0.08 (–)	-0.08 (–)
	2-9	**-0.02**	0.20	**-0.25**	-1.12	-0.11 (–)	-0.09 (–)	-0.19 (–)
NW	1	0.28	**0.53**	-0.08	0.07	-0.06 (0.32)	-0.03 (**0.39**)	-0.07 (0.08)
	2-5	-0.09	-0.06	**-0.30**	-0.37	0.14 (-0.05)	0.14 (-0.40)	0.24 (-0.08)
	6-9	0.07	0.28	**-0.09**	**-1.44**	0.13 (–)	0.20 (–)	0.17 (–)
	2-9	0.02	0.23	**-0.17**	-1.13	0.24 (–)	0.36 (–)	0.37 (–)
EA	1	-0.09	-0.59	**-0.66**	**-0.96**	0.13 (-0.06)	-0.41 (-0.27)	-0.13 (0.06)
	2-5	**-0.26**	-0.19	**-0.54**	**-4.16**	-0.05 (**-5.32**)	0.04 (-1.09)	-0.04 (**-1.10**)
	6-9	**-0.32**	-0.27	**-0.63**	**-4.92**	-0.07 (–)	-0.03 (–)	-0.09 (–)
	2-9	-0.16	-0.08	**-0.41**	**-3.69**	-0.04 (–)	0.06 (–)	-0.04 (–)
SW	1	-0.11	0.04	**-0.33**	**-7.94**	-0.10 (0.15)	-0.10 (0.34)	-0.18 (0.10)
	2-5	0.07	0.28	-0.06	-0.80	0.09 (0.13)	0.18 (0.04)	0.10 (0.12)
	6-9	-0.09	0.10	**-0.31**	**-1.12**	-0.04 (–)	0.10 (–)	-0.10 (–)
	2-9	-0.13	0.11	**-0.37**	-0.48	-0.16 (–)	-0.08 (–)	-0.28 (–)
CN	1	-0.02	0.23	**-0.28**	**-0.62**	-0.03 (-0.11)	0.06 (-0.27)	-0.11 (-0.10)
	2-5	0.06	0.31	**-0.21**	4E-03	0.10 (0.54)	0.21 (0.21)	0.04 (0.73)
	6-9	0.15	0.42	**-0.14**	-4E-03	0.10 (–)	0.16 (–)	-0.03 (–)
	2-9	0.23	0.48	**-0.01**	-0.26	0.21 (–)	0.28 (–)	0.08 (–)

many regions where the hindcasts perform well in terms of reliability, particularly in lead years 2–9, showing that the average ensemble spread is suitable to quantify the forecast uncertainty because the CRPSS results are not significantly different from zero. The best results have been achieved in the northwestern region (NW) for lead year 1, as could be expected from the maps shown in Section 4.1. NW is the only region which has a significant positive ACC, with a value of 0.53 in lead year 1, and shows a significant added value over the uninitialized simulations also in terms of ACC in lead year 1 ($\Delta \mathrm{ACC_U} = 0.39$). In addition, for this region, CB is not significantly different from zero at this lead time and the CRPSS results indicate that the hindcasts are reliable in lead years 2–5 and 2–9. Another region which occasionally showed optimistic results in the previous sections was the central north (CN). The regional averages show positive correlations for all lead times and positive $\mathrm{MSSS_C}$ values almost always, with the exception of lead year 1. However, the values are not high enough to get statistical significance. Some of the worst results have been obtained for the northeast (NE) region, as expected from the analysis at grid-point scale done in the previous sections. In this region, $\mathrm{MSSS_C}$ and CB show significant negative results. A significant loss of predictive skill is also observed when comparing with the global hindcasts in lead years 2–5 and 2–9.

The spatially averaged lead time series of the WRF-DPLE ensemble mean and AEMET for the NW region have been depicted in Figure 4.13. These representations also include the 90 % confidence interval associated to the probability of finding a single-member hindcast of the WRF-DPLE ensemble within it. This interval has been calculated by considering that the members of the ensemble follow a Gaussian distribution with a mean given by the WRF-DPLE ensemble mean (Eq. [3.5]) and a variance equal to the average ensemble spread (Eq. [3.32]), as well as was assumed to calculate the CRPSS. Note that this confidence interval cannot be interpreted as a measure of the uncertainty of the predictions for lead years 2–5 because CRPSS has a value significantly different from zero at this lead time, as opposed to lead years 1, 6–9 and 2–9. Additionally, the ensemble envelope is also represented to show the maximum deviation of the ensemble members from the mean. The largest model and observational variability have been found in lead year 1 (Figure 4.13a), when the smoothing effect of the lead time average is not as much accentuated as for the 4-year and 8-year lead time averages. The hindcasts are able to reproduce to some extent the AEMET variability at this lead time, as expected from the ACC = 0.53 result showed in Table 4.1, exhibiting skill to replicate some of the peaks observed in the AEMET time series, such as, for example, those corresponding to the start

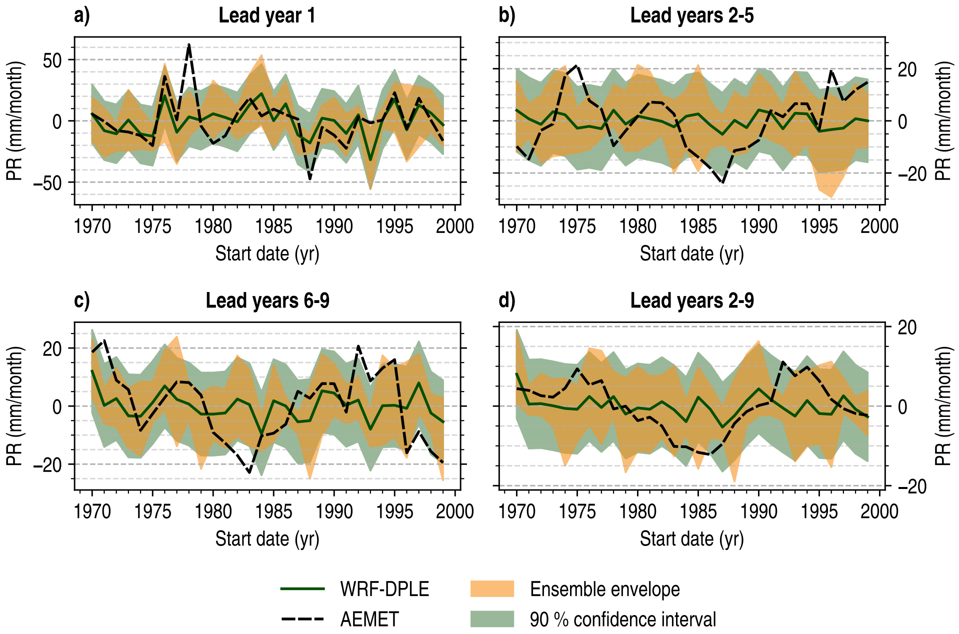

Figure 4.13: Time series of the spatially averaged multiannual mean anomalies of PR in the NW region for lead years 1, 2–5, 6–9 and 2–9 at annual scale. Solid green lines identify the WRF-DPLE ensemble mean, whereas dashed black lines correspond to AEMET. Shaded green surfaces indicate the 90 % confidence interval for a WRF-DPLE single member, calculated from the average ensemble spread (Eq. [3.32]). Shaded yellow surfaces show the ensemble envelope which encloses the trajectories followed by the members composing the WRF-DPLE ensemble.

years 1976, 1995, 1996 or 1997. In general, the WRF-DPLE time series stays close to the observational one in lead year 1, with the exception of some important peaks occurring for AEMET in start years 1978 and 1988, or another only traced by the hindcasts in 1993, explaining the not significant but positive $\text{MSSS}_C = 0.28$ showed in Table 4.1. The situation is different at the other lead times, when the low MSSS_C values already show the poor ability of the WRF-DPLE lead time series to replicate the magnitude of the variability observed for AEMET in Figures 4.13b to 4.13d, and the small correlations lead to significant negative CB values. All panels are good illustrative examples of the low signal-to-noise ratio found in the hindcasts for PR, as the width of the confidence intervals is generally much larger than the magnitude of the ensemble mean signal.

The spatially averaged time series for the CN region have been depicted in Figure 4.14. The general comments which describe the limited predictive skill of the hindcasts for PR in the region NW can also be extrapolated to this region at all lead times. In this case, the highest correlation has been obtained in lead years 2–9, with a not significant ACC = 0.48. An increase of the ensemble size would help to improve this correlation and even get a significant result. Nevertheless, the signal-to-noise ratio would still remain being not large enough to replicate some of the most accentuated maximum and minimums observed in the AEMET time series at this lead time, such as those observed for the start dates 1975, 1982 or 1987 (Figure 4.14d).

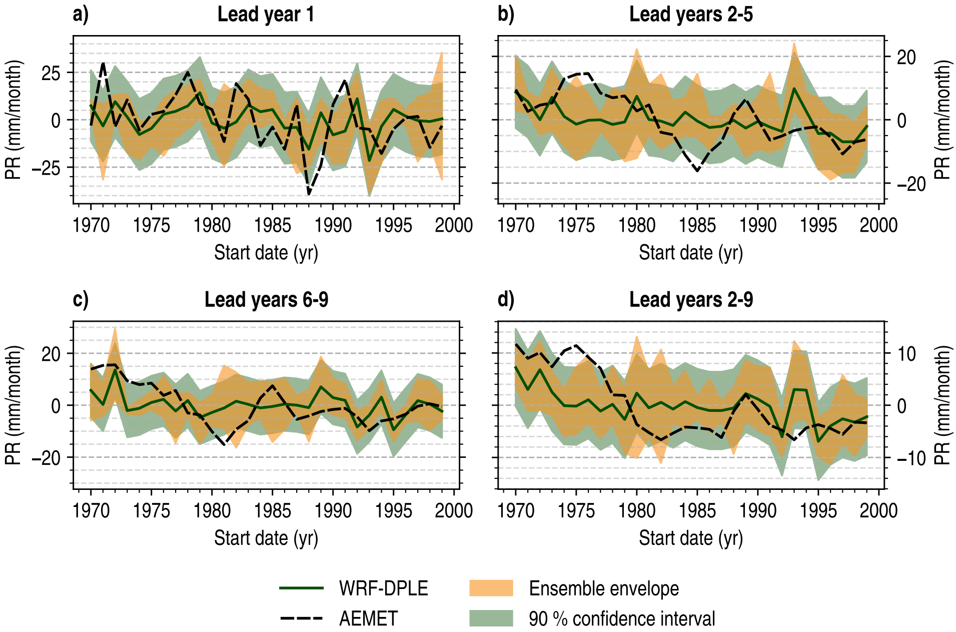

Figure 4.14: As Figure 4.13 but for the CN region.

4.5. Analysis of the spatio-temporal variability of sea level pressure in CESM-DPLE

The last part of this Chapter is devoted to briefly explore the skill of the CESM-DPLE subensemble to predict the main sea level pressure (SLP) spatio-temporal variability patterns in the Northern Hemisphere. The ability of DPSs to reproduce the mechanisms which control the atmospheric circulation has already been subject of analysis in previous studies (e.g., Dunstone et al., 2016; Smith et al., 2019, 2020). The interest to have skilful predictions of the atmospheric circulation relies on its influence on the evolution of fields such as PR or temperature at local scale. In DD simulations, the information about the atmospheric circulation represented by the GCM fields is transferred to the RCM through the ICs and LBCs, so the skill of CESM-DPLE to predict these variability modes influences on the predictive skill achieved by the WRF-DPLE output fields. So far, the research available in literature has mainly focused on the analysis of the skill to predict the North Atlantic Oscillation (NAO), the circulation pattern which most influences on weather and climate during winter in the Northern Hemisphere. The NAO is characterized by changes in SLP or geopotential height over the action centres located at the Azores and Iceland (Hurrell et al., 2003; Smith et al., 2019). It explains about a third of the SLP variance in the Northern Hemisphere during this season (Hurrell et al., 2003) and drives part of the PR and temperature variability in the southwestern regions in Europe (e.g., Queralt et al., 2009; Ríos-Cornejo et al., 2015b; Trigo et al., 2004). However, there are other circulation patterns (Wallace and Gutzler, 1981) which have not been received as much attention in the context of DCPs, but that also exerts some influence on PR and temperature in the particular case of the IP (Ríos-Cornejo et al., 2015a, 2015b).

This Section presents an analysis of the results obtained from a PCA (Preisendorfer, 1988) which has been applied to the CESM-DPLE and ERA5 SLP, being the latter used as the reference dataset. ERA5 provides SLP with a higher spatial resolution than CESM-DPLE, allowing to work on the native CESM-DPLE resolution ($\sim 1^\circ$) after interpolating ERA5 SLP onto the model grid. The PCA has been applied to the ensemble mean of the same 4-member subensemble used in DD simulations (ENS4), as well as to the ensemble mean of the 10-member CESM-DPLE subensemble (ENS10, the largest ensemble attainable for DD simulations) to evaluate the dependence of the results on the ensemble size. The PCA extracts the main spatio-temporal variability modes which determine the evolution of SLP, accounting not only for the NAO but also for other important circulation patterns in the Northern Hemisphere. The PCA has been applied to the lead time series of SLP along the control period (the decades initialized every year from 1970 to 1999) in lead years 2–9, so that the interannual variability is removed and the analysis fully focus on the decadal scale. The PCA have been computed for the covariance matrix in S-mode. The significant EOFs, selected by the North's rule (North et al., 1982), have been rotated by following a varimax approach (Preisendorfer, 1988), just as it was done in Section 3.6 as part of the process of regionalization of the IP.

The results obtained for the rotated loadings have been depicted in Figure 4.15. These rotated loadings represent the correlations between the standardized SLP anomaly series and the rotated PC (PC_R) associated to a given variability mode at each grid point. The ERA5 SLP spatio-temporal variability is represented by 5 significant modes, which explain about 92.01 % of the total SLP variance. The spatial distribution of the rotated loadings in the first ERA5 variability mode, which represent 35.75 % of the variance, shows two main action centers. One negative action center is placed over Greenland and also spans over part of the northern and northeastern Eurasia, whereas the other, positive in this case, covers the whole North Atlantic and spans part of the western Eurasia regions. Some positive correlations have been obtained also over the Pacific Ocean. This variability mode represents the PCA version of NAO in lead years 2–9. The correlation between its associated PC_R and a NAO index calculated with ERA5 by following the same approach described in Smith et al. (2019) for lead years 2–9 is 0.91 (p-value < 0.1, see Section 3.2.3[7]). In the second mode, which explains about 19.19 % of the variance, the strongest correlations of the negative action center are observed in the northwestern regions of America. On the other hand, positive correlations cover part of the western North Atlantic and the North American continent, reaching the highest values in the North Pacific. The rotated loadings of the third mode share some similarities with the pattern of the first mode, although only 13.60 % of the variance is explained in this case. Indeed, the correlation between the associated PC_R and the NAO index is 0.87 (p-value < 0.1, see Section 3.2.3). The main differences between the first and third modes, besides the explained variance, are mainly observed in the decrease of the absolute value in the correlations of the negative action center, and in the concentration of the highest North Atlantic positive correlations over the eastern North Atlantic sector, the southwestern Europe and the northern Africa. In addition, the cluster of positive correlations observed in the North Pacific Ocean has slightly shifted westward compared to the first mode. The fourth variability mode explains 12.91 % of the SLP variability. Its spatial pattern displays a strong action center over the North Pacific, to the northwest of America, and lower positive correlations along part of North America, the North Atlantic, Europe and the north of Africa. On the other hand, weaker negative correlations are observed to the south of Greenland, the north of America and in part of eastern Eurasia. The last significant variability mode presents two well-distinguishable negative action centers, one spanning North America, with the strongest correlations to the northwest of the continent, and another which covers eastern Europe and some southern and eastern Asian regions. The highest positive correlations, weaker than the negative ones, are observed mainly to the northwest of the British Isles, to the west of North America

[7] In the evaluation of the statistical significance done in this Section, the non-parametric bootstrapping over the PC_Rs calculated with the ensemble mean has been applied only for start dates.

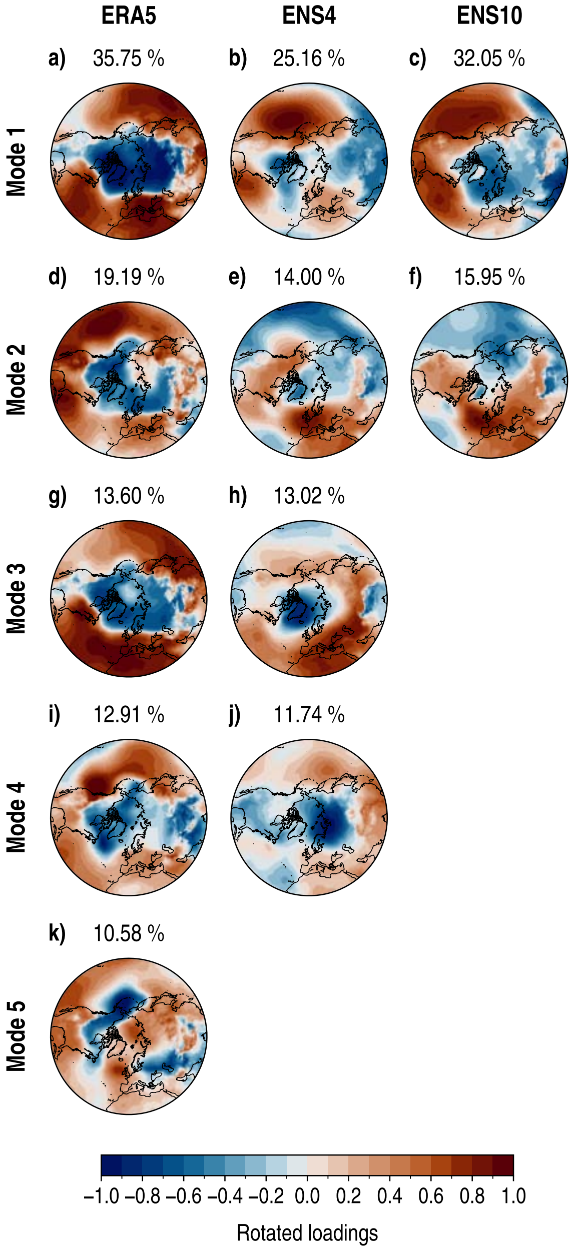

Figure 4.15: Rotated loadings of the ERA5 (left column), the 4-member CESM-DPLE ensemble mean (ENS4, center column) and the 10-member CESM-DPLE ensemble mean (ENS10, right column) SLP for each significant spatio-temporal variability mode (rows). They have been computed for lead years 2–9 in DJF. The variance ratio explained by each mode is shown in map headings.

over the North Pacific and over the Arctic Ocean.

The extent to which the CESM-DPLE hindcasts are able to reproduce the main atmospheric circulation patterns which influence on weather and climate over the North Atlantic sector is established by the skill to replicate the rotated loadings described above, alongside their associated PC_Rs. The rotated loadings obtained for the ENS4 and ENS10 ensemble means are also displayed in Figure 4.15, whereas the correlations between ERA5 and CESM-DPLE PC_Rs are shown in Table 4.2. ENS4 has 4 significant spatio-temporal variability modes, which explain around 63.92 % of the total SLP variance. This explained variance is more evenly distributed among the modes than in the case of ERA5. On the other hand, ENS10 presents only 2 significant variability modes, explaining together about 48 % of the total variance. The distribution of the explained variance among these modes is more similar to that observed for ERA5 than it was for ENS4. It may be related to the better ability to capture the climate signal and distinguish between variability modes compared to the 4-member subensemble because of its larger ensemble size. The first variability modes explain 25.16 % and 32.05 % of the variance for ENS4 and ENS10, respectively. They have rotated loadings which show some similar spatial features. Both are characterized by two positive action centers at both sides of North America over the Pacific and Atlantic Oceans. There are also negative correlations located over Greenland and part of Eurasia, which cover a larger area of the continent in the case of ENS10. These patterns have certain similarities with the second ERA5 variability mode, especially in the case of ENS10, as revealed not only by the common aspects regarding the rotated loadings but also by the significant positive correlation between the PC_Rs (ACC = 0.68). This correlation is also significant but much lower for ENS4 (ACC = 0.33). The PC_R 1 of ENS10 also shows a significant correlation of 0.39 with the PC_R 3 of ERA5. The second variability modes of ENS4 and ENS10, which explain 14–16 % of the variance, are characterized by a spatial pattern which shows positive correlations along North America, the meridional latitudes of the North Atlantic Ocean and part of Eurasia. The highest positive correlations are shown in the northwestern Europe, with a slight shift to the northwest in ENS10 compared to ENS4. The most intense negative correlations are observed along the Arctic and North Pacific Oceans, as well as in part of the northeast and southeast of Eurasia. In the case of ENS4, the area covered by these negative correlations over the north of Eurasia is larger. Both second modes show some features which are shared to some degree with the first three ERA5 modes and even with the fourth one in the case of ENS10, as the correlations in Table 4.2 indicate. While the correlations for the CESM-DPLE PC_R 2

Table 4.2: ACC calculated with the ERA5 and CESM-DPLE rotated principal components (PC$_{\text{R}}$s) for lead years 2–9 in DJF. The CESM-DPLE PC$_{\text{R}}$s have been computed for the 4-member and 10-member ensemble means (ENS4 and ENS10, respectively). The variance ratio explained by each significant spatio-temporal variability mode is shown in brackets below the PC$_{\text{R}}$s. The bold formatting indicates results different from zero at the 90 % confidence level. In the evaluation of the statistical significance, the non-parametric bootstrapping conducted with the PC$_{\text{R}}$s has been applied only for start dates (see Section 3.2.3).

ERA5	**ENS4**				**ENS10**	
	PC$_{\text{R}}$ 1 (25.16 %)	PC$_{\text{R}}$ 2 (14.00 %)	PC$_{\text{R}}$ 3 (13.02 %)	PC$_{\text{R}}$ 4 (11.74 %)	PC$_{\text{R}}$ 1 (32.05 %)	PC$_{\text{R}}$ 2 (15.95 %)
PC$_{\text{R}}$ 1 (35.75 %)	0.08	**-0.28**	-0.07	**-0.13**	0.18	-0.28
PC$_{\text{R}}$ 2 (19.19 %)	**0.33**	-0.39	**-0.36**	0.10	**0.68**	**-0.41**
PC$_{\text{R}}$ 3 (13.60 %)	0.21	**-0.34**	-0.03	0.02	**0.39**	**-0.39**
PC$_{\text{R}}$ 4 (12.91 %)	0.07	-0.01	0.15	**-0.26**	0.16	0.29
PC$_{\text{R}}$ 5 (10.58 %)	0.05	-0.14	**-0.31**	**0.28**	0.30	-0.09

are only significant with the ERA5 PC$_{\text{R}}$s 2 and 3 in the case of ENS10 (ACC = −0.41 and ACC = −0.39, respectively), they are significant with ERA5 PC$_{\text{R}}$s 1 and 3 in the case of ENS4 (ACC = −0.28 and ACC = −0.34, respectively). However, the ENS4 and ENS10 rotated loadings for the second mode do not consistently capture the location of the main action centers of any ERA5 mode. The third and fourth variability modes of ENS4 (13.02 % and 11.74 % of explained variance, respectively) are not able to fully reproduce any ERA5 mode either. The strongest correlation with the ERA5 PC$_{\text{R}}$s is observed between ERA5 PC$_{\text{R}}$ 2 and ENS4 PC$_{\text{R}}$ 3, with a statistically significant result of -0.36. Some common spatial features between their respective rotated loadings have been found but, as occurred before, the location of the main action centers is not accurate.

The results described above show some skill in the CESM-DPLE hindcasts to partially reproduce the SLP variability, although they are not able to clearly capture most part of the main spatio-temporal variability modes extracted from the PCA computed with ERA5 SLP. Despite the second ERA5 mode is in a certain way replicated by the ENS10 ensemble mean (the maximum attainable ensemble size for DD), it is not the case for ENS4. None of the CESM-DPLE ensembles have been able to consistently capture the NAO variability mode either, the mode which dominates the atmospheric circulation in the North Hemisphere during the boreal winter months, which was found in the ERA5 SLP. Smith et al. (2019) stressed the need to consider very large ensemble sizes to properly capture the NAO signal, which is strongly affected by the signal-to-noise paradox. For a 4-member multimodel ensemble mean and 46 start dates (initialized from 1960 to 2005), the authors showed that the skill to reproduce the station-based NAO index in terms of ACC is below 0.2 and not significant at the 90 % confidence level. For a multimodel ensemble size of 10-members, however, the ACC may increase up to 0.25 and show statistical significance. The limited skill to reproduce this variability pattern by ENS4, which provides the ICs and LBCs for the DD decadal simulations conducted in this Thesis, might be negatively affecting the predictive skill for PR in the IP in the results discussed in the previous sections, as NAO has a notable influence on the PR variability in this region, especially in DJF (Queralt et al., 2009; Ríos-Cornejo et al., 2015a; Trigo et al., 2004). Indeed, since NAO also drives part of the temperature variability in Spain (Ríos-Cornejo et al., 2015b), improvements in the skill to capture the NAO signal (and the other variability modes in general) may also contribute to enhancing the predictive skill of the hindcasts analyzed in the following Chapter 5. In this line, Smith et al. (2020) revealed that a multimodel ensemble composed of 169 members subjected to a post-processing adjustment could increase the ACC with NAO index up to 0.79, introducing improvements in the representation of PR, temperature and SLP. However, this attractive approach has some obstacles to overcome in the framework of DD. The first one is the aforementioned computing cost of performing DCPs at a very high resolution by means of DD simulations. The second one, also very important, is the availability of GCM data to conduct the simulations. The huge storage requirements needed to save the data generated by GCMs in DCPs leads the research groups to select a subset of the output product to be saved. This subset usually encompasses a wide variety of fields from frequencies ranging from daily to annual, but commonly lacks of most part of the 6-hourly mandatory fields to provide the ICs and LBCs of a RCM. As mentioned in Section 2.1.1, CESM-DPLE is the only DPS which publicly provides all the required fields to address this task. Nevertheless, there is still a potential to continue improving the predictive skill of the presented in this Chapter with the available data, adding new members up to an ensemble size of 10 and also increasing the number of start dates of the downscaled ensemble back to 1956 (15 additional start dates).

4.6. Concluding remarks

This Chapter has been devoted to the analysis of the WRF-DPLE downscaled hindcasts for PR. Several metrics and skill scores have been used to evaluate the performance of these experiments in terms of their accuracy, their reliability and their added value to the predictive skill over the global CESM-DPLE hindcasts and the WRF-LE uninitialized experiments. The ability of CESM-DPLE to reproduce the spatio-temporal variability of the SLP has also been explored. Before conducting the evaluation, the WRF-DPLE experiments have been recalibrated by applying the DeFoReSt method to reduce the unconditional and conditional biases as well as improve the representation of the ensemble spread (see Section 3.5). The main findings are summarized in the following:

- **The signal-to-noise paradox is strong in the WRF-DPLE hindcasts for PR.** The signal-to-noise paradox leads to the counterintuitive situation in which the model is better at predicting the real climate evolution than it is at predicting itself. This paradox is a consequence of a low signal-to-noise ratio in model predictions. It is present not only in the WRF-DPLE hindcasts but also in the 4-member CESM-DPLE subensemble, as revealed by the results obtained for the spatial distributions of the RPC in Figure 4.5. These results, which are consistent with previous studies available in literature, indicate that the predictive skill for PR would benefit from increasing the ensemble size.

- **At annual scale, the spatial distributions of ACC show positive results over most part of the domain at all lead times.** On the other hand, negative results are commonly found in the eastern flank of the IP. The results generally lack of statistical significance, with the exception of some regions with positive ACC values. The most promising outcomes have been found in the northwestern regions for lead year 1, where significant positive scores up to values around 0.7 have been obtained. The spatial distributions of RMSE_R show the lowest errors in the northern regions of the domain at all lead times because the highest PR rates are commonly observed there.

- **Some of the best results obtained at seasonal scale in terms of ACC have been found in JJA, especially in lead years 1 and 2–5.** These spatial distributions show generalized positive results in the IP, although the statistical significance is limited to specific regions. For example, the results are significant in part of the southern half of the IP for lead year 1 and in some northern and southern locations for lead years 2–5. Positive ACC values are also common in DJF and

MAM. The lowest $\mathrm{RMSE_R}$ values have been found in MAM, whereas the highest scores are shown mainly across the southern regions in JJA. These high relative errors are mainly motivated by the lower PR rates observed in this part of the domain during this season.

- **The WRF-DPLE predictive skill for PR at annual scale, with climatology as reference, is limited.** Generalized negative results have been found for $\mathrm{MSSS_C}$ at almost all lead times. Some positive scores are also observed in small regions, especially in lead year 1, but with absence of statistical significance. These results are consequence of the joint action of the low ACC and the high absolute CB values. Even if the CB were completely removed, the predictive skill would continue being low because of the results obtained for ACC, as indicated by the spatial distribution of $\mathrm{MSSS_{CBA}}$ ($\mathrm{MSSS_C}$ for CB = 0). Similar results have been obtained at seasonal scale.
- **The WRF-DPLE hindcasts for PR are generally reliable over wide areas of the IP.** This means that the average ensemble spread can be used to quantify the uncertainty of the predictions in those regions. Almost the whole domain shows reliable hindcasts in lead years 2–9 at annual scale. Nevertheless, the areas with not significant CRPSS outcomes are smaller at the other lead times. The regions in the northern half of the domain are generally among those showing hindcast reliability at all lead times. This reliability is caused by the results obtained for LESS, which are commonly not significant, indicating that there are not significant differences between the average ensemble spread and the squared standard errors in those locations. At seasonal scale, the hindcast reliability has been found over almost the whole domain for all lead times in MAM. Very good results have also been obtained in JJA and, to a lesser extent, in SON.
- **The highest predictive skill of WRF-DPLE for PR at annual scale with CESM-DPLE as reference has been found in lead years 6–9.** The generalized positive $\Delta\mathrm{ACC_G}$ results obtained at this lead time, along with the improvement also observed in terms of $\Delta\mathrm{CB_G}$, lead to these positive $\mathrm{MSSS_G}$ values over large areas. However, the statistical significance is constrained to very small regions in the central eastern part of the domain. The lack of statistical significance has also been observed at the other lead times. In those cases, the areas with positive scores are smaller. The northern, northwestern and central eastern areas of the IP have generally obtained the best results in lead years 2–5, 6–9 and 2–9. In lead year 1, they are fundamentally found along the Mediterranean coast and

part of the Northern Subplateau.

- **The best results obtained for the WRF-DPLE predictive skill at seasonal scale, with CESM-DPLE as reference, have been found in JJA.** In this season, the positive $MSSS_G$ results are predominant in lead years 2–5, 6–9 and, especially, 2–9. Nevertheless, as at annual scale, the regions not showing statistical significance are widespread. These outcomes are a consequence of two concurring factors: the generalized positive results obtained for ΔACC_G and ΔCB_G at these lead times. Both metrics show positive results spanning most part of the domain in lead years 2–5 and 2–9, with smaller areas in lead years 6–9.
- **The PR hindcast reliability is higher for WRF-DPLE than for CESM-DPLE.** Although these results lack of statistical significance, the positive $\Delta CRPSS_G$ values are clearly predominant for the lead years 1, 2–5 and 6–9 at annual scale, being also found in the coastal and some inner regions in lead years 2–9. There is a large improvement in the representation of the ensemble spread in the WRF-DPLE hindcasts, as revealed by the positive results obtained for $LESSS_G$, which are the cause of the results observed for $\Delta CRPSS_G$. At seasonal scale, the highest improvements in the hindcast reliability have been found in DJF and SON for the same reasons.
- **The highest predictive skill of WRF-DPLE, with the WRF-LE uninitalized experiments as reference, has been mainly found for lead year 1 in the western part of the domain.** Large areas with positive $MSSS_U$ results are observed over those regions at this lead time. These results are caused by the positive scores found for ΔACC_U and ΔCB_U. On the other hand, the added value obtained in terms of $MSSS_U$ is fundamentally constrained to some northern and southern regions in lead years 2–5. The most promising outcomes obtained for this metric at seasonal scale are shown in JJA and SON, when the added value of WRF-DPLE is observed over large areas of the domain, especially in lead year 1, but with a general lack of statistical significance. This evaluation has been constrained to the decade 1990-2005 because of the unavailability of CESM-LE data outside this period, so a large sampling error affect these results. Therefore, they should be taken with caution.
- **The performance of the WRF-DPLE hindcasts for PR, from a regional perspective, reflects what has been obtained at grid-point scale.** There is a general poor representation of PR at all lead times in terms of accuracy. Nevertheless, in regard to reliability, the performance is promising in many regions, especially

for lead years 2–9. The NW and CN regions show some of the best results in general. There, the hindcasts are able to reproduce part of the relative minimums and maximums present in the observational time series. In general, the magnitude of the ensemble mean signal is lower than that of the observational time series and the amplitude of the confidence intervals, as a consequence of the signal-to-noise paradox.

- **The limited ability of the 4-member CESM-DPLE subensemble to represent the spatio-temporal variability of SLP may be partially responsible of the results obtained for the WRF-DPLE PR.** The 4-member CESM-DPLE subensemble can reproduce part of the SLP variability, but it cannot capture most part of the spatio-temporal variability modes extracted from the PCA of the ERA5 SLP. Indeed, neither the 4-member nor the 10-member subensembles are able to consistently simulate the NAO. Further improvements carried out in this line could contribute to enhancing the predictive skill of the downscaled product.

5

Retrospective decadal climate predictions for near-surface air temperature

This Chapter is dedicated to the analysis of the WRF-DPLE predictive skill for NSAT. The structure is very similar to that followed in Chapter 4 for the analysis of PR, but this Chapter includes results for three NSAT variables: T_{max}, T_{min} and T_{mean}.

In the section devoted to each variable, at first, the predictive skill of the ensemble of WRF-DPLE hindcasts is evaluated against the AEMET observational datasets to examine the accuracy and reliability. Secondly, the performance of WRF-DPLE ensemble is compared to that of CESM-DPLE subensemble, the source which provides ICs and LBCs to conduct the DD experiments with WRF. Afterwards, the performance of WRF-DPLE is compared to that of WRF-LE instead, the dynamically downscaled uninitialized experiments. Finally, the WRF-DPLE predictive skill of the regional averages of the variables is also evaluated.

5.1. Daily maximum near-surface air temperature

5.1.1. *Predictive skill of the WRF-DPLE ensemble*

The following results have been achieved from the analysis of the WRF-DPLE T_{max} ensemble after being recalibrated by applying the DeFoReSt approach (Pasternack et al., 2018, 2021) with an additional filtering to adjust the inherent biases in DCPs. See Section 3.5 for more information.

❦ *Accuracy analysis*

The spatial distributions of the RMSE and ACC calculated for the multiannual mean anomalies of T_{max} have been depicted in Figure 5.1. A pattern in RMSE maps similar to that observed in Figure 4.1 for PR has also been found here, as expected, in the sense that the highest RMSE values are found for lead year 1 and decrease by increasing the length of the lead time window. In lead year 1, the maximum values are above 0.9 K, mainly in some mountain regions located in the southern half of the domain. This is the case of those located to the south of the Northern Subplateau, in Sierra Morena, in regions close to the Strait of Gibraltar and belonging to the Baetic System. The regions with the lowest errors, around 0.45 K, are commonly those with lower terrain height, such as the Ebro, Guadalquivir or Tagus valleys, or the locations situated along the Mediterranean coast (a terrain elevation map is available in Figure 1.4a). The differences between RMSE values in lead years 2–5 and 6–9 are more noticeable for $T_{\max}$ than for PR. In this case, the errors slightly decrease with lead time, generally getting differences about 0.1 K between these lead time windows. The highest RMSE values are primarily observed in some of the highest mountain locations of the IP, with the Baetic System depicting the worst results. The errors over the Baetic System are around 0.75 K and 0.65 K in lead years 2–5 and 6–9, respectively. A qualitatively similar spatial distribution is observed in lead years 2–9, but being the maximum RMSE, found over the Baetic System again, around 0.55 K. In these high mountain regions, the complexity of the topography directly affects the installation and maintenance of observational stations, complicating the monitoring of variables such as temperature and, consequently, negatively impacting on the availability of quality spatio-temporal observational information (Esteban-Parra et al., 2022). This might explain, at least to some extent, the high errors observed in these regions. On the other hand, the rest of the domain generally exhibits a consistent decrease for the RMSE compared to the other lead times, with a large fraction of the domain showing results below 0.2 K.

The results obtained for ACC in the analysis of $T_{\max}$ are certainly more promising than those found in Section 4.1 for PR. In this case, the results show positive ACC values spanning almost the whole domain regardless of the lead time. As opposed to PR, the signal is dominated by upwards climate trends in temperature-based variables, positively contributing to the ACC when the models are able to reproduce them reasonably well (Goddard et al., 2013). In lead year 1, statistically significant positive results have been found in some northwestern regions, a stripe of northeastern locations over the southern Pyrenees and surrounding regions, the Balearic Islands and an area covering part of the Baetic System in the south. These values are in the range from 0.4 to 0.8. The rest of the domain, including those regions with negative outcomes, does not show statistical significance. As for RMSE, the lead time has some influence on the ACC results, which are better in lead years 6–9 compared to lead years 2–5, fundamentally in regard to the statistical significance. While the outcomes

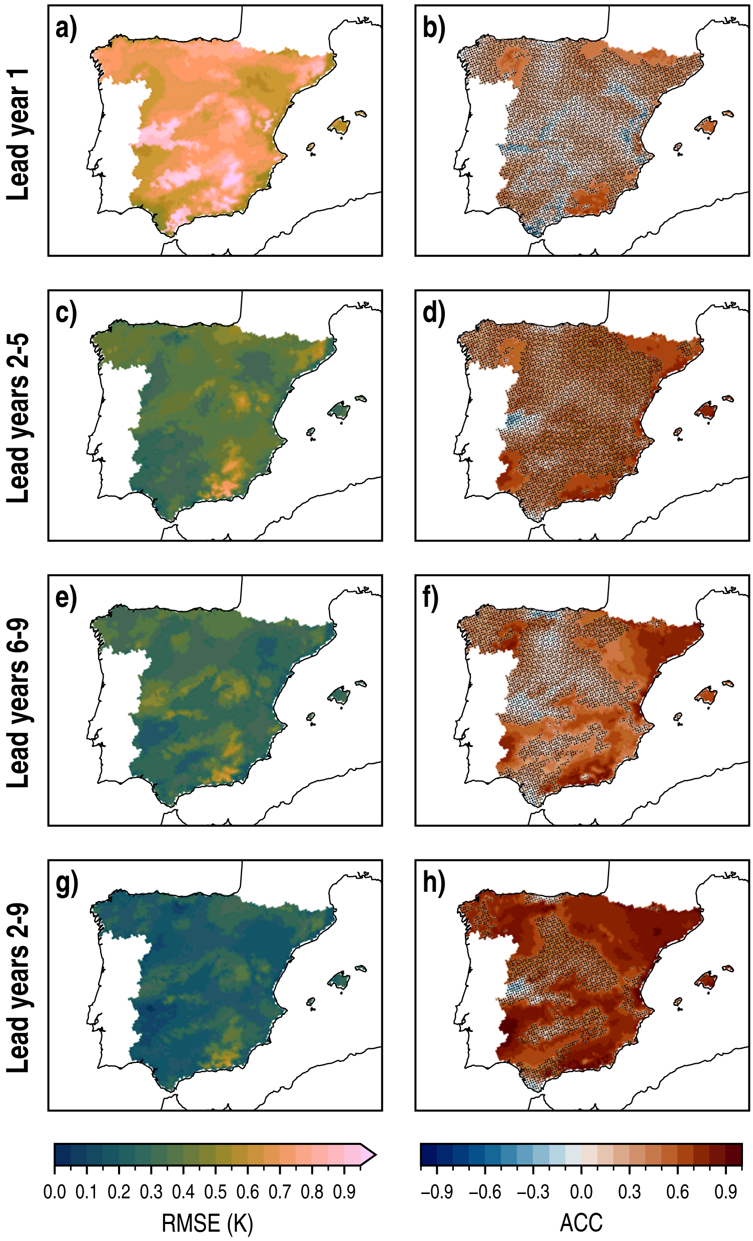

Figure 5.1: Spatial distributions of RMSE (left column) and ACC (right column) for the WRF-DPLE multiannual mean anomalies of $T_{\max}$ in lead years 1, 2–5, 6–9 and 2–9 (rows) at annual scale. In ACC maps, the absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

in lead years 2-5 show slightly more robustness that at the first year of the decade, the performance in lead years 6–9 is even better, getting wider areas of the domain with significant positive results. These regions are mainly situated in the eastern and southern halves of the domain, and show correlations between 0.4 and 0.9. In this case, the northeastern regions show the most promising results. The best outcomes have been obtained in lead years 2–9, when the statistical significance for positive correlations is widely spread along the domain. The significant positive ACC values are above 0.5, reaching results even above 0.9 in some small areas situated in the northeast and south.

At seasonal scale, the results for RMSE (Figure B.12 in Appendix B.2.1) vary depending on the season, although the analysis follows the same fundamental line observed at annual scale. The RMSE spatial distributions mostly show similar patterns to that for the standard deviation of the AEMET T_{max} (Figure B.13 in Appendix B.2.1), in the sense that locations with higher RMSE values also are those which show the higher variability. This does not occur in lead year 1 at annual scale for some of the aforementioned regions which have the largest errors. Only those in the Baetic System have a high variability, the rest does not show errors which can be attributed to this, only to a bad representation of the magnitude of the variable by the hindcasts. The results depicted later in Figure 5.4 suggest that this misrepresentation might be due to a very negative CB, caused by the imbalance between the low correlations and the ratio of the hindcast and observational standard deviations.

The contrast among the seasons is more noticeable in the results for the seasonal ACC (Figure 5.2). The best results have been found for MAM, followed by JJA. In general, there are almost not statistically significant results in lead year 1, even less than at annual scale. In MAM, the situation drastically changes in lead years 2–5 because the whole domain is covered by significant positive results, excepting some regions scattered along the IP, mainly located in the southeast and the northwest. The largest correlation coefficients can be observed in the northeastern regions, with values ranging from 0.6 to 0.9. The amount of significant results decreases in lead years 6-9, but are maintained along the east and part of the southern regions of the IP. The most promising results are depicted in lead years 2–9, when the scores are essentially significant and positive and the highest correlations are located along the Mediterranean coast with values from 0.8 to 0.9. On the other hand, the worst represented season is clearly SON. Not significant negative results are generally predominant at all lead times with the exception of lead years 2–9. At this lead time, significant negative ACC values cover large portions of the western part of the domain,

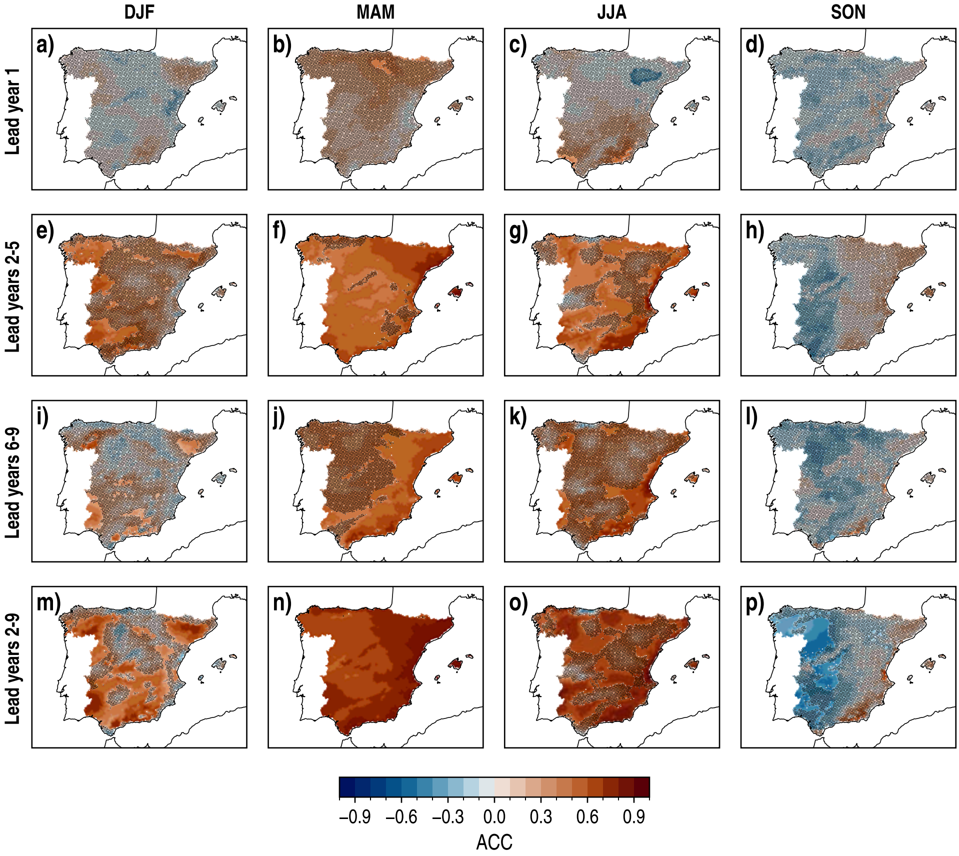

Figure 5.2: Spatial distributions of ACC for the WRF-DPLE multiannual mean anomalies of T_{max} for lead years 1, 2-5, 6-9 and 2-9 (rows) in DJF, MAM, JJA and SON (columns). The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

especially across the northwestern sector. The good results obtained in MAM and JJA, relative to the other seasons, are in part due to the fact that T_{max} trends are more pronounced. Tables A.3 to A.6, available in Appendix A.2, can be consulted to examine the seasonal trends corresponding to each region resulting from the regionalization done in Section 3.6.

Although the signal-to-noise ratio is generally higher in temperature hindcasts than for other variables such as PR or SLP (Smith et al., 2019), the global and dynamically downscaled decadal experiments analyzed in this Thesis also show the existence of the signal-to-noise paradox in T_{max} forecasts (Figure 5.3). In this case, the RPC spatial distributions show lower values than those found for PR (Figure 4.5), but they are still significantly larger than 1 (the ideal value; see Section 3.2). The paradox is stronger in the downscaled hindcasts than in their global counterparts. RPC values above 2 has been found mainly in the regions of the southwestern quarter of the IP for the WRF-DPLE T_{max}. The prevalence of the unpredictable background noise in the downscaled forecasts might explain the absence of robustness in part of the ACC results showed above. Note that a high RPC does not necessarily implies a total lack of statistical significance for correlations with observations. Indeed, a high ACC with observations would really contribute to increasing the RPC value if the model signal-to-noise ratio is small (see Eq. [3.27]). For example, regions over the Guadalquivir Valley and the southwestern IP exhibit a high RPC (Figure 5.3b) along with significant and positive ACC values (Figure 5.1h). The presence of the signal-to-noise paradox indicates that the ensemble size needed to remove the model background noise is larger than it would be if the model signal-to-noise ratio were higher (Smith et al., 2019). The addition of more members to the WRF-DPLE ensemble may help to improve these results by reducing the noise, as shown by Reyers et al. (2019) or Sienz et al. (2016) for other temperature-related variables. However, the benefits of increasing the number of members would be lower in NSAT than in PR (Reyers et al., 2019), as could be expected because the signal-to-noise paradox for PR is more pronounced than for NSAT. Therefore, more members than in the case of PR would be needed to get an equivalent improvement in NSAT ACC values, which would imply the use of notably larger amounts of computing resources.

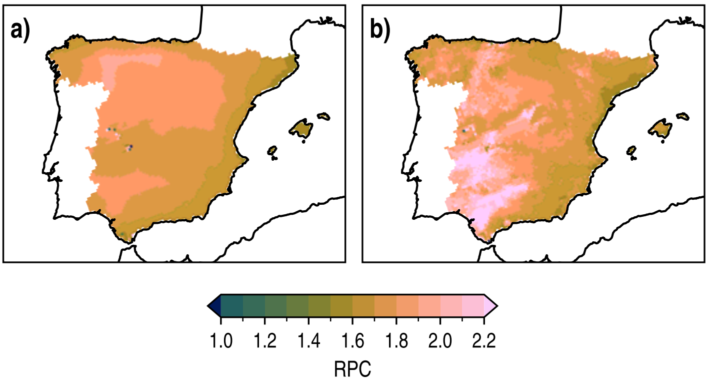

Figure 5.3: Signal-to-noise paradox in the hindcasts for T_{max} over the IP. **a**) Spatial distribution of RPC for the multiannual mean anomalies of the CESM-DPLE T_{max} in lead years 2–9 at annual scale. **b**) As **a**) but for WRF-DPLE. The absence (presence) of black dots indicates (not) statistically significant results different from 1 at the 90 % confidence level.

The spatial distributions of $\mathrm{MSSS_C}$ (see Eq. [3.24]), i.e., the MSSS calculated for WRF-DPLE with climatology as reference, are depicted in Figure 5.4. Generally, better results have been obtained for T_{max} than for PR (Figure 4.6). The influence of the climate change trends on temperature-based variables is also visible here. The results in lead year 1 show an IP mainly covered by positive but not significant results. The significant results found to the south of the Northern Subplateau, in Sierra Morena and some regions close to the Strait of Gibraltar, however, depict strong negative $\mathrm{MSSS_C}$ scores, with values even below -0.8. Note that these three regions also showed very high RMSE values in Figure 5.1 which cannot be attributed to a high T_{max} variability. $\mathrm{MSSS_C}$ results are slightly better in lead years 2–5. These clusters of very negative scores are more moderated and the magnitude of the positive values generally increases. Nevertheless, the regions with statistically significant positive results are very small. They are placed in the southern Mediterranean coast with values from 0.3 to 0.6. The worst results have been found in lead years 6–9, when the negative results continue dominating among the statistically significant scores, but in this case they cover larger portions of the IP than in lead years 1 and 2–5. On the other hand, the best results are shown in lead years 2–9. Positive $\mathrm{MSSS_C}$ values span most part of the domain, although those with statistical significance are reduced to small areas in the northwest, northeast, southwest and some regions in the Baetic System. The highest scores are observed in the northern regions, with values starting from 0.6 and surpassing 0.8. With regard to the significant negative results, in addition to the aforementioned three regions, which show scores below -0.8, there is another one close to the Cantabrian coast with similar outcomes.

According to Eq. [3.24], the results obtained for $\mathrm{MSSS_C}$ can be explained by those achieved for ACC and CB, depicted in Figure 5.1 and Figure 5.4, respectively. The spatial distribution of CB in lead year 1 shows important significant negative values in the same regions which show the worst results in terms of $\mathrm{MSSS_C}$. These results for CB are a consequence of the imbalance between ACC and the ratio of the hindcast and observations standard deviations (see Eq. [3.17]), which is caused by the low correlations observed in Figure 5.1 over these regions and the hindcasts variance being slightly larger than for observations. A similar situation is observed for the other lead times. In general, positive $\mathrm{MSSS_C}$ values are found in regions which also have a small CB. The largest areas covered by not significant CB values are mostly observed in the eastern half of the domain for lead years 2–9, followed by lead years 2–5. At all lead times, the locations having significant results in $\mathrm{MSSS_C}$ are often the same which show high ACC and not significant CB values. See, for instance, the regions in the

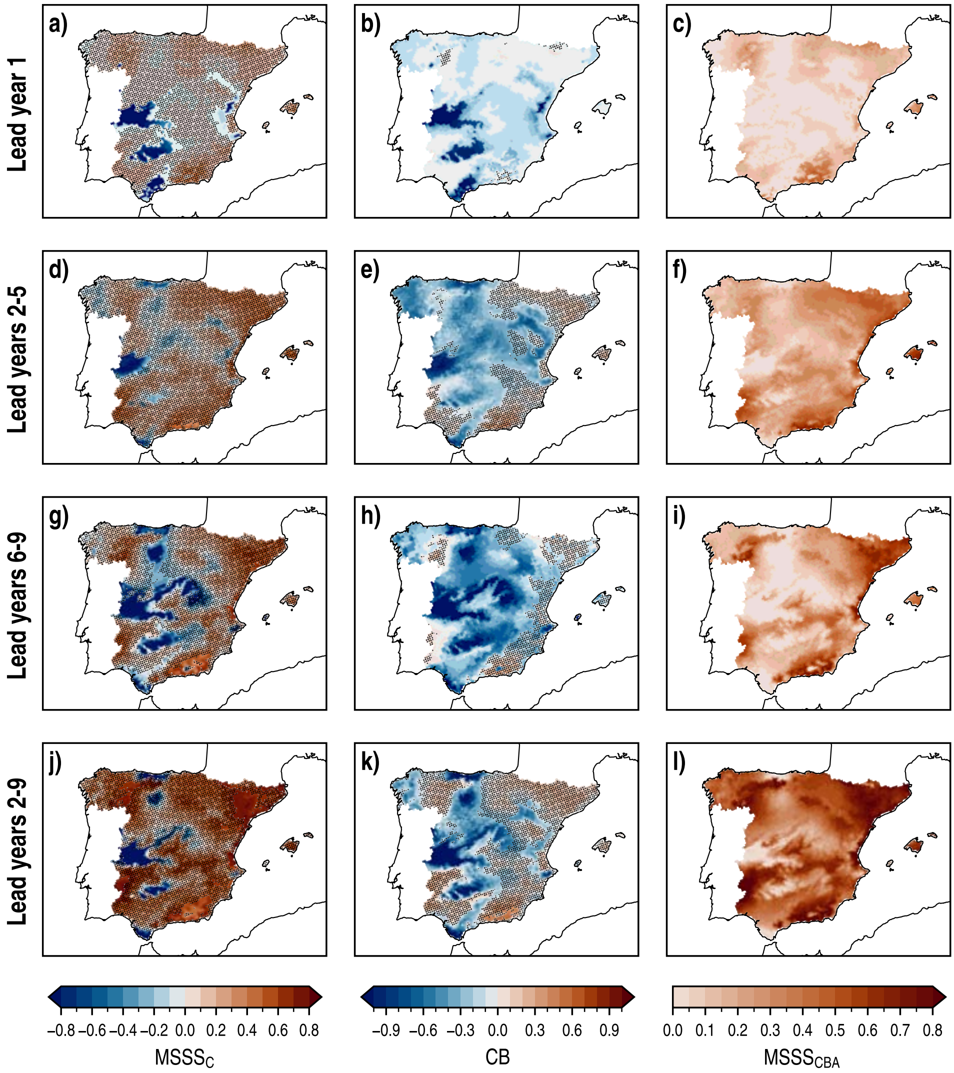

Figure 5.4: Spatial distributions of MSSS_C (left column), with climatology as reference, CB (center column) and the same MSSS_C calculated for lead time series with an adjusted CB, i.e., equal to zero (MSSS_CBA; right column), for the WRF-DPLE multiannual mean anomalies of T_max in lead years 1, 2–5, 6–9 and 2–9 (rows) at annual scale. In MSSS_C and CB maps, the absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

northwest or northeast in lead years 2–9. If the CB were completely removed from the lead time series (CB = 0 in Eq. [3.24]), the MSSS_C spatial distributions would look like the MSSS_CBA maps for the conditional bias-adjusted MSSS in Figure 5.4. Since the strongest negative CB values which cause the significant negative results for

$\mathrm{MSSS_C}$ are associated to very low ACC scores, no skill would be gained from the CB removal over those regions. Even though, there is a potential to continue improving the accuracy of the forecasts in regions which show low or moderate CB values if an effective conditional bias adjustment is carried out. Some of these regions are the Northern Subplateau, the Baetic System or the northeastern regions, for example.

At seasonal scale, the results achieved for $\mathrm{MSSS_C}$, CB and $\mathrm{MSSS_{CBA}}$ vary depending on the season (Figures B.14 to B.16, respectively, in Appendix B.2.1), but a similar analysis to that done at annual scale can be carried out. In this case, the $\mathrm{MSSS_C}$ results are also very influenced by the seasonal ACC scores (Figure 5.2), so the most promising results have been obtained for MAM, followed by JJA. Notwithstanding, seasonal $\mathrm{MSSS_C}$ is mainly not significant in these seasons, excepting the east side of the IP for lead years 2–9 in MAM.

❦ *Reliability analysis*

As part of the assessment of the predictive skill for T_{max}, the accuracy analysis must be complemented with the reliability analysis, which tests if the WRF-DPLE average ensemble spread $\overline{\sigma_Y^2}$ (Eq. [3.32]) is adequate to represent the uncertainty of the forecasts. A probabilistic metric, the CRPSS (Eq. [3.29]), has been used to address this task.

The spatial distributions of the CRPSS, calculated with the multiannual mean anomalies of T_{max} for each lead time, are depicted in Figure 5.5. As occurred for PR in Figure 4.7, very promising results have been obtained. Large portions of the domain are covered by not significant CRPSS values at all lead times, which is the ideal result for this metric (see Section 3.2). In lead year 1, the not significant results mainly span the northern half of the IP and some areas in the south over the Guadalquivir and Guadiana Valleys. Significant negative scores have been found in the same regions which showed very pessimistic RMSE and $\mathrm{MSSS_C}$ outcomes, with values below -0.05. The general picture is improved in lead years 2–5 with the increase of the amount of not significant results across the domain, but some significant negative clusters appear in the northeast and southeast of the IP. The absolute value of these negative results decreases in lead years 6–9, tuning into not significant values in some northeastern and southeastern locations. The best results are observed in lead years 2–9, when not significant CRPSS outcomes are found across a large part of the domain. The reliability of the hindcasts is closely related to the similitude between $\overline{\sigma_Y^2}$ and the squared standard error σ_X^2 (Eq. [3.33]). The LESS (Eq. [3.34]) reflects how they

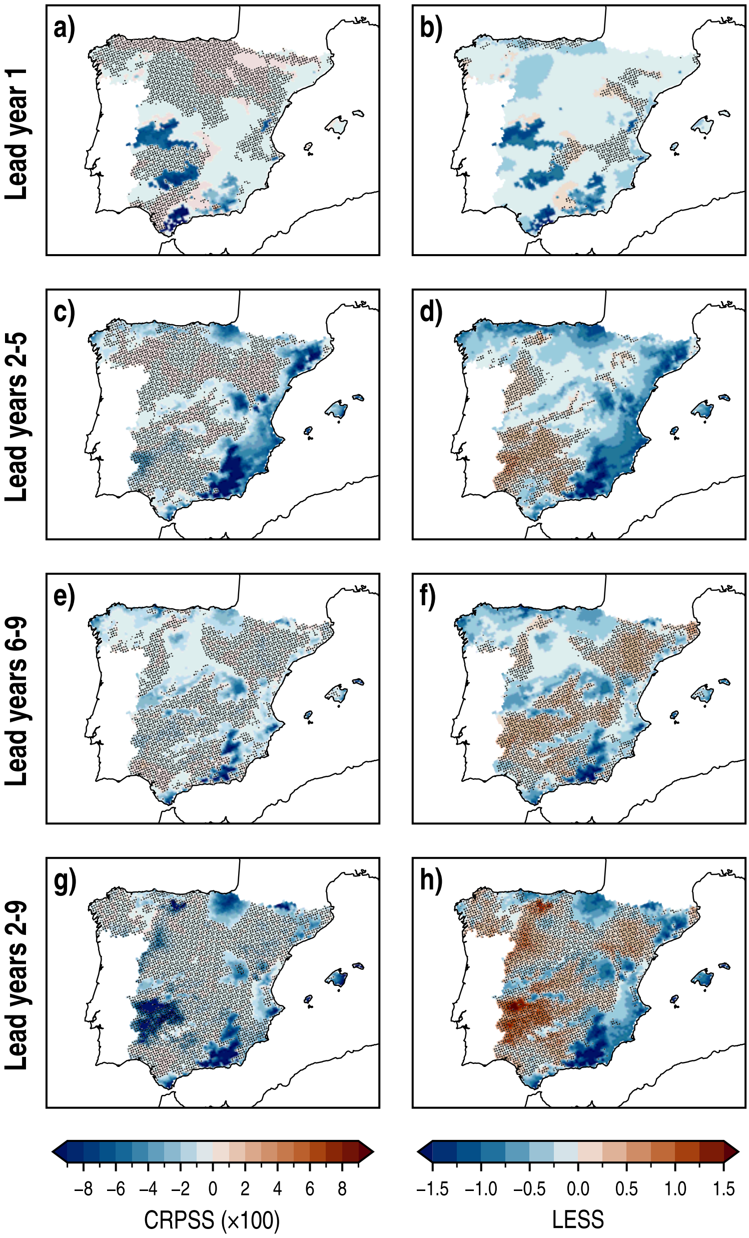

Figure 5.5: Spatial distributions of CRPSS (left column) and LESS (right column) for the WRF-DPLE multiannual mean anomalies of $T_{\max}$ in lead years 1, 2–5, 6–9 and 2–9 (rows) at annual scale. The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

are related and explains the results obtained for CRPSS. The influence of the LESS on the CRPSS distributions is evident because of the similarities between the lead time pairs of maps: the locations showing a not significant LESS also often has a not significant CRPSS. In general, there is a predominance of negative LESS at all lead times, with the exception of lead years 2–9, which shows that the hindcasts are usually underdispersive. The largest amount of regions showing overdispersion has been found in lead years 2–9, although it is also considerable in lead years 2–5 and 6–9, being typically not significant at all lead times. A few locations show a significant overdispersion in lead years 2–9 in the west and north of the IP, which is translated into significant results in CRPSS maps.

The results at seasonal scale for CRPSS and LESS can be consulted in Figures B.17 and B.18 in Appendix B.2.1. The best performance has been obtained in DJF and JJA, when the domain is mostly covered by not significant CRPSS results. However, if there is not predictive skill in terms of accuracy, as shown in Figure B.14 in DJF, the benefits provided by a not significant CRPSS are very limited. On the other hand, the season which exhibits the highest hindcast accuracy, MAM, only shows generalized not significant CRPSS outcomes in lead years 2–5. A lead year 1, there are also very promising results in the northern part of the domain, although they are restricted to smaller areas scattered along the IP in lead years 6–9 and 2–9.

5.1.2. *Comparison with the CESM-DPLE subensemble*

The results presented in the following have been obtained by comparing the performance of the WRF-DPLE ensemble, recalibrated with the DeFoReSt approach, to that of the CESM-DPLE, recalibrated with the same drift correction methodology applied on the ICs and LBCs used in DD simulations.

❦ *Accuracy analysis*

The spatial distributions of $\mathrm{MSSS_G}$ calculated for the multiannual mean anomalies of T_{max}, with the CESM-DPLE as reference, are depicted in Figure 5.6. At first sight, the added value provided by these downscaled hindcasts is clearly higher than in the case of PR (Section 4.2). The T_{max} is not only better represented than PR by the downscaled hindcasts with regard to skill to predict the observed climate, it is also when it is compared to the global hindcasts. In lead year 1, there are wide areas of significant positive $\mathrm{MSSS_G}$ scores along in the east of the IP, especially in the northeastern sector, with maximum values from 0.4 to 0.5. However, there are also

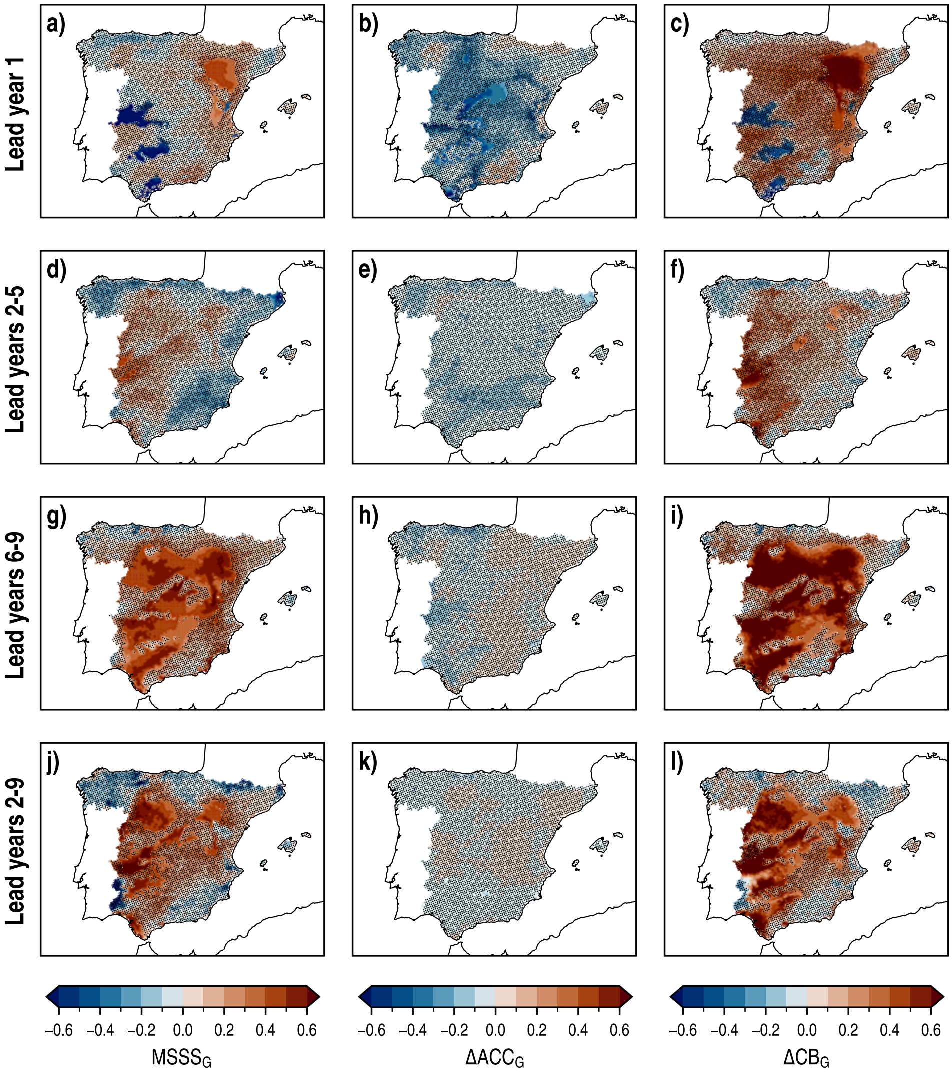

Figure 5.6: Spatial distributions of $MSSS_G$ (left column), ΔACC_G (center column) and ΔCB_G (right column), with CESM-DPLE as reference, for the WRF-DPLE multiannual mean anomalies of T_{max} in lead years 1, 2–5, 6–9 and 2–9 (rows) at annual scale. The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

three clusters of significant negative scores below -0.6, the same which showed very high errors in Section 5.1.1. The maximum significant positive $MSSS_G$ results in lead years 2–5 are concentrated in a central western area, with values ranging from 0.2 to 0.6. In this case, the significant negative outcomes spread across regions mostly close to the Mediterranean and the Cantabrian coasts. These values are mainly between -0.3 and -0.1, although there is a very small region in the northeast showing scores below -0.6. Very robust and positive results have been obtained in lead years 6–9. Excluding some regions close to the coast and some inner locations, $\mathrm{MSSS_G}$ shows a generalized added value of the downscaled hindcasts over the global experiments. These results are statistically significant and are mainly between 0.2 and 0.5. In lead years 2–9, the improvement provided by the downscaled hindcasts is generally more moderate than in lead years 6–9, since part of the significant positive results have turned into not significant and even negative in some areas. Nevertheless, the added value in terms of $\mathrm{MSSS_G}$ continues being predominant over large portions of the domain. There are some locations which show higher scores than in lead years 6–9, such as those situated over the central west of the IP, with values above 0.6.

These $\mathrm{MSSS_G}$ outcomes can be explained by the dependence they have on the ACC and CB calculated for the WRF-DPLE and CESM-DPLE hindcasts, which is described by Eq. [3.16]. The $\Delta\mathrm{CB_G}$ is generally positive at all lead times, which means that the CB magnitude for the downscaled hindcasts is lower than that for the GCM. There are important improvements of the CB in the northeastern regions for lead year 1, in the central west for lead years 2–5 and covering very large areas of the domain for lead years 6–9 and 2–9. The results obtained for $\Delta\mathrm{CB_G}$ are the main factor leading to the positive $\mathrm{MSSS_G}$ scores because the $\Delta\mathrm{ACC_G}$ maps mostly show not significant results, excepting for lead year 1. In this first year, the WRF-DPLE hindcasts experience a degradation of the predictive skill in terms of ACC compared to the CESM-DPLE. Additionally, in the same regions where this occurs, the $\Delta\mathrm{CB_G}$ maps do not show significant values, thus the $\mathrm{MSSS_G}$ maps do not show any significant outcomes either.

Very promising results have also been obtained in the seasonal analysis of $\mathrm{MSSS_G}$ (Figure 5.7). Most part of the statistically significant results show an added value of the WRF-DPLE ensemble. This added value reaches the highest level in SON (despite being the season with the worst results in terms of $\mathrm{MSSS_C}$ and ACC), especially in lead years 6–9 and 2–9, when almost the whole domain is covered by significant values starting from 0.1 and surpassing 0.6 in certain locations. In general, the statistically significant results are predominantly positive. However, there are some regions that still show an important degradation of the predictive skill, such as, for example, those located over the Central System, the Iberian System and the Cantabrian Range in DJF for lead years 2–5, with $\mathrm{MSSS_G}$ values below -0.5. Likewise at annual scale, the positive $\mathrm{MSSS_G}$ scores are mainly caused by the positive outcomes achieved for the seasonal $\Delta\mathrm{CB_G}$, with the contribution of the seasonal $\Delta\mathrm{ACC_G}$ in some cases, although

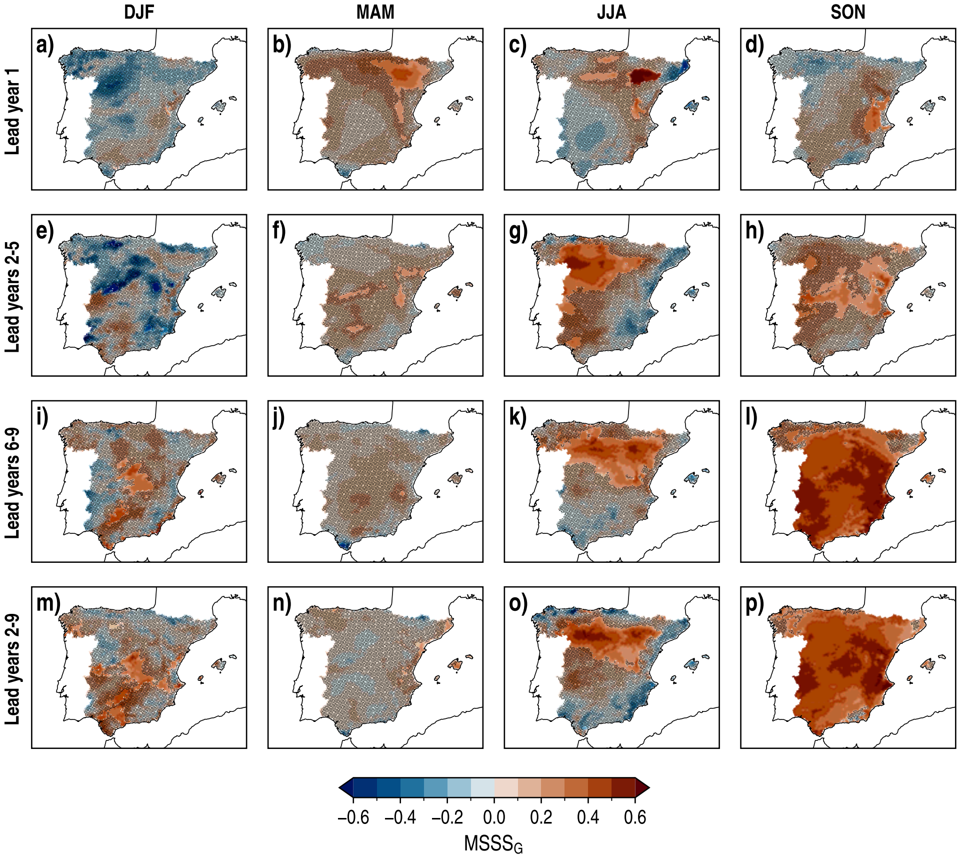

Figure 5.7 : Spatial distributions of $MSSS_G$ for the WRF-DPLE multiannual mean anomalies of T_{max}, with CESM-DPLE as reference, for lead years 1, 2–5, 6–9 and 2–9 (rows) in DJF, MAM, JJA and SON (columns). The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

the not significant changes in correlations are very frequent also at seasonal scale. The results for the seasonal ΔACC_G and ΔCB_G are depicted in Figures B.19 and B.20 in Appendix B.2.1, respectively.

❦ *Reliability analysis*

The extent to which the downscaled hindcasts improve or degrade the reliability of the CESM-DPLE subensemble is assessed by comparing the CRPSS obtained by both ensembles through the calculation of ΔCRPSS (Eq. [3.35]). The spatial distributions of ΔCRPSS for each lead time are depicted in Figure 5.8. As occurred

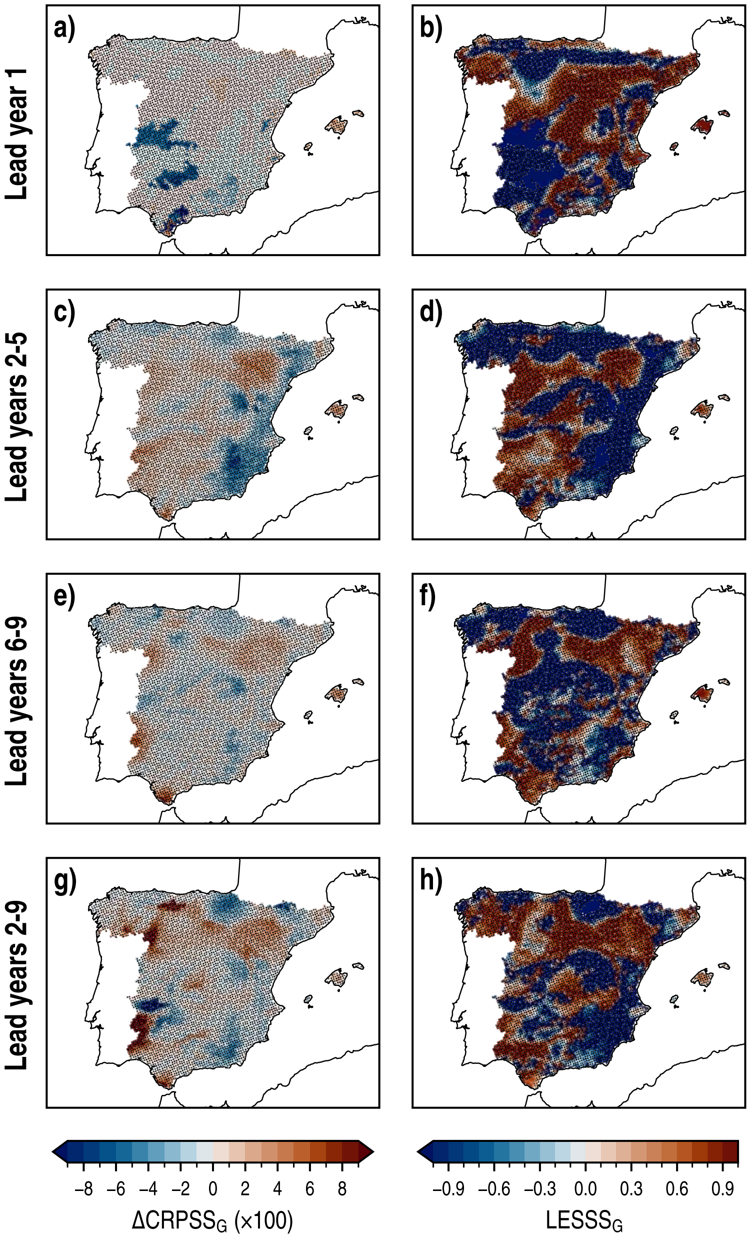

Figure 5.8: Spatial distributions of ΔCRPSS$_G$ (left column), and LESSS$_G$ (right column) for the WRF-DPLE multiannual mean anomalies of T_{max}, with CESM-DPLE as reference, in lead years 1, 2–5, 6–9 and 2–9 (rows) at annual scale. The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

for PR (Figure 4.10), the differences in terms of CRPSS are not large enough to get statistically significant results at any lead time window. The strongest negative scores are observed in lead year 1, with values below -0.05 over the same regions where the T_{max} hindcasts showed the worst accuracy results in the previous sections. Over the rest of the domain, the differences are very small in general. The highest scores have been found in lead years 2–9, when some locations situated in the northwest and southwest show maximum values above 0.09. The results of $\Delta\mathrm{CRPSS_G}$ can be explained by the degree to which the WRF-DPLE ensemble improves or degrades the average ensemble spread present in the CESM-DPLE hindcasts. How well the downscaled hindcasts perform in this respect is quantified by $\mathrm{LESSS_G}$ (Eq. [3.36]). In this line, very similar patterns can be observed between the $\Delta\mathrm{CRPSS_G}$ and $\mathrm{LESSS_G}$ pairs of maps at all lead times, showing that a better representation of the spread leads to a better representation of the forecast uncertainty by the ensemble. The $\mathrm{LESSS_G}$ results highly depend on the lead time, although there are some regions which always show an added value of the downscaled hindcasts, such as some locations in the Northern Subplateau or in the northeast. As found for PR in Figure 4.10, there is a strong presence of very extreme positive and negative results, which can be related to the sensitivity of the metric to small changes in the LESS. In this case, the statistically significant results are mainly negative and are mostly found in lead year 1.

The results obtained in the seasonal analysis are depicted in Figures B.21 and B.22 in Appendix B.2.1. Although the results for both $\Delta\mathrm{CRPSS_G}$ and $\mathrm{LESSS_G}$ vary depending on the season and the lead time, the relationship between both scores is the same described above. The best results in terms of $\Delta\mathrm{CRPSS_G}$ have been obtained in JJA (lead years 2–5, 6–9 and 2–9) and SON (lead years 1, 6–9 and 2–9), when the positive results predominate over the negative scores. However, the statistical significance is again almost nonexistent.

5.1.3. *Comparison with the WRF-LE ensemble*

When comparing the performances between the WRF-DPLE and WRF-LE ensembles, only the accuracy analysis is conducted since the uninitialized simulations are not probabilistic in essence, as the DCPs are (see Section 1.1.2). In addition, note that the period simulated to generate the WRF-LE ensemble only spans the interval 1990–2005 because the data needed to provide the RCM with ICs and LBCs are only available for these years. As in the analysis done for PR in Section 4.3, the sample of start dates analyzed in this section is only composed of those initialized every year from 1990 to 1999 (10 start dates), and only the lead years 1 and 2–5 have been examined. The large reduction of start dates compared to the analyses shown above leads to take these results with caution, since they may be affected by a large sampling bias.

❦ *Accuracy analysis*

The spatial distributions of the MSSS_U calculated for the WRF-DPLE multiannual mean anomalies of T_max, with WRF-LE as reference, are available in the Figure 5.9. There are large differences between the added value provided by the downscaled initialized experiments over the uninitialized counterpart in lead years 1 and 2–5. While the results are mostly positive (but not significant) in the first year of the decade, a very wide fraction of the IP is covered by significant negative results with values below -0.8 in lead years 2–5. The regions showing significant positive results in lead year 1 are smaller and are located in the southern part of the IP. The MSSS_U values in these locations range from 0.4 to 0.7. There are also significant positive results in lead years 2–5, but they are much scarcer than the negative values. These are observed over the Baetic System in the south, with values around 0.5. The added value of the hindcasts to the results in terms of CB is the main responsible factor leading to

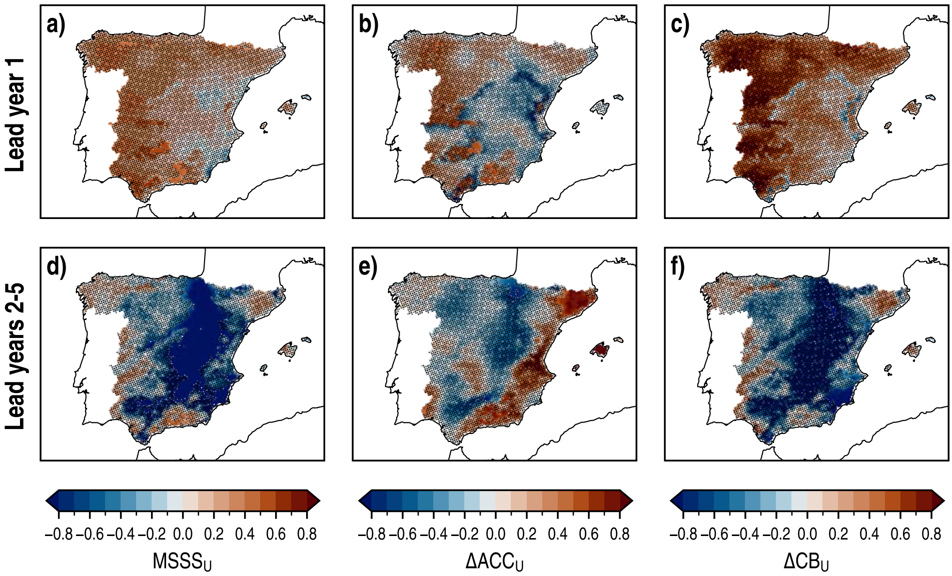

Figure 5.9: Spatial distributions of MSSS_U (left column), $\Delta\text{ACC}_\text{U}$ (center column) and ΔCB_U (right column) for the WRF-DPLE multiannual mean anomalies of T_max, with WRF-LE as reference, in lead years 1, 2–5, 6–9 and 2–9 (rows) at annual scale. The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

the positive results observed in MSSS_U map for lead year 1, which counteracts the degradation in ACC observed mainly in eastern and also southern regions. At the same lead time, there is a confluence of statistically significant positive $\Delta\text{ACC}_\text{U}$ and high ΔCB_U values which cause the aforementioned significant results observed for MSSS_U. The same influence of ΔCB_U on MSSS_U is observed in lead year 2–5, but causing negative results in this case. Although the added value observed in terms of ACC over the southern regions positively contributes to obtaining significant results for MSSS_U, the high values of $\Delta\text{ACC}_\text{U}$ achieved in the northeast, mostly above 0.7, are not enough to produce significant positive MSSS_U scores in that sector.

At seasonal scale, the best results for MSSS_U have been obtained in MAM (Figure 5.10). Significant positive outcomes span the northern part of the IP, with high values above 0.7 in the northeast. The rest of the domain is also covered by positive results, although they are not significant. A very similar spatial distribution is observed in lead years 2–5, but generally with lower scores. Part of the statistically significant results in the northwest has disappeared, but the values over the regions along the Mediterranean coast have become statistically significant. On the other hand, the most pessimistic outcomes have been found in DJF. In lead year 1, the domain is almost fully covered by negative scores, with values below -0.8 in some locations. The situation is slightly better for some western regions in lead years 2–5, where

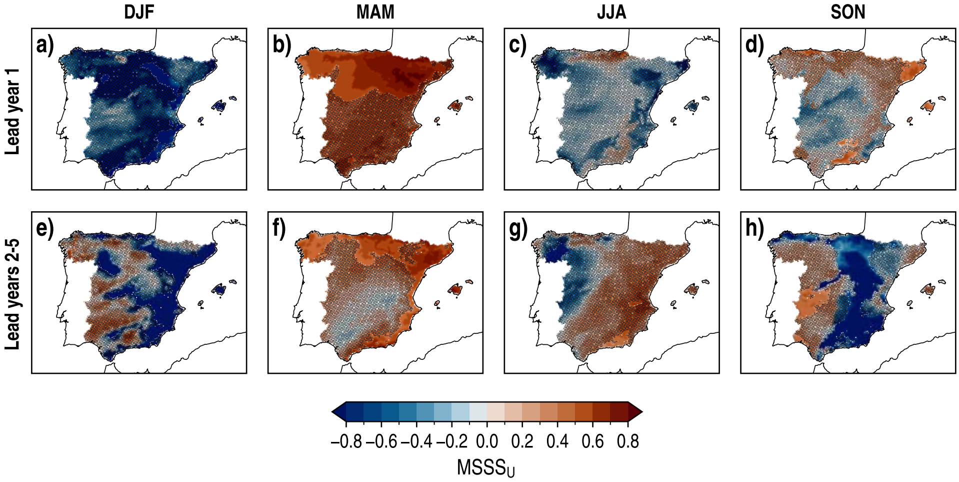

Figure 5.10: Spatial distributions of MSSS_U for the WRF-DPLE multiannual mean anomalies of T_{max}, with WRF-LE as reference, for lead years 1, 2–5, 6–9 and 2–9 (rows) in DJF, MAM, JJA and SON. The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

some positive results have been achieved, but mostly without statistical significance. The results in terms of seasonal $\Delta\mathrm{ACC_U}$ and $\Delta\mathrm{CB_U}$, which $\mathrm{MSSS_U}$ depends on, are available in Figures B.23 and B.24 in Appendix B.2.1.

5.1.4. *Predictive skill for regional averages*

The multiannual mean anomalies of T_{max} have been spatially averaged in every region resulting from the regionalization procedure applied over NSAT variables in Section 3.6 (see Figure 3.5b). Afterwards, a similar analysis to that done in the previous sections has been carried out to examine the predictive skill. Since the regionalization of the domain groups together those locations which share similar NSAT variability patterns, the skill scores calculated with these spatially averaged lead time series represent the general predictive skill achieved over a given region. The results of the analysis are summarized in Table 5.1.

As could be expected from the analysis done at a grid-point scale, the results obtained for the spatially averaged T_{max} are much more promising than for PR. The regional averages of T_{max} show a better ability to reproduce the observed variability, as indicated by the positive and often significant ACC scores. This improvement can be partially attributed to the typical positive trend observed in the lead time series of temperature-related variables (see Table A.2 in Appendix A.2). Moreover, there is a majority of regions which have achieved a CRPSS not significantly different from zero at least at three lead time windows, meaning that the hindcasts can be used to quantify the forecast uncertainty in these cases. Some of the best accuracy results have been found in the central south (CS) region, where significant positive ACC values have been obtained for all lead times. Futhermore, the CB almost always shows values not significantly different from zero in this region. A significant negative CB = −0.17 has been found only in lead year 1, where the ratio $s_{\{Y\}}/s_X$, whose value is below 1, cannot be counterbalanced by a lower ACC value, equal to 0.51 (see Eq. [3.17]). The CS region is also the only one which has a significant positive $\mathrm{MSSS_C}$, observed in lead years 2–9 with a value of 0.63. However, there is not any significant added value neither over the global hindcasts nor the uninitialized simulations, although some improvements can be observed especially in the latter case, mainly for correlations. In this region, the average ensemble spread $\overline{\sigma_Y^2}$ is able to represent the forecast uncertainty only in lead years 6–9 because the CRPSS values are significantly different from zero at other lead times. Another region that shows high accuracy scores is the high mountain (MT) region, which agglomerates some of the areas with the highest terrain elevation in the IP, especially in lead years 2–9. At this lead time, it has obtained the largest

Table 5.1: Skill scores for the spatially averaged WRF-DPLE multiannual mean anomalies of T_{max} in lead years 1, 2–5, 6–9 and 2–9 at annual scale. The subscripts C, G and U denote the reference data used to calculate the skill score: AEMET climatology, CESM-DPLE global hindcasts and WRF-LE uninitialized experiments, respectively. The bold formatting indicates results different from zero at the 90 % confidence level. Dashes denote data unavailability at that lead time.

Region	**Lead years**	**MSSS$_\mathbf{C}$**	**ACC**	**CB**	**CRPSS (×100)**	**MSSS$_\mathbf{G(U)}$**	**ΔACC$_\mathbf{G(U)}$**	**ΔCB$_\mathbf{G(U)}$**
SW	1	0.04	0.28	**-0.21**	**-0.08**	0.06 (0.43)	-0.13 (0.43)	0.23 (0.61)
	2-5	0.25	0.53	**-0.19**	-1.25	0.23 (-0.24)	-0.09 (-0.11)	0.41 (-0.32)
	6-9	0.13	**0.49**	**-0.32**	-0.79	**0.50** (–)	-0.05 (–)	**0.69** (–)
	2-9	0.61	**0.79**	**-0.10**	-9.43	**0.59** (–)	-0.02 (–)	**0.67** (–)
NO	1	0.12	0.35	**-0.06**	0.10	-0.10 (0.35)	-0.13 (0.23)	0.11 (0.55)
	2-5	0.15	0.47	**-0.27**	**-0.97**	-0.17 (-0.10)	-0.08 (-0.13)	-0.08 (-0.14)
	6-9	0.11	0.45	**-0.31**	**-1.10**	0.06 (–)	-0.05 (–)	0.14 (–)
	2-9	0.52	**0.73**	-0.12	-0.37	-0.12 (–)	-0.04 (–)	0.01 (–)
CI	1	0.03	0.23	**-0.14**	**-0.05**	0.04 (0.22)	-0.20 (0.07)	0.30 (0.44)
	2-5	0.13	0.47	**-0.29**	-0.13	0.06 (-1.00)	-0.08 (-0.28)	0.18 (-0.98)
	6-9	-0.16	0.41	**-0.57**	0.01	**0.45** (–)	0.01 (–)	**0.56** (–)
	2-9	0.36	**0.65**	**-0.26**	0.01	0.35 (–)	0.01 (–)	**0.37** (–)
NE	1	0.14	0.37	**-0.04**	0.03	0.15 (0.26)	-0.05 (0.11)	0.41 (0.54)
	2-5	0.41	0.64	-0.02	-0.86	-0.15 (-0.33)	-0.06 (0.23)	0.06 (-0.29)
	6-9	0.48	**0.71**	-0.14	-0.01	0.30 (–)	0.04 (–)	**0.28** (–)
	2-9	0.69	**0.84**	0.11	-0.06	-0.01 (–)	0.00 (–)	-0.05 (–)
CS	1	0.23	**0.51**	**-0.17**	**-0.92**	0.09 (0.35)	0.07 (0.40)	0.05 (0.17)
	2-5	0.44	**0.68**	0.15	**-4.46**	-0.22 (0.12)	-0.08 (0.57)	0.02 (-0.18)
	6-9	0.51	**0.72**	-0.01	0.30	0.14 (–)	0.04 (–)	0.12 (–)
	2-9	**0.63**	**0.83**	0.25	**-2.68**	-0.13 (–)	-0.01 (–)	-0.04 (–)
EA	1	0.06	0.28	**-0.14**	-0.02	0.18 (0.05)	-0.08 (-0.11)	0.38 (0.25)
	2-5	0.34	0.60	-0.14	**-1.90**	-0.21 (-0.79)	-0.08 (0.56)	-0.03 (-0.53)
	6-9	0.37	**0.65**	-0.22	0.03	0.40 (–)	0.02 (–)	**0.44** (–)
	2-9	0.62	**0.79**	-0.04	-0.24	0.10 (–)	-0.00 (–)	0.19 (–)
MT	1	0.15	0.40	**-0.10**	0.01	0.02 (0.38)	-0.09 (0.33)	0.23 (0.68)
	2-5	0.28	0.56	**-0.19**	0.10	-0.04 (-0.21)	-0.06 (-0.07)	0.10 (-0.25)
	6-9	0.45	**0.68**	**-0.13**	-3E-03	0.17 (–)	0.00 (–)	0.23 (–)
	2-9	0.75	**0.87**	0.06	-1.99	-0.10 (–)	-0.02 (–)	0.02 (–)
WI	1	0.07	0.28	**-0.10**	-0.07	-0.00 (0.38)	-0.21 (0.35)	0.31 (0.76)
	2-5	0.08	0.45	**-0.34**	0.05	0.15 (-0.20)	-0.05 (-0.22)	0.23 (-0.27)
	6-9	-0.15	0.32	**-0.51**	-0.08	**0.43** (–)	-0.05 (–)	**0.56** (–)
	2-9	0.37	**0.68**	**-0.32**	-0.89	0.42 (–)	0.00 (–)	**0.43** (–)

scores for MSSS_C and ACC (0.75 and 0.87, respectively), although the former without statistical significance. Additionally, the WRF-DPLE hindcasts also perform very well in terms of reliability at all lead times. Nevertheless, this region does not show any significant added value of WRF-DPLE neither over CESM-DPLE nor WRF-LE. The highest added value over the CESM-DPLE subensemble is observed in the south west (SW) region, particularly in lead years 6–9 and 2–9. Very robust results have been obtained for MSSS_G at these lead times, with significant values of 0.50 and 0.59, respectively, which have been caused by the large improvement observed in the CB. High ACC scores are also shown at these lead times and a CRPSS not significantly different from zero is found at almost all lead times, excepting in the first year of the decade. There are other regions such as the central interior (CI) or the western interior (WI), for example, that also show significant improvements in CB, which in some cases lead to an added value in terms of MSSS_G. There are not regions showing an added value of WRF-DPLE over the uninitialized WRF-LE ensemble in spite of high scores sometimes observed (e.g., in SW and lead year 1), which may be due to the small sample size composed of only 10 start dates.

The lead time series of the WRF-DPLE ensemble mean and AEMET in the CS region have been depicted in Figure 5.11. These representations include the 90 % confidence intervals associated to a single model member, which can be interpreted as a measure of the forecast uncertainty at those lead times when the downscaled hindcasts are reliable (for all excepting at lead year 1 in this region). As usual, the highest variance in the time series is observed for lead year 1, experiencing a progressive decrease as the length of the lead time window increases. The significant positive ACC results show the skill of the hindcasts to reproduce the observational variability. For example, the lead years 6–9 series of WRF-DPLE is able to replicate some of the relative minimums and maximums described by the AEMET series during the whole period. There is also relative skill to capture these peaks in lead years 2–5, although their magnitudes are commonly misrepresented. The same situation occurs to some degree in lead year 1 during the second half of the control period. The high correlations observed at all lead times are favoured by the positive trends obtained in both WRF-DPLE and AEMET series. However, despite these trends share the same sign, the AEMET T_max trends are underestimated by WRF-DPLE in most part of the regions at several lead times, as can be seen in Table A.2, available in Appendix A.2 (see Section 3.2 to consult the methodology applied to compute the trends). Although the differences between trends are typically not statistically significant, they may partially explain the overestimation of the observational T_max

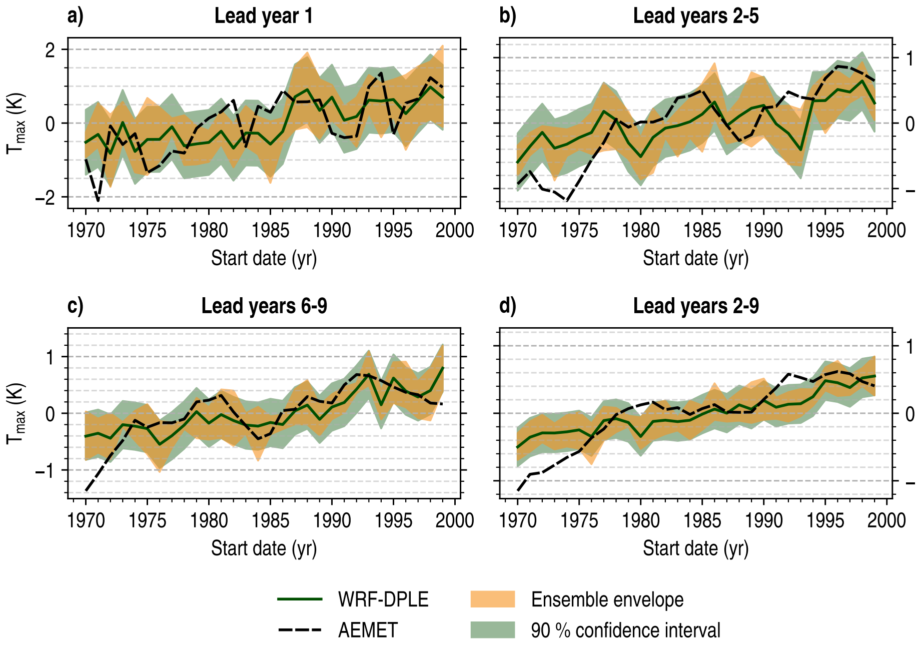

Figure 5.11: Time series of the spatially averaged multiannual mean anomalies of T_{max} in the CS region for lead years 1, 2–5, 6–9 and 2–9 at annual scale. Solid green lines identify the WRF-DPLE ensemble mean, whereas dashed black lines correspond to AEMET. Shaded green surfaces indicate the 90 % confidence interval for a WRF-DPLE single member, calculated from the average ensemble spread (Eq. [3.32]). Shaded yellow surfaces show the ensemble envelope which encloses the trajectories followed by the members composing the WRF-DPLE ensemble.

at the first start dates, shown for almost all regions, which generally decreases over the course of the control period. For instance, see Figure 5.11 for the CS region and Figures B.25 and B.26 in Appendix B.2.1 for a few additional examples. In the particular case of the CS region, this overestimation is more accentuated in lead years 2–5 and 2–9, when it is prolonged up to start years 1978 and 1976, respectively. Indeed, there are statistically significant differences between the trends of the WRF-DPLE and AEMET series in lead years 2–5, although they are not in lead years 2–9. The trends of the difference series are about -0.33 and -0.21 K/decade, respectively, with the negative sign indicating the underestimation of the AEMET trend by WRF-DPLE.

The lead time series of the CI region have been represented in Figure B.27 (Appendix B.2.1) with the purpose of showing the only case of significant positive differences between the WRF-DPLE and AEMET trends, found in lead years 6–9 with a value about 0.19 K/decade for the trend of the difference series, pointing to an overestimation of the observational trend. Although this region shows high added value in terms of accuracy over the global hindcasts at this lead time, as mentioned above, its performance is actually limited for MSSS_C due to its significant negative CB result. The CI region is also characterized by showing an overestimation of the AEMET T_{max} anomaly along the last start dates of the control period, mainly observed in lead years 6–9 (although it is also present, to a lesser extent, in lead years 2–5 and 2–9), which positively contributes to that overestimation of the observed trend.

In general, the width of the confidence intervals, relative to the magnitude of the signal, in these lead time series is smaller for T_{max} than for PR in all regions, especially in lead years 2–9. This finding is consistent with the results depicted in Figures 4.5 and 5.3 for the RPCs of PR and T_{max}, respectively, in lead years 2–9, which showed that the presence of the signal-to-noise paradox in PR hindcasts is stronger than in T_{max}.

5.2. Daily minimum near-surface air temperature

5.2.1. *Predictive skill of the WRF-DPLE ensemble*

❦ *Accuracy analysis*

The results obtained for the RMSE (Eq. [3.10]) calculated with the WRF-DPLE multiannual mean anomalies of T_{min} have been depicted in Figure 5.12. The highest RMSE values are generally found for lead year 1 in the west of the Northern Subplateau, with values above 0.6 K. Similar errors are observed in the Iberian System, but they are constrained to a smaller area. The lowest errors are shown along the Mediterranean coast, with values between 0.35 K and 0.4 K in the northeasternmost regions of the IP and between 0.4 K and 0.45 K in the southeast. Also in the southeast, there is a few locations over the Baetic System with RMSE outcomes above 0.7 K. These locations belong to the highest areas of Sierra Nevada, the place with the highest terrain elevation in the IP (see Figure 1.4a). As in the case of T_{max}, the concentration of the highest errors in these high mountain regions may be related to the challenges associated to the production of quality observational data in the zone (see Section 5.1.1). The RMSE generally decreases in lead years 2–5, and the structure of the spatial distribution is slightly different, in a qualitative sense, from that observed in lead year 1. The largest values are located again in the Baetic System, but the cluster in the west of the Northern Subplateau, which gathered high errors in lead year 1, has disappeared. The lowest values are observed in this case along the Ebro Valley, with errors around 0.25 K. This structure is essentially maintained in lead years 6–9 and 2–9, but with

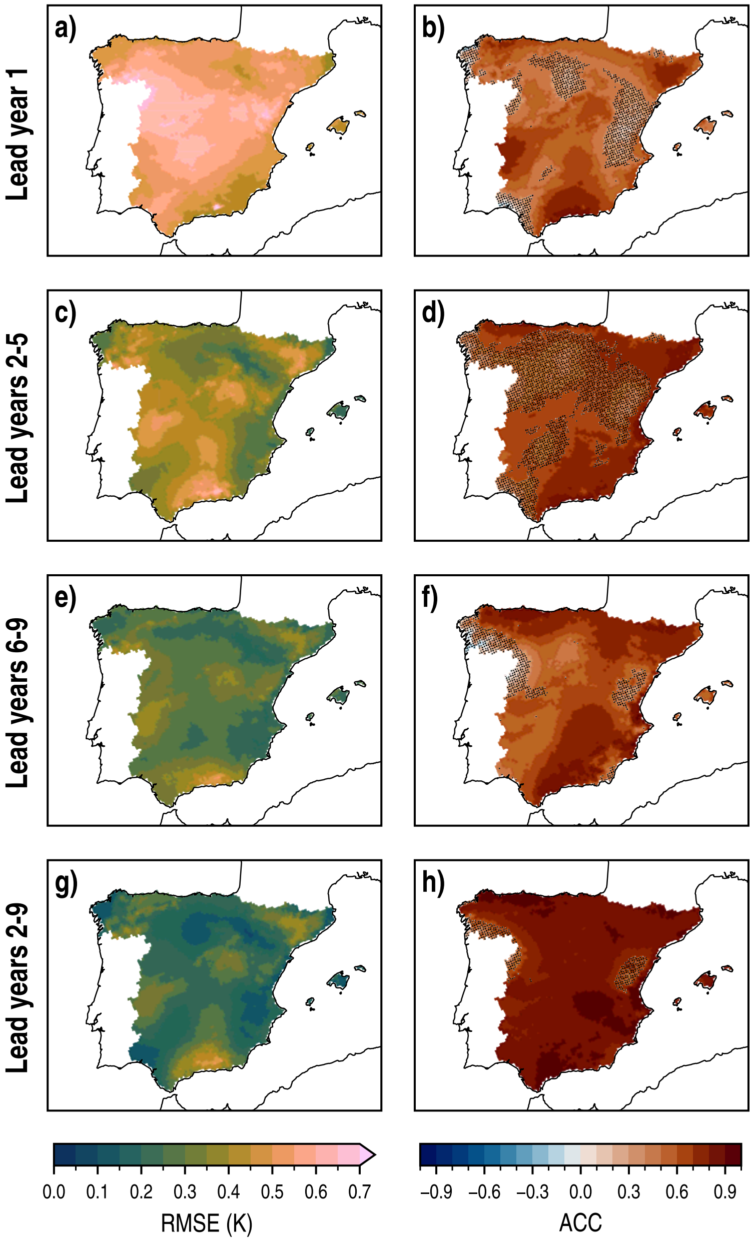

Figure 5.12: Spatial distributions of RMSE (left column) and ACC (right column) for the WRF-DPLE multiannual mean anomalies of T_{min} in lead years 1, 2–5, 6–9 and 2–9 (rows) at annual scale. In ACC maps, the absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

lower RMSE results.

In the same line of the results shown for T_{max} in Figure 5.1, the trend component which drives the evolution of the temperature-based variables is the main responsible of the high ACC (Eq. [3.13]) values observed for T_{min} and the differences between them and those for PR. In this case, the correlations are positive along almost the whole domain across all lead times, with the statistically significant results being predominant. A lead year 1, the highest values, which are above 0.7, are mainly present in southern, western and northeastern areas. There are some not significant results in the central east, in some northern regions and in the southwesternmost locations of the IP. The correlation coefficients are even larger in lead years 2–5. In this case, significant positive values above 0.7 span a large fraction of the domain. However, the area covered by not significant results is higher than that observed in lead year 1, and it is situated mainly in the northern part of the IP and in some smaller regions further south. Very high ACC values above 0.8 are located along the Mediterranean coast. Although the ACC values slightly decrease for lead years 6–9 in general compared to the previous lead time window, the not significant positive results are constrained to small regions close the Mediterranean coast and the northwest of the IP. The highest values are observed in lead years 2–9, when the climate trend component is better captured after filtering the interannual variability with the 8-year lead time averages. The spatial distribution shows a domain mostly covered by correlations above 0.8, with values even above 0.9 in the regions close to the Strait of Gibraltar, in the east and some northern locations. Not significant results are also observed in the same places as for lead years 6–9, but covering narrower areas in this case.

The seasonal RMSE results are depicted in Figure B.28, available in Appendix B.2.2. While the highest errors have been found in DJF, with values above 1.6 K in the Northern Subplateau, the lowest are obtained in MAM, which are generally below 0.8 K. As at annual scale, the RMSE decreases at the other lead times. The spatial distributions of RMSE generally resemble to that of the standard deviation of AEMET T_{min} available in Figure B.29, so the magnitude of the errors is strongly related to the magnitude of the observational signal.

With respect to the seasonal ACC (Figure 5.13), the best representation of the T_{min} variability has been found in MAM and JJA. In both seasons, the worst results can be observed in lead year 1, when the domain is generally covered by not significant outcomes. These results are negative in the northern part of the IP in JJA. The situation

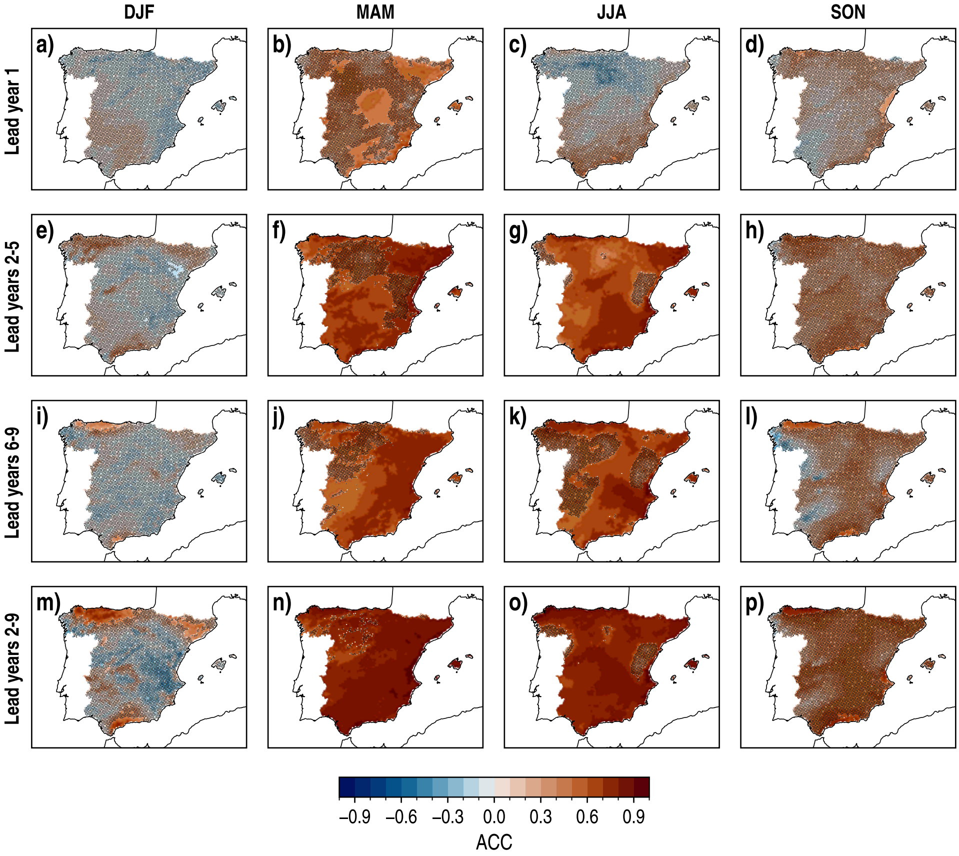

Figure 5.13: Spatial distributions of RMSE for the WRF-DPLE multiannual mean anomalies of T_{min} for lead years 1, 2-5, 6-9 and 2-9 (rows) in DJF, MAM, JJA and SON (columns). The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

changes at the other lead times. The spatial distributions mainly show significant positive results, which are frequently above 0.4. In lead years 2–5 and 6–9, there are not significant results mostly along the Northern Subplateau and some eastern regions close to the Mediterranean coast. The best results have been obtained in lead years 2–9 again, when the correlations are predominantly above 0.7 and the areas showing not significant results are constrained to very small locations in the northwestern flank (for MAM and JJA) and to the east (for JJA). In DJF and SON, the domain is largely covered by not significant results. However, it is worth remarking that, despite this lack of statistical significance, the correlations in SON are predominantly positive at all lead times, with the exception of some western and northeastern regions mainly in lead years 6–9. As occurred in Section 5.1.1, the results observed in MAM and JJA, better than in other seasons, are favoured by the stronger T_{min} trends found in MAM and JJA (see Tables A.3 to A.6 in Appendix A.2). The RPC (Eq. [3.27]) in lead years 2–9 is depicted in Figure 5.14. Although the values are still significantly higher than 1 over the whole domain for both WRF-DPLE and CESM-DPLE, the presence of the signal-to-noise paradox is not as strong here as in the PR (Figure 4.5) or T_{max} (Figure 5.3) downscaled hindcasts. In this case, the RPC spatial distribution generally shows lower values for WRF-DPLE than for CESM-DPLE. The signal-to-noise ratio is still low and the addition of new members to the ensemble would contribute to reducing the background noise and enhance the predictive skill, especially in terms of ACC, but this improvement achieved for the ACC scores, compared to the previous variables, may be partially due to this more discrete role of the signal-to-noise paradox.

The spatial distributions of the $\mathrm{MSSS_C}$ (Eq. [3.24]) for T_{min}, the MSSS calculated with the climatology as reference, are shown in Figure 5.15. As happened for PR (Figure 4.6) and T_{max} (Figure 5.4), there are large areas with not statistically significant results, covering almost the whole domain in lead years 1 and 2–5. Nevertheless, the results for T_{min} are essentially positive across all lead times. In lead year 1, there are small regions showing significant results with values ranging from 0.4 to 0.7 along the periphery of the domain in the northeast, south and west. In lead years 2–5, the significant values are slightly lower and are situated along the Mediterranean coast. The area covered by significant positive values is wider in lead years 6–9. In this case,

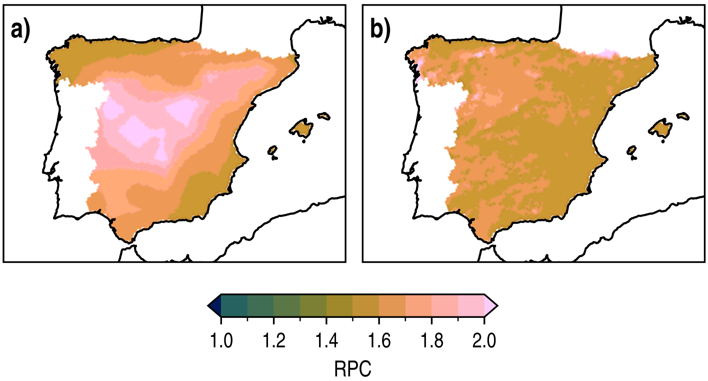

Figure 5.14: Signal-to-noise paradox in the hindcasts for T_{min} over the IP. **a**) Spatial distribution of RPC for the multiannual mean anomalies of the CESM-DPLE T_{min} in lead years 2–9 at annual scale. **b**) As **a**) but for WRF-DPLE. The absence (presence) of black dots indicates (not) statistically significant results different from 1 at the 90 % confidence level.

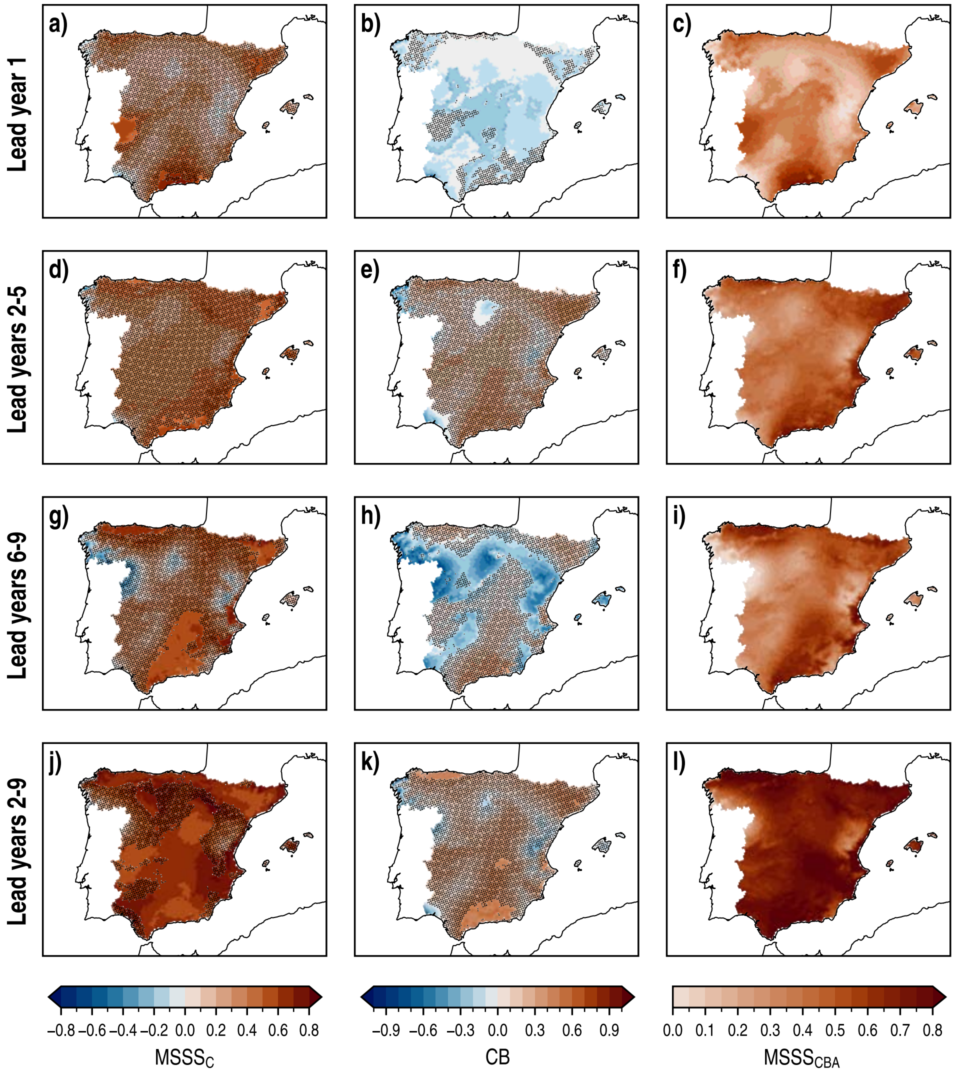

Figure 5.15: Spatial distributions of $\mathrm{MSSS_C}$ (left column), with climatology as reference, CB (center column) and the same $\mathrm{MSSS_C}$ calculated for lead time series with an adjusted CB, i.e., equal to zero ($\mathrm{MSSS_{CBA}}$; right column), for the WRF-DPLE multiannual mean anomalies of T_{min} in lead years 1, 2–5, 6–9 and 2–9 (rows) at annual scale. In $\mathrm{MSSS_C}$ and CB maps, the absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

there is a large region in the southern half of the domain with values ranging from 0.3 to 0.7. The positive scores bordering the Mediterranean Sea are higher than in lead year 2-5, with maximum values between 0.7 and 0.8. Another cluster of positive significant results is shown along the Cantabrian coast. There are also some locations which show significant negative results in the northwest, with values between -0.7 and -0.1. The best results have been found in lead years 2–9, when the IP is mostly covered by statistically significant positive results. These scores are mainly in the range 0.5–0.8, with the maximum values above 0.8 situated in small areas along the Mediterranean coast and the Ebro Valley.

The results obtained for $MSSS_C$ are a consequence of those shown for ACC and CB (Eq. [3.17]), as determined by Eq. [3.24]. The confluence of high ACC and close-to-zero CB values positively contributes to the increase of $MSSS_C$. The opposite occurs when a certain location presents low correlations and a large absolute CB. The significant CB values observed in Figure 5.15 are mainly negative, excepting in some northern and southern regions in lead years 2–9. This means that, in general, there is a significant imbalance between the correlation and the ratio between the hindcast and observational standard deviations in which the former is dominated by the latter, especially in those regions showing low correlations. The negative $MSSS_C$ values observed in lead years 6–9 are caused by the significant negative CB results and the low ACC scores found in the same region (Figure 5.12f). In the same vein, the low correlations, although positive and significant, along with the significant negative CB values in lead year 1, produce the not significant $MSSS_C$ results in Figure 5.15a. On the other hand, Figure 5.15j depicts very promising results for $MSSS_C$ because of the generally not significant CB shown in Figure 5.15k and the high correlations observed in Figure 5.12h. The $MSSS_{CBA}$ spatial distributions show the extent to which the predictive skill in terms of $MSSS_C$ could be enhanced by totally eliminating the CB. Some of the regions where the predictive skill could be improved the most are located in the south of the IP for lead years 2–9, where the $MSSS_C$ has values around 0.6 and the $MSSS_{CBA}$ shows results above 0.8. Others, such as the regions in the east of the IP in lead year 1, have absolute ACC values too small to experience any improvement with the correction of the CB.

The results obtained for $MSSS_C$ at seasonal scale have been depicted in Figure B.30 (Appendix B.2.2). As occurred for the seasonal ACC, the best results have been obtained in MAM and JJA, with the exception of lead year 1, because of the confluence of significant positive correlations and not significant CB values (Figure B.31). On the other hand, the negative seasonal values observed for $MSSS_C$ in DJF are ultimately motivated by the not significant correlations shown in Figure 5.12, which in turn cause the very negative seasonal CB values. The seasonal $MSSS_{CBA}$ (Figure B.32) spatial distributions show that the potential predictive skill which would be gained after adjusting the CB is practically nonexistent in DJF because of the same not significant correlations. Since the CB outcomes are mainly not significant in MAM and JJA, the contribution of such adjustment to improving the predictive skill is not huge, but it exists. In the Sourthern Subplateau, for example, there are regions where the values in the interval 0.6–0.7 increase up to 0.7–0.8 for lead years 2–9 in JJA. Some regions over the Central System also show values changing from 0.5–0.6 to 0.7–0.8 at the same lead time and season.

❦ *Reliability analysis*

The spatial distributions of the CRPSS calculated for the multiannual mean anomalies of T_{min}, needed to assess the reliability of the hindcasts, are depicted in Figure 5.16 (left column). The results obtained for this variable are not as promising as those for PR and T_{max} (Figures 4.7 and 5.5, respectively). The largest areas covered by not statistically significant results different from zero, which indicate that $\overline{\sigma_Y^2}$ is appropriate to quantify the forecast uncertainty, have been found in lead year 1 mainly in the northern part of the IP, in some southern regions and in the Balearic Islands. In general, the smallest CRPSS absolute values have been found at this lead time. In lead years 2–5 and 6–9, the amount of not significant outcomes decreases and the scores are generally higher in absolute value, especially in lead years 2–5. While the hindcasts are reliable in part of the southwestern sector of the IP in lead years 2–5, the regions with reliable hindcasts are mainly concentrated along a stripe of not significant CRPSS results spanning the Ebro Valley in lead years 6–9. In lead years 2–9, there are clusters of very negative results in the northeast, the inner and western central regions and over the Baetic System. On the other hand, the largest areas showing not significant outcomes are found in the east of the Northern Subplateau, the southwestern sector of the IP and some regions in the Mediterranean coast. The Ebro Valley and a few locations in the northwest also show not significant CRPSS outcomes.

As for the other variables, the reliability of the hindcasts is strongly influenced by the relation between $\overline{\sigma_Y^2}$ and σ_X^2, which is quantified by the LESS, depicted in the same Figure 5.16 (right column). The downscaled ensemble generally shows a robust underdispersion over almost the whole domain. The lowest values are observed in lead years 2–5 and 2–9. At the latter lead time, the minimum values are below -2 in some regions in the south and northeast, meaning that σ_X^2 is larger than $\overline{\sigma_Y^2}$ by more than a factor of 7. The regions with the closest-to-zero LESS values, sometimes still significant, are commonly those which show not significant CRPSS values at each lead time. Those regions generally having the most negative values of LESS, also

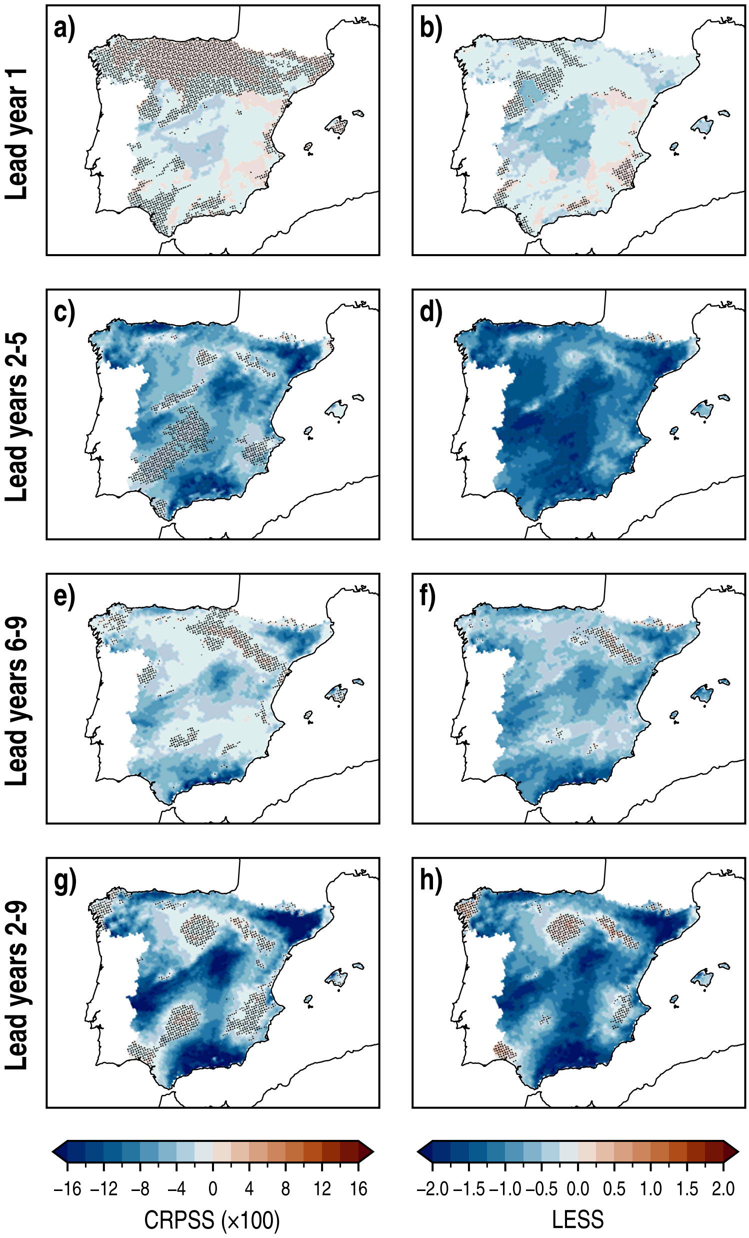

Figure 5.16: Spatial distributions of CRPSS (left column) and LESS (right column) for the WRF-DPLE multiannual mean anomalies of T_{min} in lead years 1, 2–5, 6–9 and 2–9 (rows) at annual scale. The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

have statistically significant CRPSS results different from zero.

Slightly better results have been found at seasonal scale in the analysis of the CRPSS (Figure B.33 in Appendix B.2.1). For example, not significant CRPSS cover large portions of the domain in JJA at all lead times. The LESS spatial distributions also show the most promising results in this season. The hindcasts continue being predominantly underdispersive also at seasonal scale, regardless of the season. In MAM and DJF for lead years 1 and 2–9, respectively, there are regions showing hindcast overdispersion, indicated by the positive LESS. The same is observed in some places in JJA for lead years 2–5 and 2–9. However, these results lack of statistical significance.

5.2.2. *Comparison with the CESM-DPLE subensemble*

❦ *Accuracy analysis*

After examining the predictive skill of the WRF-DPLE downscaled hindcasts, their performance has been compared with that of the CESM-DPLE subensemble. The results obtained for the accuracy scores have been depicted in Figure 5.17. In general, the spatial distributions of $\mathrm{MSSS_G}$ (Eq. [3.16]) at all lead times show the added value of the downscaled hindcasts to the predictive skill, although its magnitude depends on the lead time. Very high $\mathrm{MSSS_G}$ values are present in a large fraction of the IP for lead year 1. The significant positive values range from 0.3 to above 0.6, with the maximum scores situated in the central and northeastern parts of the IP. The northeastern cluster of high $\mathrm{MSSS_G}$ is maintained in lead years 2–5, but the amount of significant positive values drastically decreases. Some regions with not significant negative outcomes have been found in the north, west and southeast of the domain. The significant negative results are almost nonexistent, constrained to a few grid points in the north and south, with values below -0.2. As the lead time increases, the added value of the WRF-DPLE is higher. The regions having negative results in lead years 2–5 show higher scores in lead years 6–9, even turning into significant positive outcomes over the Northern Subplateau. The northeastern regions with significant positive outcomes are also present at this lead time and, along the northwest, experience the largest improvement compared to the CESM-DPLE hindcasts, with values above 0.5. The spatial distribution of $\mathrm{MSSS_G}$ in lead years 2–9 is similar to that observed in lead years 2–5, but with higher scores in absolute value. The northeastern regions, the east of the Northern Subplateau and the northwestern regions gather the highest results, with values above 0.6. The negative scores are mainly found in the southern half of

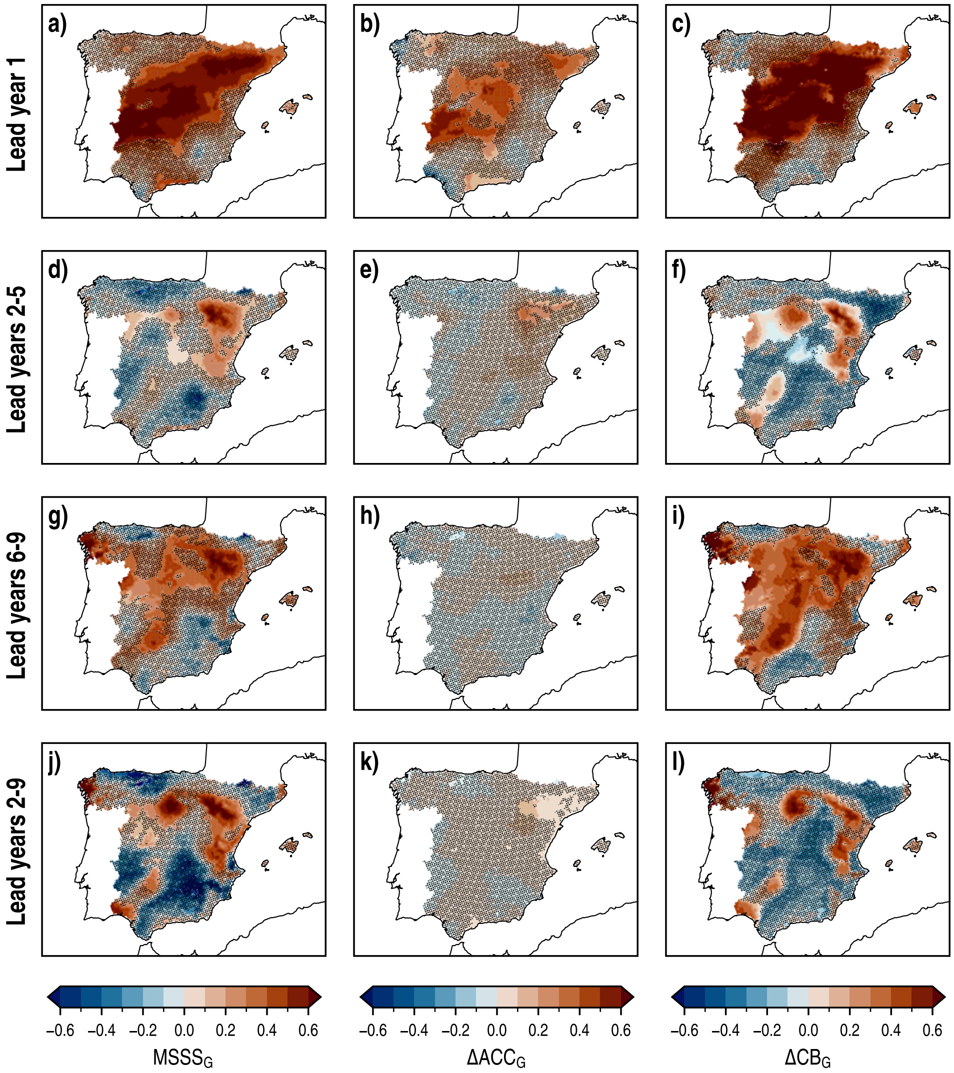

Figure 5.17: Spatial distributions of $MSSS_G$ (left column), ΔACC_G (center column) and ΔCB_G (right column), with CESM-DPLE as reference, for the WRF-DPLE multiannual mean anomalies of T_{min} in lead years 1, 2–5, 6–9 and 2–9 (rows) at annual scale. The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

the IP, along with some northern and northwestern regions, some with values below -0.6 in the Baetic System, but practically without statistical significance.

The relation described in Eq. [3.16] indicates that $MSSS_G$ is determined by the results obtained in the analysis of the added value of the downscaled hindcasts to both

ACC and CB with the global experiments as reference. There is a robust enhancement of the correlation in lead year 1, as shown in Figure 5.17b. Significant positive results are observed in the central regions of the IP and in the northeast, with values mainly above 0.3. The maximum added value to correlation is between 0.5 and 0.6, observed mostly to the south of the Central System. The results at the other lead times hardly show any significant improvement, only for lead years 2–5 and 2–9 in the northeast of the IP. The largest improvement in terms of CB, with values above 0.6 widely spread across the IP, is observed in lead year 1. These results, alongside those obtained for $\Delta \mathrm{ACC}_{\mathrm{G}}$, are the responsible of the large added value to the predictive skill in terms of the $\mathrm{MSSS}_{\mathrm{G}}$ observed in the first year of the decade. In the same line of the that observed for PR and T_{max}, the differences in the level of accuracy between the downscaled and global hindcasts are mainly determined by the differences between their spatial distributions of CB, especially as of lead years 2–5 onwards. This can be verified by comparing the $\mathrm{MSSS}_{\mathrm{G}}$ and $\Delta \mathrm{CB}_{\mathrm{G}}$ maps at each lead time, which closely match each other. For example, very good results for CB have been also found in lead years 6–9, when significant positive results cover a very large fraction of the domain (Figure 5.17i).

The spatial distributions of $\mathrm{MSSS}_{\mathrm{G}}$ calculated with the multiannual mean anomalies of T_{min} at seasonal scale have been depicted in Figure 5.18. Important improvements are mainly observed in DJF, when the domain is almost entirely covered by positive scores at all lead times. In this season, the largest areas with significant results have been found in lead years 1 and 2–5, whereas the maximum values above 0.6 are mainly shown in the northeast for lead years 2–9. In MAM, there is a very strong contribution of the downscaled hindcasts to the predictive skill, in comparison with the global experiments, in lead year 1. There is still some degradation of the predictive skill, although mainly not significant, at the other lead times. The positive $\mathrm{MSSS}_{\mathrm{G}}$ scores are predominant in SON, although not always with significant results. The most promising outcomes have been found for lead years 6–9 and 2–9 in this season. The northern half of the IP is mostly covered by significant positive values in both lead times, with other regions also showing significant and positive results in the southwestern IP for lead years 2–9. On the other hand, the extension of the areas with negative results is larger in JJA, although they are generally not significant. Almost all statistically significant scores observed in JJA are positive, and the best results have been found in lead years 6–9. Note that the performance of the WRF-DPLE in terms of $\mathrm{MSSS}_{\mathrm{C}}$ in DJF was certainly poor, whereas better results was found in MAM and JJA (Figure B.30). When the performance of the global hindcasts is suboptimal,

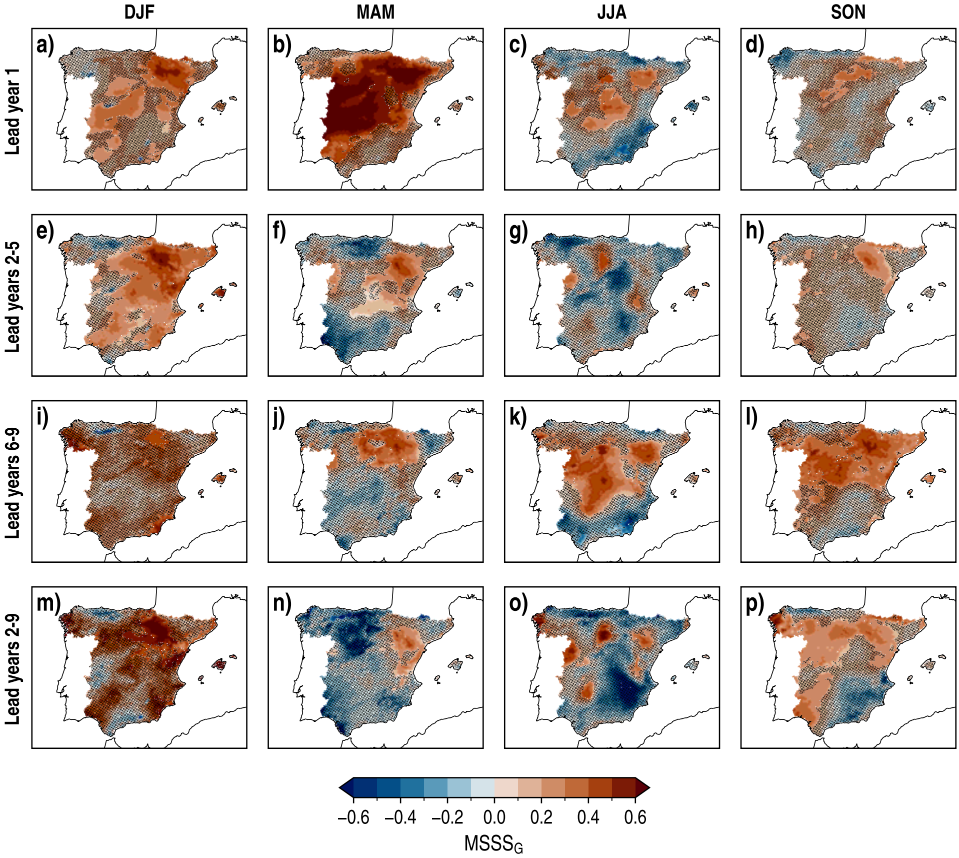

Figure 5.18: Spatial distributions of MSSS$_G$ for the WRF-DPLE multiannual mean anomalies of T_{min}, with CESM-DPLE as reference, for lead years 1, 2–5, 6–9 and 2–9 (rows) in DJF, MAM, JJA and SON (columns). The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

the downscaled hindcasts tend to be relatively more able to improve the predictive skill, both at annual and seasonal scale. Likewise at annual scale, the added value observed in MSSS$_G$ at seasonal scale is mainly motivated by the improvement in terms of CB (Figure B.35), since the results obtained for ΔACC$_G$ (Figure B.36) are mostly not significant.

❦ *Reliability analysis*

The spatial distributions of ΔCRPSS$_G$ are available in Figure 5.19 (left column). As usual, the differences between WRF-DPLE and CESM-DPLE in terms of CRPSS

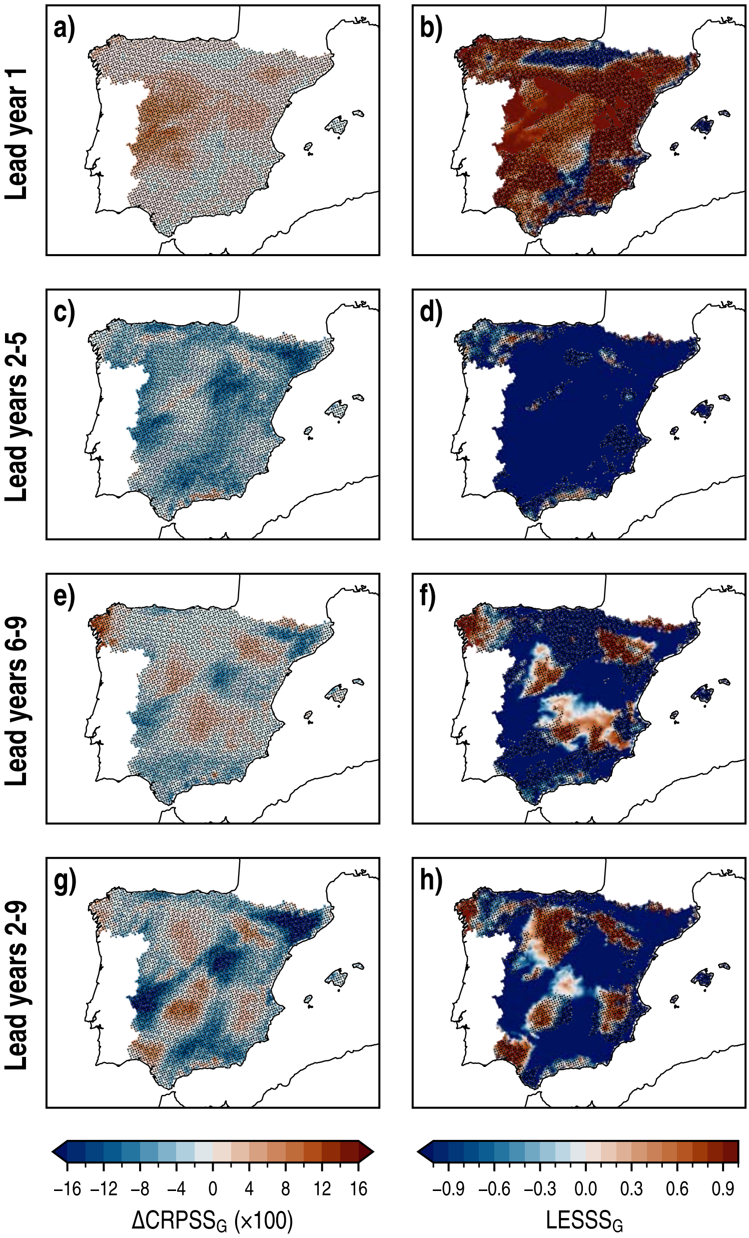

Figure 5.19: Spatial distributions of ΔCRPSS$_G$ (left column), and LESSS$_G$ (right column) for the WRF-DPLE multiannual mean anomalies of T_{min}, with CESM-DPLE as reference, in lead years 1, 2–5, 6–9 and 2–9 (rows). The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

are not large enough to observe statistically significant results different from zero. Positive results, indicating a better performance of the WRF-DPLE ensemble in this respect, are shown in lead year 1, mostly covering the whole IP, with the maximum values mainly in the central western part. At the other lead times, the negative results are predominant, especially in lead years 2–5, when they span over almost the whole domain. In lead years 6–9 and 2–9, there are some regions showing positive results. The highest added value of WRF-DPLE is observed in lead year 6–9 in the northwesternmost regions, with values above 0.08. A deterioration of the predictive skill, relative to CESM-DPLE, has been found in lead years 2–9 mainly in the northeast, with outcomes below -0.16. The distributions of LESSS_{G}, which indicate the extent to which the over- or underdispersion of the ensemble is corrected or enhanced in the downscaled experiments compared to the global hindcasts, explain the results obtained for the $\Delta\text{CRPSS}_{\text{G}}$. In lead year 1, there is a large improvement in the representation of the average ensemble variance, with significant positive values in the central and eastern part of the IP. These promising results lead to the generalized positive results observed in the $\Delta\text{CRPSS}_{\text{G}}$ maps, and the highest $\Delta\text{CRPSS}_{\text{G}}$ values are mainly shown in those regions where LESSS_{G} is significantly positive. There is a considerable deterioration in the representation of the ensemble spread in lead years 2–5, with scores mostly below -1 in the whole domain, explaining the negative $\Delta\text{CRPSS}_{\text{G}}$ results already mentioned. In the same line, the regions showing positive LESSS_{G} outcomes in lead years 6–9 and 2–9, are the same depicting an added value in terms of $\Delta\text{CRPSS}_{\text{G}}$. The opposed situation is shown for the negative values.

The results obtained for the $\Delta\text{CRPSS}_{\text{G}}$ and LESSS (Figures B.37 and B.38, respectively, in Appendix B.2.2) calculated with the multiannual mean anomalies of seasonal T_{min} slightly vary depending on the season, but the findings are in line with those discussed at annual scale. The differences between the downscaled and global CRPSS are not significant and their spatial distributions are driven by those obtained for the LESSS_{G}.

5.2.3. *Comparison with the WRF-LE ensemble*

The assessment of the predictive skill of the WRF-DPLE ensemble has been completed with the comparison between its performance and the performance of the uninitialized WRF-LE experiments. As mentioned in the analyses of PR and T_{max}, this assessment only includes the evaluation of the accuracy because the uninitialized experiments are not probabilistic, as the decadal predictions are.

❦ *Accuracy analysis*

The results for the $\mathrm{MSSS_U}$ calculated with the WRF-DPLE multiannual mean anomalies of T_{min}, considering WRF-LE as reference, are depicted in Figure 5.20. The added value provided by the downscaled hindcasts is mostly present in lead year 1. Significant positive results have been found in the southern part of the IP and over the Cantabrian Range and the Pyrenees, with values starting from 0.3. The maximum scores are observed in the southern regions along the Mediterranean coast, with values above 0.6. However, most part of the domain is covered by not significant results, with negative values mainly spanning the IP from the northwest to the eastern flank and spreading along the Mediterranean coast. In lead years 2–5, there is a generalized degradation of the predictive skill in the downscaled experiments. Significant negative results have been found in the northern half of the IP, in the Balearic Islands and in the southeasternmost regions, with values fundamentally lower than -0.6. There is an area with significant positive scores in the Pyrenees, with values above 0.6, but its area is very small.

The results obtained for $\mathrm{MSSS_U}$ are explained by the spatial distributions of

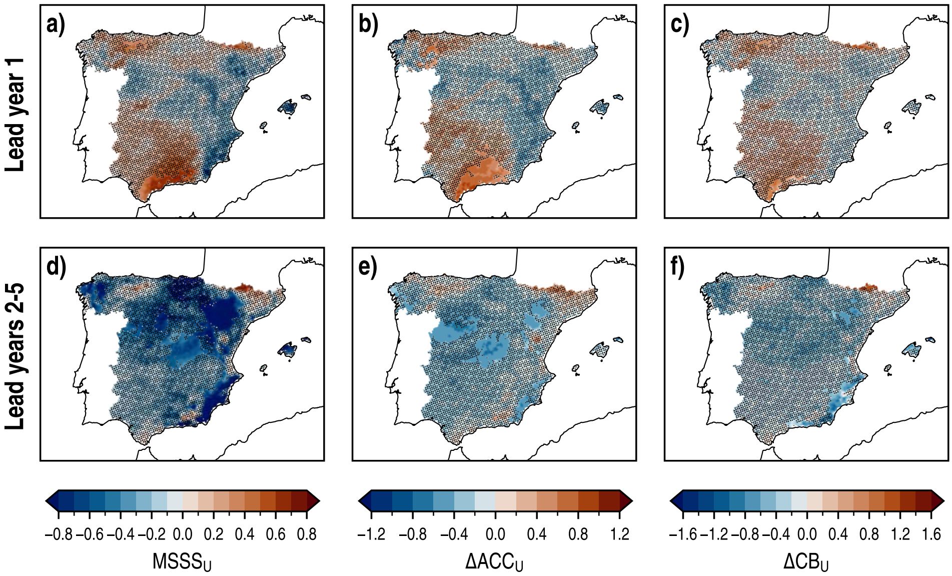

Figure 5.20: Spatial distributions of $\mathrm{MSSS_U}$ (left column), $\Delta\mathrm{ACC_U}$ and $\Delta\mathrm{CB_U}$ (right column) for the WRF-DPLE multiannual mean anomalies of T_{min}, with WRF-LE as reference, in lead years 1, 2–5, 6–9 and 2–9 (rows) at annual scale. The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

ΔACC$_U$ and ΔCB$_U$ also depicted in Figure 5.20, because of the relationship between these magnitudes determined by Eq. [3.16]. In lead year 1, there are positive values which indicate an added value of the hindcasts to ACC in the same regions which experience an added value in MSSS$_U$. Significant positive ΔACC$_U$ are shown in the Cantabrian Range and in the southern regions, with values mostly 0.6. The map of ΔCB$_U$ shows a very similar spatial distribution. Given the similarities between the ΔACC$_U$ and ΔCB$_U$ maps, the improvement in terms of CB, which depends on correlation and the ratio between the hindcast and observational standard deviations, is ultimately caused by the improvement found in ACC in most part of the domain. Therefore, the added value to predictive skill is in this case mainly driven by this added value to ACC.

The results obtained in the analysis at seasonal scale for MSSS$_U$ are shown in Figure 5.21. The most promising findings are observed in MAM, when the whole domain is covered by positive results, always above 0.4, in lead years 1 and 2–5. While large fraction of these results are not significant in lead year 1, they are fundamentally significant in lead years 2–5. The spatial distributions of the ΔACC$_U$ and ΔCB$_U$ in MAM (Figures B.39 and B.40, respectively, in Appendix B.2.2) show similar results, although the absence of statistical significance predominates for both at all lead times. In this case, not only the improvement in terms of ΔACC$_G$, but also the ratio between

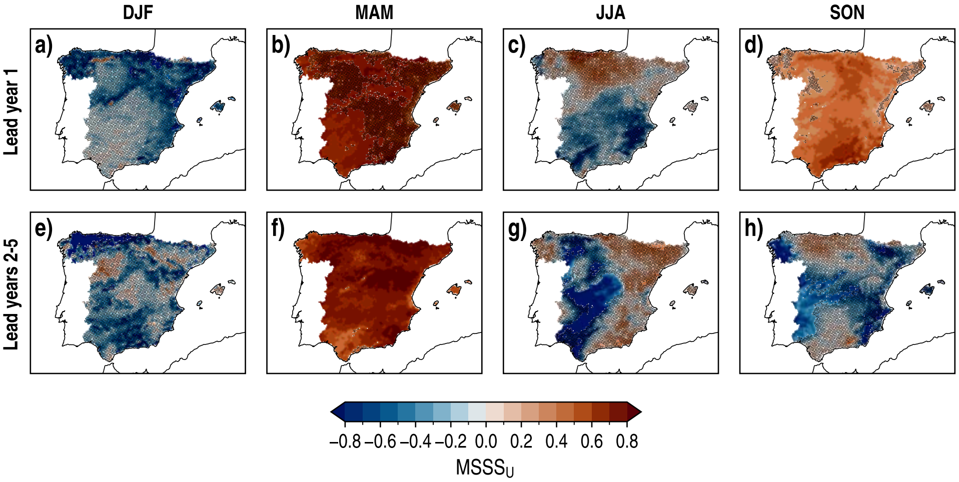

Figure 5.21: Spatial distributions of MSSS$_U$ for the WRF-DPLE multiannual mean anomalies of $T_{\min}$, with WRF-LE as reference, for lead years 1, 2–5, 6–9 and 2–9 (rows) in DJF, MAM, JJA and SON. The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

standard deviations contribute to improving the predictive skill via CB. For example, there are regions in the west showing CB improvements above 1.6 along with ACC improvements around 0.8. There are also regions showing significant negative results in $\mathrm{MSSS_U}$, being the most pronounced degradation of the predictive skill observed over the western sector in JJA for lead years 2–5, with values below -0.8.

5.2.4. *Predictive skill for regional averages*

The analysis of the predictive skill has been carried out also for the regions resulting from the regionalization process applied over NSAT in Section 3.6 (see Figure 3.5b). The lead time series have been spatially averaged in each region to afterwards calculate some of the skill scores used in the previous sections. The results are summarized in Table 5.2.

Excepting a few cases, the ACC generally shows significant positive results in all regions at all lead times, as could be expected from the outcomes depicted in Figure 5.12. These results imply an improvement compared to those observed for T_{max} and, especially, for PR (Tables 4.1 and 5.1, respectively). There are also positive $\mathrm{MSSS_C}$ values at all lead times for all regions. These scores are generally significant in lead years 2–9, with the exception of the WI region. Very good results have been also found in terms of CB, in which the not significant outcomes predominate over the significant ones. As opposed to that obtained for PR and T_{max}, there is a relative abundance of positive CB results in the case of T_{min}. This is a consequence of the high ACC scores obtained for this variable, which dominate the expression in Eq. [3.17] against the ratio $s_{\{Y\}}/s_X$. The analysis of the CRPSS reveals that the average ensemble spread $\overline{\sigma_Y^2}$ is generally not adequate to quantify the uncertainty of the forecasts, since most part of the regions show values significantly different from zero. The not significant results for CRPSS are mainly found in lead year 1, with the north (NO) region as the only one which shows not significant results also in lead years 6–9. The outcomes shown for the added value of WRF-DPLE over the global hindcasts and uninitialized experiments are in the line of those observed for PR and T_{max}, since there is a strong presence of not significant results. The northeast (NE) is the region which shows the best results in this respect, where positive values have been found for $\mathrm{MSSS_G}$ and $\Delta\mathrm{ACC_G}$ at all lead times, although they are only significant in lead years 1 and 2–5 for $\mathrm{MSSS_G}$.

One of the regions showing the best overall performance is the CS region, whose lead time series for the WRF-DPLE ensemble mean and AEMET have been repre-

Table 5.2: Skill scores for the spatially averaged WRF-DPLE multiannual mean anomalies of $T_{\min}$ in lead years 1, 2–5, 6–9 and 2–9 at annual scale. The subscripts C, G and U denote the reference data used to calculate the skill score: AEMET climatology, CESM-DPLE global hindcasts and WRF-LE uninitialized experiments, respectively. The bold formatting indicates results different from zero at the 90 % confidence level. Dashes denote data unavailability at that lead time.

Region	Lead years	MSSS$_{C}$	ACC	CB	CRPSS (×100)	MSSS$_{G(U)}$	ΔACC$_{G(U)}$	ΔCB$_{G(U)}$
SW	1	0.38	**0.62**	-0.09	**-0.22**	0.33 (0.24)	0.19 (0.42)	0.25 (0.44)
	2-5	0.40	**0.67**	0.21	**-5.24**	-0.10 (-0.26)	-0.01 (-0.31)	-0.19 (-0.28)
	6-9	0.47	**0.68**	0.03	**-3.50**	0.08 (–)	-0.01 (–)	0.22 (–)
	2-9	**0.69**	**0.88**	0.29	**-5.27**	-0.36 (–)	0.00 (–)	-0.24 (–)
NO	1	0.31	**0.56**	-0.04	0.14	0.14 (0.07)	0.09 (0.07)	0.08 (0.10)
	2-5	0.43	**0.68**	0.17	**-5.99**	-0.15 (-0.62)	-0.04 (-0.27)	-0.06 (-0.49)
	6-9	0.49	**0.71**	-0.09	-0.60	0.31 (–)	0.00 (–)	**0.39** (–)
	2-9	**0.70**	**0.86**	0.21	**-3.05**	-0.06 (–)	0.02 (–)	-0.16 (–)
CI	1	0.32	**0.59**	**-0.18**	**-0.90**	**0.52** (0.01)	0.33 (-0.01)	**0.52** (0.02)
	2-5	0.39	0.68	0.26	**-5.87**	0.02 (-0.44)	0.06 (-0.35)	-0.22 (-0.41)
	6-9	0.49	**0.70**	-0.03	**-2.51**	0.21 (–)	-0.01 (–)	0.36 (–)
	2-9	**0.66**	**0.87**	0.31	**-7.03**	-0.16 (–)	0.03 (–)	-0.29 (–)
NE	1	0.36	**0.60**	**-0.08**	-0.03	**0.50** (-0.09)	0.30 (-0.15)	**0.54** (-0.07)
	2-5	0.49	**0.77**	0.33	**-6.72**	**0.18** (-0.44)	0.16 (-0.14)	-0.26 (-0.32)
	6-9	0.56	**0.75**	-0.01	**-2.92**	0.29 (–)	0.02 (–)	0.37 (–)
	2-9	**0.68**	**0.87**	0.28	**-9.00**	0.06 (–)	0.06 (–)	-0.26 (–)
CS	1	0.58	**0.77**	-0.07	0.02	0.23 (0.53)	0.09 (**0.49**)	-0.06 (0.28)
	2-5	0.51	**0.80**	0.37	**-10.74**	-0.15 (-0.27)	-0.02 (-0.06)	-0.04 (-0.28)
	6-9	**0.58**	**0.80**	0.24	**-5.89**	-0.11 (–)	0.00 (–)	-0.14 (–)
	2-9	**0.63**	**0.90**	**0.43**	**-13.63**	-0.24 (–)	0.00 (–)	-0.10 (–)
EA	1	0.22	**0.48**	**-0.11**	**-0.02**	0.23 (-0.26)	0.10 (-0.30)	0.29 (-0.15)
	2-5	0.50	**0.73**	0.19	**-4.32**	0.12 (-0.51)	0.07 (-0.21)	-0.13 (-0.39)
	6-9	0.46	**0.71**	-0.20	**-1.06**	0.21 (–)	-0.04 (–)	0.30 (–)
	2-9	**0.74**	**0.87**	0.11	**-1.82**	0.19 (–)	0.03 (–)	0.06 (–)
MT	1	0.33	**0.58**	**-0.08**	0.04	0.23 (0.31)	0.14 (0.29)	0.17 (0.44)
	2-5	0.42	0.67	0.15	**-3.04**	-0.11 (-0.13)	-0.03 (0.01)	-0.07 (-0.12)
	6-9	0.50	**0.71**	-0.06	**-0.80**	0.09 (–)	-0.02 (–)	0.24 (–)
	2-9	**0.68**	**0.86**	0.24	**-3.13**	-0.08 (–)	0.02 (–)	-0.19 (–)
WI	1	0.22	**0.48**	**-0.11**	-0.18	0.35 (-0.06)	0.23 (-0.13)	0.41 (-0.08)
	2-5	0.33	0.58	0.10	**-5.03**	-0.04 (-0.54)	-0.02 (-0.45)	0.00 (-0.46)
	6-9	0.19	**0.54**	**-0.32**	**-1.41**	0.29 (–)	-0.01 (–)	**0.34** (–)
	2-9	0.65	**0.82**	0.12	**-3.38**	0.13 (–)	0.02 (–)	0.07 (–)

sented in Figure 5.22. The high ACC scores reflect the ability of the WRF-DPLE ensemble mean to reproduce the observed variability at all lead times. Although the marked positive trends contribute to enhancing these results, the downscaled hindcasts are also able to partially replicate some of the relative maximums and minimums described by the AEMET series. In lead year 1, for example, this ability is mainly found in the last third of the control period, but also at other start years such as 1973, 1975, 1979 or 1981. The magnitude of these peaks, however, is not always well represented. The lack of reliability in lead years 2–5, 6–9 and 2–9, highlighted by the results shown for CRPSS in Table 5.2, can be now observed here. There are clear differences between WRF-DPLE and AEMET trends which lead to an overestimation (underestimation) of the AEMET anomalies at the beginning (end) of the control period, especially in lead years 2–5 and 2–9. With the exception of lead year 1, the trends of the difference between WRF-DPLE and AEMET series are statistically significant

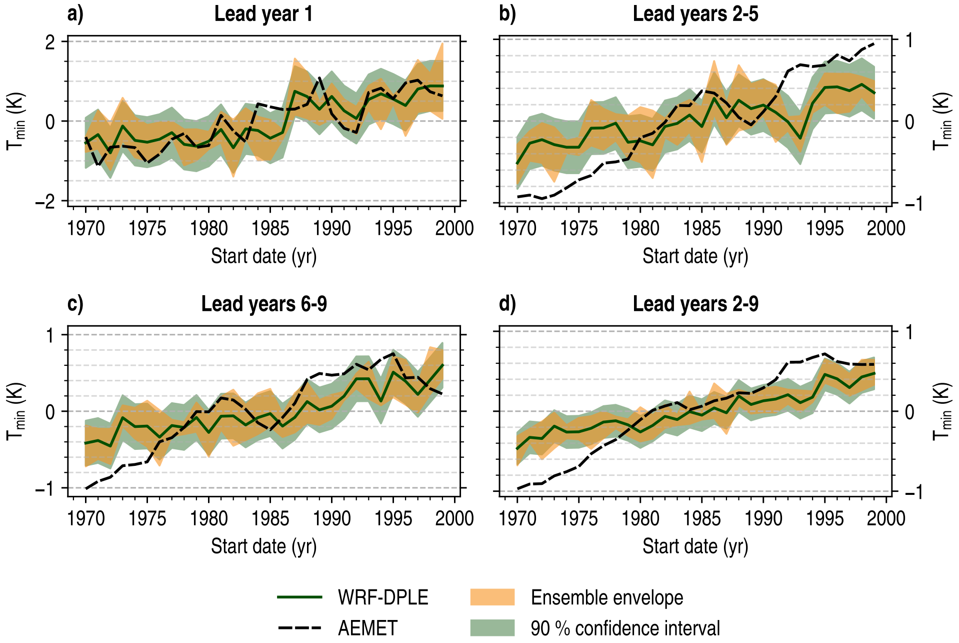

Figure 5.22: Time series of the spatially averaged multiannual mean anomalies of T_{min} in the CS region for lead years 1, 2–5, 6–9 and 2–9 at annual scale. Solid green lines identify the WRF-DPLE ensemble mean, whereas dashed black lines correspond to AEMET. Shaded green surfaces indicate the 90 % confidence interval for a WRF-DPLE single member, calculated from the average ensemble spread (Eq. [3.32]). Shaded yellow surfaces show the ensemble envelope which encloses the trajectories followed by the members composing the WRF-DPLE ensemble.

and have values of -0.42, -0.25 and -0.34 K/decade for lead years 2–5, 6–9 and 2–9, respectively (see Table A.2 in Appendix A.2 to consult the results and Section 3.2 for further information about the methodology applied to compute the trends). These results lead to values of the AEMET series clearly outside the confidence intervals of the hindcasts at the beginning and end of the control period (excepting for some start dates) in lead years 2–5 and 2–9. The impact that these discrepancies have on the predictive skill directly affects the results for CRPSS in the CS region, which shows the worst results among all regions at each lead time for this probabilistic score, with the exception of lead year 1.

The trends of the difference series are not so pronounced in other regions, although the general overestimation of the anomalies at the beginning of the control period, observed in the case of T_{max} in Section 5.1.4, is commonly observed also for T_{min}. For instance, see the lead time series for the NO region in Figure 5.23 and some additional examples in Figures B.41 and B.42 (Appendix B.2.2). In the NO region, the WRF-DPLE lead time series are closer to the AEMET series at the end of the control period than they were in the CS region. In this line, there is a reduction of the differences between trends, although they are still significant in lead years 2–5 (Table A.2). This

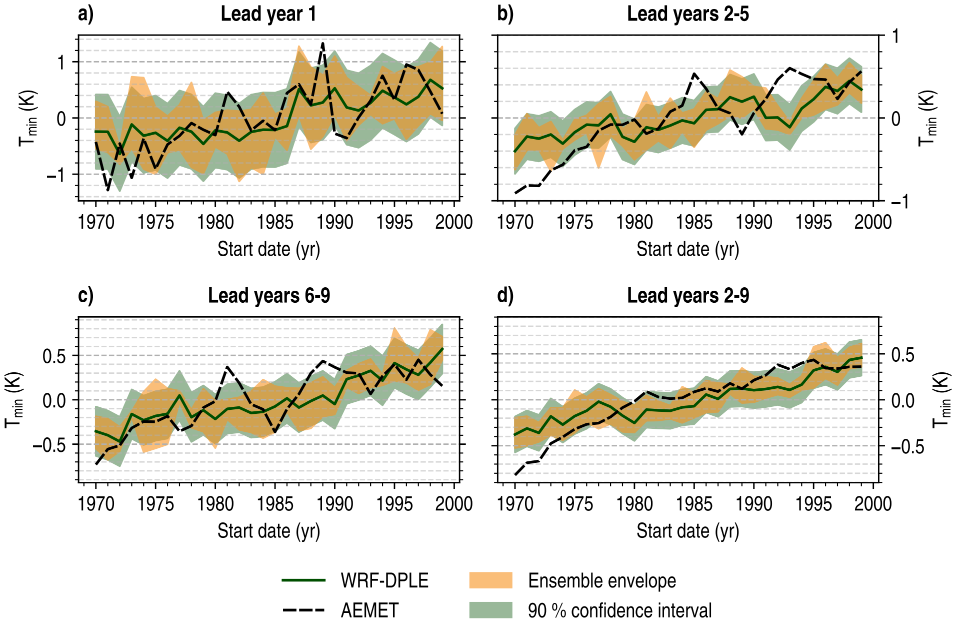

Figure 5.23: As Figure 5.22 but for the NO region.

improvement in the representation of the trends in the NO region, compared to the CS region, is transferred to the predictive skill in general, in terms of both accuracy and reliability. In the NO region, the CRPSS results are not significant for lead years 1 and 6–9, meaning that the hindcasts can be used to quantify the forecast uncertainty at these lead times. Note that the width of the confidence intervals in the T_{min} series, relative to the magnitude of the signal, is narrower than it was mainly in PR, but also in T_{max} to a lesser degree, especially in lead years 2–9. This could be anticipated by the RPC results displayed in Figures 4.5, 5.3 and 5.14, which showed a weaker signal-to-noise paradox in the WRF-DPLE hindcasts of T_{min} than in those of PR or T_{max}.

5.3. Daily mean near-surface air temperature

This Section is devoted to assess the WRF-DPLE predictive skill for T_{mean}. While the T_{mean} provided by WRF and CESM is calculated as the daily average of the instantaneous temperature at every model time step, the observational T_{mean} has been calculated by computing the arithmetic mean with the AEMET daily T_{max} and T_{min}, as stated in Section 2.3.

5.3.1. *Predictive skill of the WRF-DPLE ensemble*

The multiannual mean anomalies of T_{mean} have been used to calculate the spatial distributions of RMSE (Eq. [3.10]), which have been depicted in Figure 5.24 (left column). These distributions are, along side those for T_{max} and T_{min}, more homogeneous than those for PR. The highest errors, as usual, are observed in lead year 1 and decrease with the increase of length of the lead time window. For this lead year 1, the highest values have been observed in the mountain regions belonging to the Central, Iberian and Baetic Systems, with the maximum values above 0.7 K in the last two. The lowest errors have been found mainly along the Mediterranean coast and the southwestern regions, showing the latter the minimum values between 0.4 K and 0.45 K. The same spatial structure is maintained to some extent at the other lead times but with lower errors. Some dependence on lead time is observed when comparing the results in lead years 2–5 and 6–9, as the latter shows slightly lower errors, with differences generally about 0.1 K between them. The largest errors are shown over the Baetic System, with values around 0.65 K and 0.55 K in lead years 2–5 and 6–9, respectively. In lead years 2–9, there is a general decrease of RMSE around 0.1 K compared to lead years 6–9, although the maximum values about 0.55 K are still observed in the Baetic System.

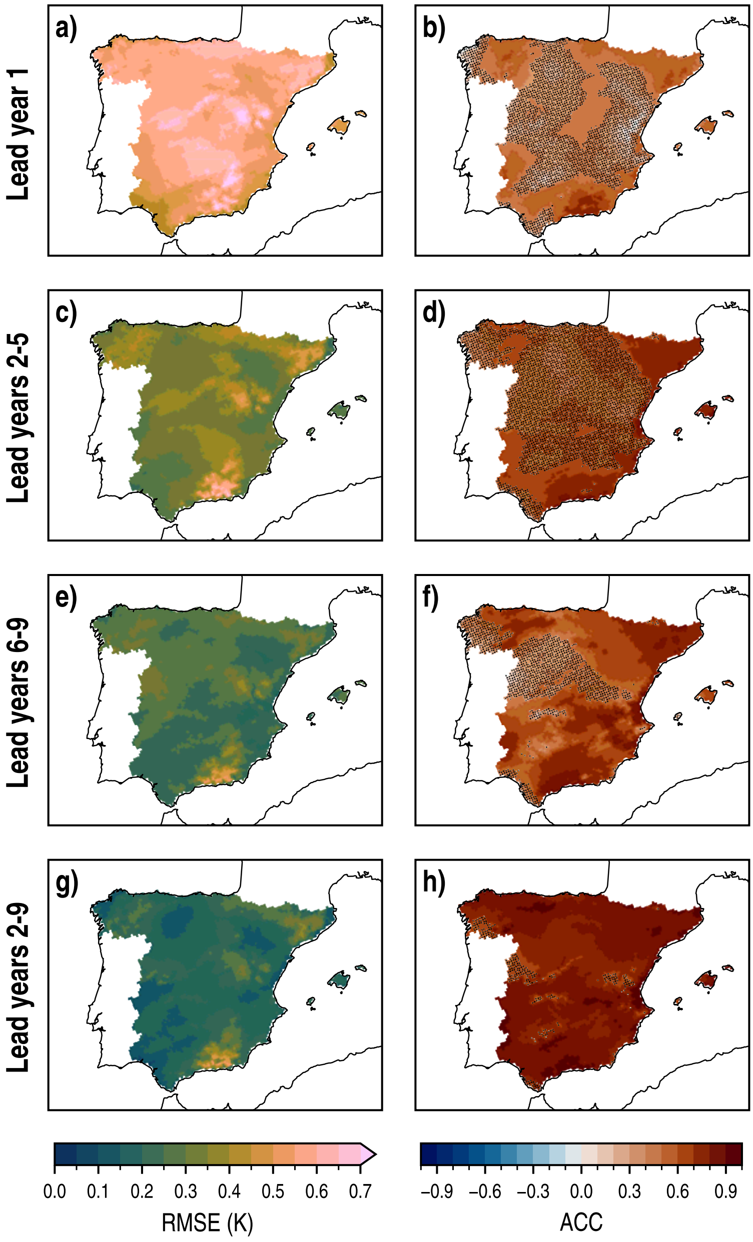

Figure 5.24: Spatial distributions of RMSE (left column) and ACC (right column) for the WRF-DPLE multiannual mean anomalies of T_{mean} in lead years 1, 2–5, 6–9 and 2–9 (rows). In ACC maps, the absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

The results obtained for ACC (Eq. [3.13]; Figure 5.24, right column) are in line with those achieved for T_{min} and, to a lesser extent, for T_{max}, in the sense that the generalized positive correlations are motivated by the positive climate trend present in temperature-based variables. The lowest correlations are observed in lead year 1, with significant positive values ranging from 0.4 to 0.8. The strongest correlations have been found over the Baetic System. On the other hand, there are large areas covered by not significant results across the IP from north to south. Although the significant correlations are higher in lead years 2–5, the surface covered by not significant results is also wider. Some southern regions, the Mediterranean coast, some locations to the northwest of the Northern Subplateau, the northeastern IP and the Balearic Islands are the areas which show statistical significant results, with values between 0.6 and 0.8. In lead years 6–9, the not significant results are observed only over the Northern Subplateau, the Iberian and Central Systems, some regions in the northwest and some smaller areas in the south. The significant positive ACC values are slightly higher than in lead years 2–5, with scores above 0.8 situated in regions close to the Mediterranean coast. The highest correlations are displayed in lead years 2–9, when the climate trend is better captured after eliminating the interannual and higher-frequency variability. The domain is almost fully covered by significant positive results, with values fundamentally above 0.7 and even 0.9 in some locations scattered throughout the IP. It is worth remarking that there are several regions with significant and positive results at all lead times. They are the southern and southwestern regions, an area in the northwest and another in the northeast of the IP.

The results achieved for the seasonal RMSE are available in Figure B.43 (Appendix B.2.3). As at annual scale, the errors are higher in lead year 1 and decrease with the increase of the length of the lead time window. The spatial distributions of RMSE vary depending on the season. For example, in DJF for lead year 1, the highest (lowest) values have been found in the north (south). The Baetic System shows the largest errors in MAM, but they are also situated over the Iberian System in JJA. In SON, the highest values are located in some of the main mountain regions, such as the previously mentioned Iberian and Baetic Systems, alongside the Central System and Sierra Morena. The spatial distributions are to some extent maintained across lead time and show similarities with those of the AEMET T_{min} standard deviation (Figure B.44 in Appendix B.2.3). As seen in the analyses of T_{max} and T_{min}, the highest errors are commonly observed in those regions which show the highest variability.

The results obtained for the seasonal ACC (Figure 5.25) follow the same path observed in the analyses of T_{max} and T_{min}, with the best results found in MAM and JJA.

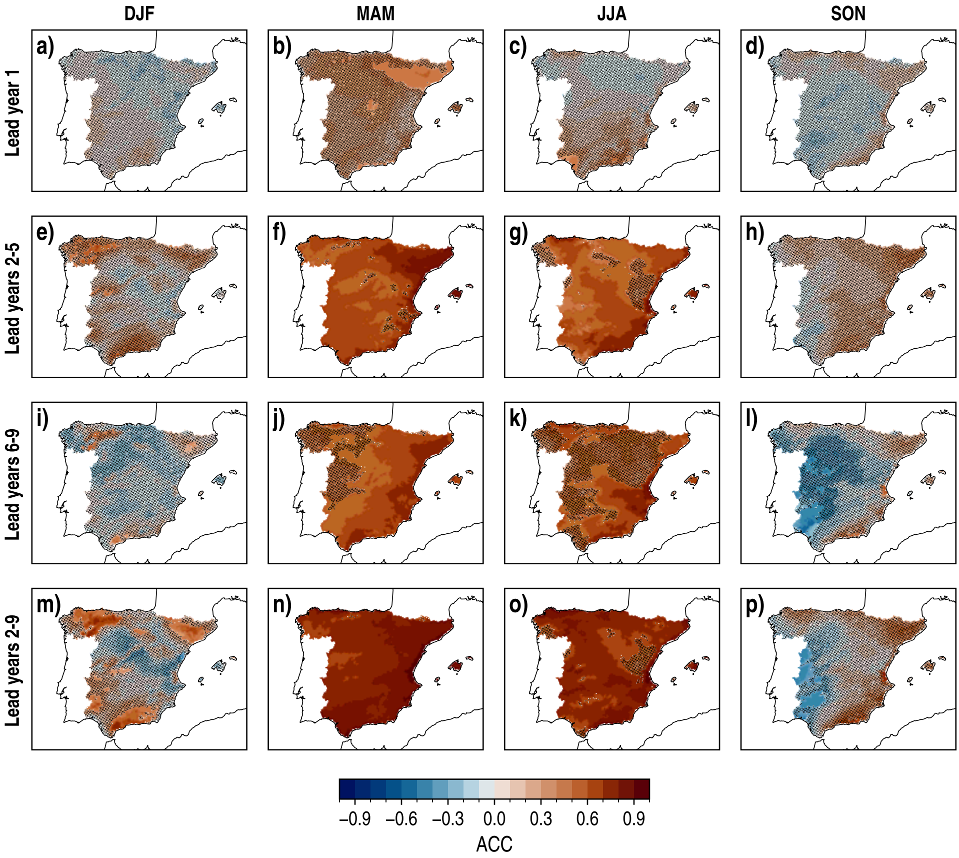

Figure 5.25: Spatial distributions of RMSE for the WRF-DPLE multiannual mean anomalies of T_{mean} for lead years 1, 2-5, 6-9 and 2-9 (rows) in DJF, MAM, JJA and SON (columns). The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

The domain is almost fully covered by not significant results for lead year 1 in these seasons, but the results clearly improve at other lead times. The highest correlations are observed in lead years 2–9, when the values are mainly above 0.6 in both MAM and JJA, especially in the former. In DJF and SON, the results are generally not significant. The regions having significant results in DJF mostly show positive values (e.g., Figure 5.25m), whereas they are negative in SON (e.g., Figure 5.25p). Even though, in SON, there are positive (but not significant) correlations at all lead times in some regions mainly situated in the eastern flank and in the north of the domain. The signal-to-noise paradox is also present in T_{mean}, as shown by the spatial distribution of the RPC for lead years 2–9 in Figure 5.26, because the RPC values are significantly higher than 1 over the whole domain. In the same vein as the other variables, although

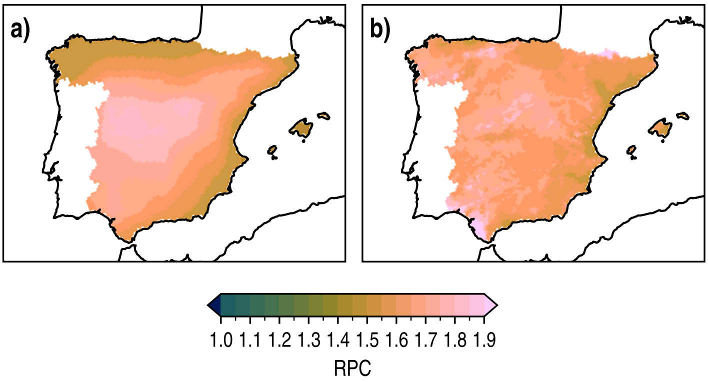

Figure 5.26: Signal-to-noise paradox in the hindcasts for T_{mean} over the IP. **a**) Spatial distribution of RPC for the multiannual mean anomalies of the CESM-DPLE T_{mean} in lead years 2–9 at annual scale. **b**) As **a**) but for WRF-DPLE. The absence (presence) of black dots indicates (not) statistically significant results different from 1 at the 90 % confidence level.

the ensemble size of 4 members is enough to achieve some predictive skill, especially in temperature-based variables, there is a potential to improve the performance by removing the unpredictable background noise with the addition of new members. However, although the RPC for T_{mean} is higher than 1, it is generally lower than that for PR, even sometimes by a factor of 2 (compare Figure 5.26 with Figure 4.5 for lead years 2–9). This means that, as mentioned in Section 5.1.1, the addition of a fixed number of members would not contribute to improving the ACC for T_{mean} as much as it would do for PR. This can be seen in Reyers et al. (2019, Figures 5d and 5e), who showed that, in their experimental framework, the increase of the ensemble size from 4 to 10 members enhances the ACC for T_{mean} by about 0.04 in the IP, in contrast to the increase around 0.4 for PR. A larger ensemble size may also help to reduce the RMSE in Figures 5.24 and B.43 through the increase of MSSS, as shown in Reyers et al. (2019, Figure 5a) by comparing the downscaled decadal predictions with an uninitialized fixed-size ensemble of global experiments.

The spatial distributions of MSSS_{C} (Eq. [3.24]) for each lead time are depicted in Figure 5.27 (left column). The outcomes are generally positive, although they are predominantly not significant at all lead times. In lead year 1, with not significant positive values across the whole domain, the highest MSSS_{C} values, above 0.4, have been found over the Baetic System. The general view is similar in lead years 2–5, but with slightly higher positive scores over the whole IP. At this lead time, there are a few locations to the south of the Baetic System, along the Mediterranean coast, depicting statistically significant results ranging from 0.3 to 0.7. There is an area showing negative scores over the Northern Subplateau and its sourthern vicinities in lead years 6–9, although they are mainly not significant. The significant negative results are below -0.7. The regions which show significant positive results have a larger extension than at the other lead times. These locations are situated along the Mediterranean coast, with $\mathrm{MSSS_C}$ values between 0.4 and 0.7. The highest scores have been found in lead years 2–9, the lead time window which has the largest amount of significant positive results, although the area showing not significant results continues being large. The highest $\mathrm{MSSS_C}$ values, above 0.7, are observed in the northeast, the northwest and some inner and coastal regions in the southern part of the IP.

The spatial distributions of $\mathrm{MSSS_C}$ are determined by those of ACC and CB (Eq. [3.17]), as indicated by Eq. [3.24]. In general, the results obtained in terms of CB (Figure 5.27, central column), which are significantly negative mostly in lead years 1 and 6–9, indicate an imbalance between ACC and the ratio $s_{\{}Y\}/s_X$ in which the former is dominated by the latter, commonly due to low ACC values and not to the existence of locations with $s_{\{}Y\}$ being higher than s_X. The worst CB results are observed in lead years 6–9, when values mainly between -0.8 and -0.4 have been found over the Northern Subplateau and its surroundings to the southwest. There are a few locations in the west with values lower than -0.8. On the other hand, the ACC values over these regions are in the range between 0.1 and 0.4. The confluence of these results lead to the negative $\mathrm{MSSS_C}$ outcomes found in Figure 5.27g. Contrarily, the regions with high ACC and not significant low CB generally depict a high $\mathrm{MSSS_C}$. See, for example, the Guadalquivir Valley for lead years 2–9 in Figures 5.24h, 5.27j and 5.27k. The spatial distributions of $\mathrm{MSSS_{CBA}}$ (Figure 5.27, right column) show the maximum $\mathrm{MSSS_C}$ which could be achieved if CB were completely removed (CB = 0 in Eq. [3.24]). The regions which would benefit the most are those showing positive CB values, significant or not, which are commonly the same that have the highest ACC outcomes. The Baetic System in lead years 2–9 is a good example to illustrate this. The ACC values over there are around 0.8 (Figure 5.24h), whereas the CB is positive and significant with values from 0.4 to 0.6 (Figure 5.27k). Thus, the difference between $\mathrm{MSSS_{CBA}}$ and $\mathrm{MSSS_C}$ is around 0.2. There are another regions, fundamentally in lead years 2–5 and 2–9, which would also experience a relatively large improvement in terms of $\mathrm{MSSS_C}$ after ajusting CB. Contrarily, there are also regions which would not benefit very much from the CB adjustment because of their low correlations, such as those places along the Mediterranean coast in lead year 1, with $\mathrm{MSSS_{CBA}}$ values below 0.1. In Reyers et al. (2019, Figures 2a and 2b), significant positive $\mathrm{MSSS_C}$ values

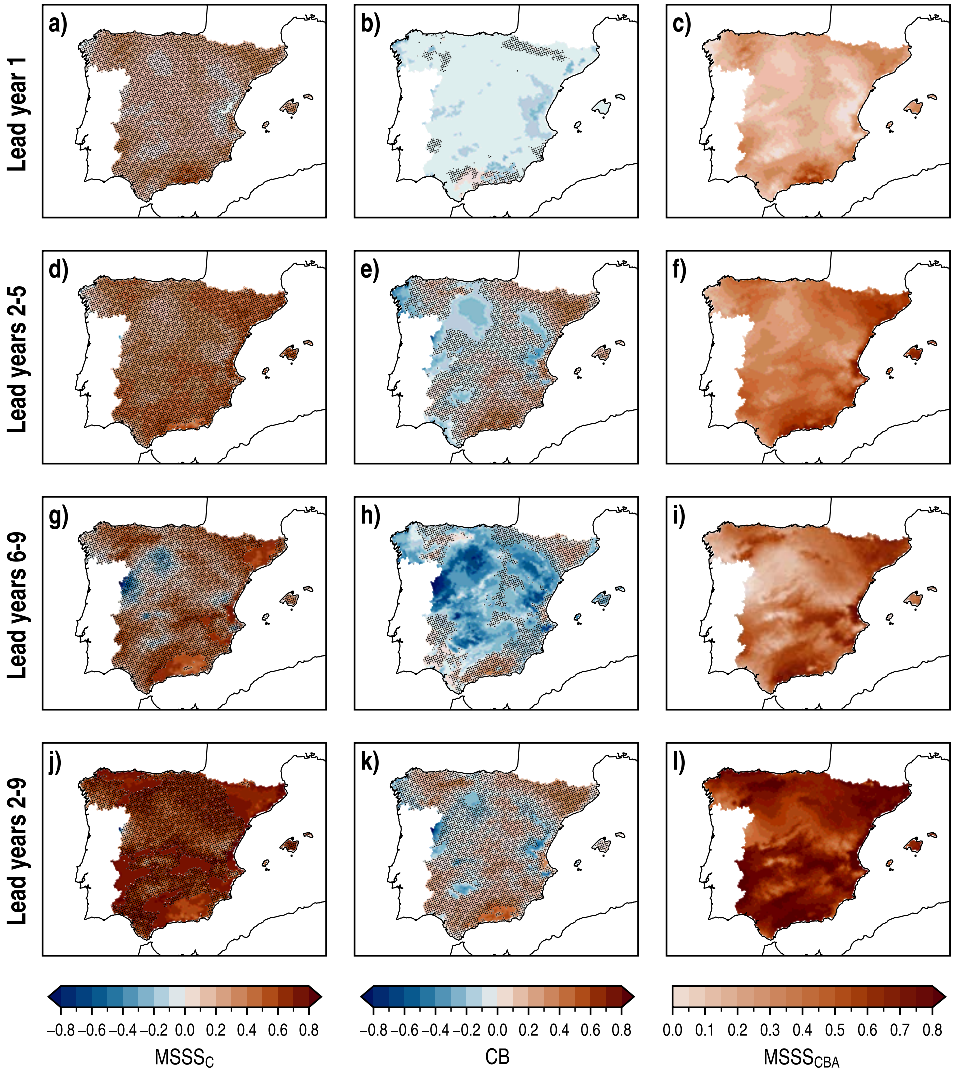

Figure 5.27 : Spatial distributions of $\mathrm{MSSS_C}$ (left column), with climatology as reference, CB (center column) and the same $\mathrm{MSSS_C}$ calculated for lead time series with an adjusted CB, i.e., equal to zero ($\mathrm{MSSS_{CBA}}$; right column), for the WRF-DPLE multiannual mean anomalies of T_{mean} in lead years 1, 2–5, 6–9 and 2–9 (rows) at annual scale. In $\mathrm{MSSS_C}$ and CB maps, the absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

were found across the IP from the southwest to the northwest for an ensemble size of 10 members. In this Thesis, the aforementioned reduction of the RMSE because of the increase of the ensemble size would contribute to improving the $\mathrm{MSSS_C}$ results.

The spatial distributions of the MSSS$_C$ calculated for the multiannual mean anomalies of T_{mean} at seasonal scale are available in Figure B.45 (Appendix B.2.3). As observed for the seasonal ACC (Figure 5.25), the best results have been found in MAM and JJA. In these seasons, the MSSS$_C$ outcomes are mostly positive, with the exception of lead year 1 in JJA, but the statistical significance is constrained to a few regions whose location vary depending on the season and the lead time. In JJA, the Baetic System shows significant positive results in lead years 2–5 and 2–9. The northeast and the Balearic Islands have significant results in MAM for lead years 2–5, whereas the most promising outcomes are widespread over the eastern and southern sectors of the domain in lead years 2–9. Again, these positive results are a consequence of the confluence of high ACC and close-to-zero CB values (Figure B.46) in those regions. These small absolute values of the seasonal CB (for example, for lead years 2–5 and 2–9 in MAM) are, in turn, also caused by high ACC scores which counteract the magnitude of the seasonal ratio $s_{\{}Y\}/s_X$. When this ratio is larger than ACC, the CB becomes very low and negatively impacts on MSSS$_C$ (see DJF and SON in Figures B.45 and B.46) The seasonal MSSS$_{CBA}$ is depicted in Figure B.47 (Appendix B.2.3) and its interpretation is similar to that provided at annual scale.

❦ *Reliability analysis*

The analysis of the predictive skill in terms of accuracy has been complemented by the assessment of the hindcasts reliability. The spatial distributions of CRPSS (Eq. [3.29]; Figure 5.28, left column) show that the average ensemble spread $\overline{\sigma_Y^2}$ (Eq. [3.32]) is adequate to quantify the forecast uncertainty at all lead times in large areas, as determined by the not statistically significant CRPSS values. In lead year 1, the not significant CRPSS results are mainly concentrated in the northwestern sector of the IP, with the addition of some regions in the central west, the northeast and close to the Strait of Gibraltar. In lead years 2–5, the not significant results mainly cover the inner regions of the IP, whereas the significant values spread along the Mediterranean and Cantabrian coasts, the northeast and some mountain inner regions such as the Iberian and the Central Systems. The lowest CRPSS values have been found in the Baetic System and the northeast. The absolute value of the scores is generally lower in lead years 6–9. The spatial distribution of the not significant results is very similar to that observed in lead years 2–5, although there are still some differences, such as new not significant values in the eastern part of the domain or the loss of reliability in the southwest. The Baetic System continues showing the lowest scores also at this lead time. The spatial distribution in lead years 2–9 is very similar to that in lead

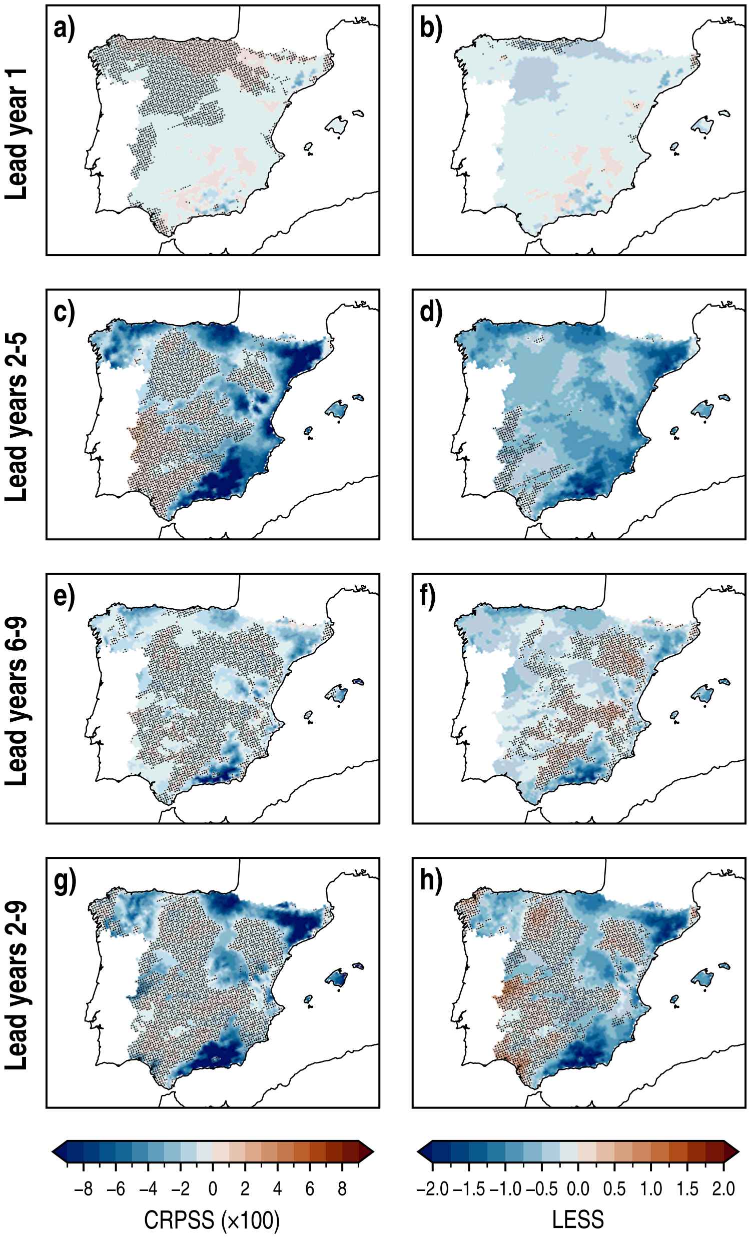

Figure 5.28: Spatial distributions of CRPSS (left column) and LESS (right column) for the WRF-DPLE multiannual mean anomalies of T_{mean} in lead years 1, 2–5, 6–9 and 2–9 (rows) at annual scale. The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

years 2–5, but with a greater abundance of not significant results. These outcomes are determined by the extent to which the hindcasts are over- or underdispersive, i.e., $\overline{\sigma_Y^2} > \sigma_X^2$ or $\overline{\sigma_Y^2} < \sigma_X^2$, respectively. The relation between $\overline{\sigma_Y^2}$ and σ_X^2 is addressed by the LESS (Eq. [3.34]; Figure 5.28, right column). The generalized presence of significant negative LESS values shows that the hindcasts are underdispersive at all lead times. There are significant positive LESS values very close to zero in lead year 1, but those observed in lead years 6–9 and 2–9 are not significant. In lead years 1 and 2–5, almost the whole domain is covered by significant values, with higher negative results at the latter lead time. In lead years 6–9 and 2–9, the not significant results are more frequent. The regions with the closest-to-zero LESS values, although they may be significant, are those which show the not significant CRPSS results. Indeed, the structure of the spatial distributions of both metrics matches almost perfectly.

The best results in the seasonal analysis of CRPSS have been obtained in JJA (Figure B.48 in Appendix B.2.3), when the not significant results are predominant at all lead times, excepting in lead year 1. These scores are a consequence of the not significant values also observed in the LESS spatial distributions (Figure B.49). In general, the LESS indicates that the hindcasts are mainly underdispersive also at seasonal scale, since the presence of significant positive LESS values is very uncommon. The worst results have been found in SON, when a strong underdispersion leads to significant CRPSS outcomes which cover the whole domain at all lead times.

5.3.2. *Comparison with the CESM-DPLE subensemble*

❦ *Accuracy analysis*

The spatial distributions of $\mathrm{MSSS_G}$ at annual scale, calculated for the WRF-DPLE multiannual mean anomalies of T_{mean} with CESM-DPLE as reference, are depicted in Figure 5.29 (left column). These results are characterized by a notable absence of statistically significant results at all lead times. In lead year 1, there is a band of positive scores from east to west, with the results reaching maximum values around 0.4 in the central regions, the north of the IP and the Baetic System. However, in spite of the magnitude of these scores, they are not significant. Negative but not significant results are observed mainly in the south of the domain and in some regions in the northwest. The scores decrease in lead years 2–5, so the area covered by positive results is lower than in lead year 1. The largest fraction of significant positive $\mathrm{MSSS_G}$ outcomes is observed in lead years 6–9. Scores between 0.3 and 0.6 are situated in central regions and locations in the southern part of the IP. A few locations over the

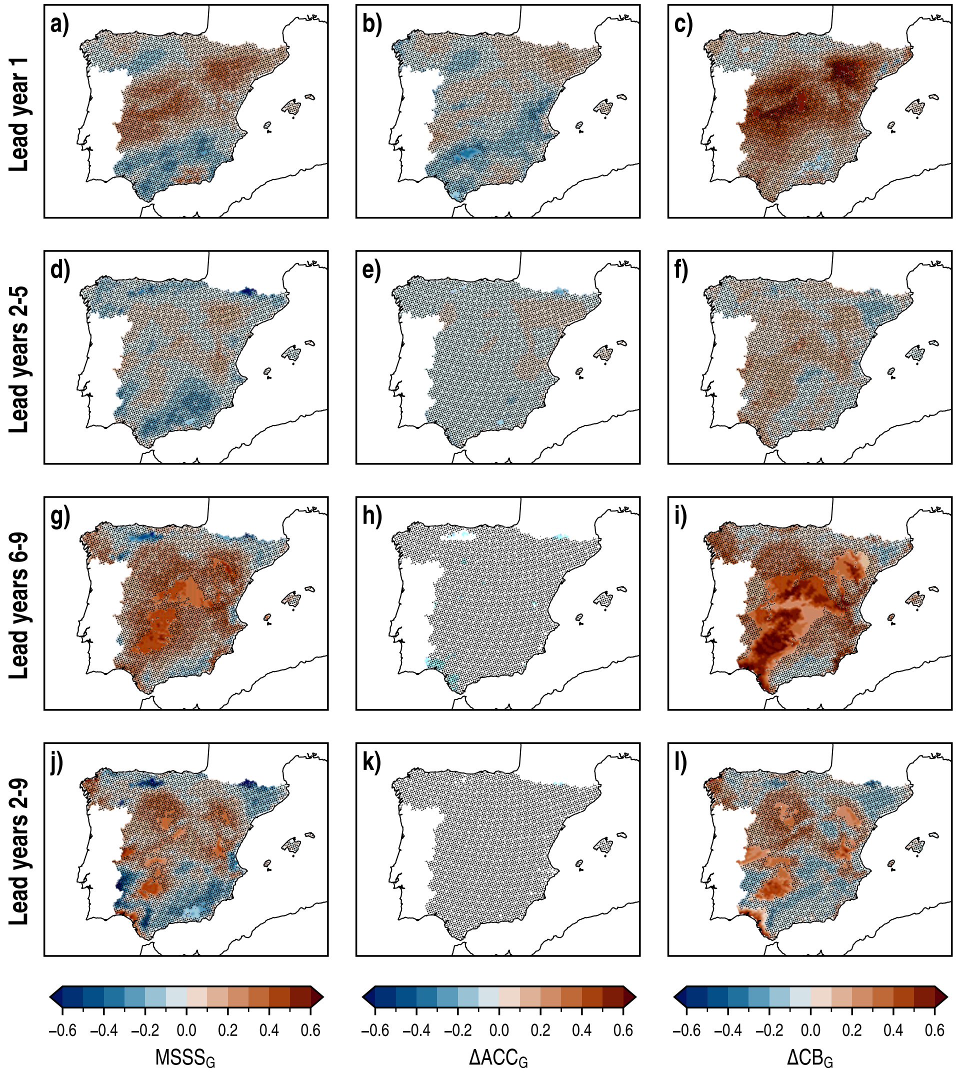

Figure 5.29: Spatial distributions of $MSSS_G$ (left column), ΔACC_G (center column) and ΔCB_G (right column), with CESM-DPLE as reference, for the WRF-DPLE multiannual mean anomalies of T_{mean} in lead years 1, 2–5, 6–9 and 2–9 (rows). The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

Ebro Valley also show very high significant scores up to 0.6. At this lead time, the negative results are mainly restricted to small areas over mountain regions, such as those to the south of the Pyrenees, the Baetic System and the Cantabrian Range. The latter contain the lowest significant negative results, with values starting from -0.2 and even surpassing -0.6. The scores are less promising in lead years 2–9, when both the value and the area covered by significant results decrease compared to lead years 6–9. These results are directly determined by those obtained for the $\Delta\mathrm{ACC}_{\mathrm{G}}$ and $\Delta\mathrm{CB}_{\mathrm{G}}$ spatial distributions displayed in Figure 5.29 (central and right columns, respectively). The differences between the performance of the downscaled and global hindcasts in terms of ACC are not statistically significant, likewise for the other variables. In this case, the contribution of the downscaled hindcasts to the predictive skill for T_{mean} is made through the correction of the imbalance between ACC and $s_{\{Y\}}/s_X$, reflected by the improvement achieved in terms of CB. The largest added value to the CB is observed in lead years 6–9, when a large portion of the IP shows significant positive results, reaching values above 0.6 mostly over the Central System and some southern regions. Very high improvements have been also found in lead year 1, but the significant results are constrained to small regions over the Central System and Ebro Valley. This influence of the $\Delta\mathrm{CB}_{\mathrm{G}}$ on $\mathrm{MSSS}_{\mathrm{G}}$ can be visually verified by comparing the structure of the spatial distributions at each lead time.

At seasonal scale, the most promising results for $\mathrm{MSSS}_{\mathrm{G}}$ (Figure 5.30) have been found in SON. In lead years 6–9, the eastern and western regions show significant positive results mostly between 0.2 and 0.5, although most part of the domain does not show any statistical significance. In lead years 2–9, almost the whole domain depicts a robust improvement compared to the global experiments, with the exception of the southeast and some northern regions. Wide areas covered by significant positive results have also been found in JJA for lead years 6–9 and 2–9. Most part of the significant results are positive, whereas the largest region which shows significant negative results is seen over the Baetic System in lead years 2–9 in DJF. The differences in terms of ACC (Figure B.50) are generally not significant. Only in DJF for lead years 2–9, significant and negative results are depicted over the Northern Subplateau with minimum values up to -0.4. At at annual scale, the $\Delta\mathrm{CB}_{\mathrm{G}}$ (Figure B.51) results are the main factor leading to the positive outcomes observed in Figure 5.30, as the regions which show the best results for $\Delta\mathrm{CB}_{\mathrm{G}}$ are generally the same with the best results for $\mathrm{MSSS}_{\mathrm{G}}$.

The results obtained by Reyers et al. (2019, Figures 4a, b and Table 1) hardly show any added value of their downscaled hindcasts to the T_{mean} predictive skill neither in terms of $\mathrm{MSSS}_{\mathrm{G}}$ nor for $\Delta\mathrm{ACC}_{\mathrm{G}}$. Depending on the initialization method of the GCM which provides the ICs and LBCs, a few locations with significant positive or negative $\mathrm{MSSS}_{\mathrm{G}}$ values are found over the IP in lead years 1–5 (at annual scale). They found some added value in regions with low to medium predictive skill in the global

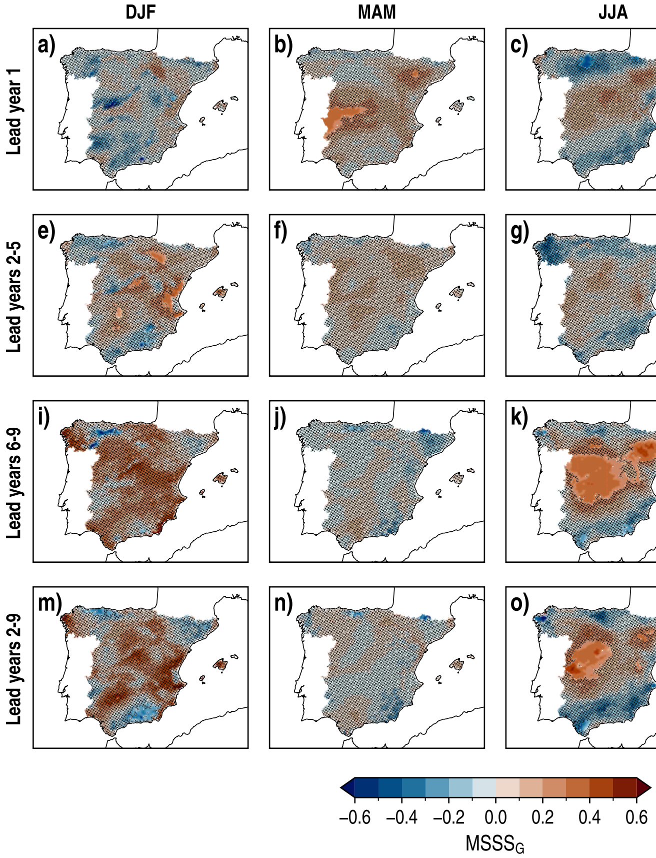

Figure 5.30: Spatial distributions of MSSS$_G$ for the WRF-DPLE multiannual mean anomalies of T_{mean}, with CESM-DPLE as reference, in lead years 1, 2–5, 6–9 and 2–9 (rows) in DJF, MAM, JJA and SON (columns). The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

hindcasts (e.g., in Scandinavia), but the improvement is more limited or directly missing in regions where the GCM already showed relatively high skill, such as the IP. Given the results depicted in Figures 5.27 and 5.28, which show a high skill in terms of accuracy, similar conclusions might be drawn here. Indeed, the highest added value observed in Figure 5.30 is found in SON, one of the seasons which showed the worst results in Figure B.45.

❦ *Reliability analysis*

The difference between the WRF-DPLE and CESM-DPLE performances in terms of reliability is measured by the ΔCRPSS$_G$, whose spatial distributions are depicted in Figure 5.31 (left column). There are not significant differences between both CRPSS, so there is not neither a significant added value nor a degradation of the hindcasts reliability in the downscaled hindcasts compared to the global experiments at any lead time. However, in lead year 1, the domain is almost fully covered by positive results, with the maximum values observed in the northwestern regions. The presence of negative results dominates in lead years 2–5, when the positive outcomes are mainly found in the northern regions along the Cantabrian coast and to the south of the Pyrenees. These northern positive results are maintained in lead years 6–9 and another locations showing an improvement by the downscaled hindcasts mainly spread across the central and eastern parts of the domain. The largest added values, although not significant, have been found in lead years 2–9, situated to the south of the Central System. The Guadalquivir and Guadiana Valleys, the northwest and other northern regions also show an added value in terms of CRPSS.

The results shown for ΔCRPSS$_G$ are determined by those achieved for LESSS$_G$ (Figure 5.31, right column), that indicate the extent to which the ensemble underdispersion or overdispersion is attenuated in the downscaled hindcasts compared to the global experiments. The regions which show an improvement in the representation of the average ensemble spread (positive LESSS$_G$ values) are the same which experience an improvement of the hindcast reliability in the ΔCRPSS$_G$ spatial distributions. As seen for the other variables analyzed in the previous sections, the absolute values of LESSS$_G$ are very large because of the definition of the score (Eq. [3.36]). The dependence on the squared forms of the downscaled and global LESS makes LESSS$_G$ very sensitive to changes in both, especially when the absolute LESS is small.

The results for the seasonal ΔCRPSS$_G$ and LESSS$_G$ have been depicted in Figures B.52 and B.53 (Appendix B.2.3). The general picture is similar to that at annual scale, since there is not a robust added (or lost) value in the hindcasts reliability of the WRF-DPLE compared to the CESM-DPLE subensemble. The best results have been obtained in SON, when positive ΔCRPSS$_G$ (but not significant) have been found over the whole domain in lead years 6–9 and 2–9, as consequence of the generalized attenuation of the ensemble underdispersion or overdispersion at those lead times.

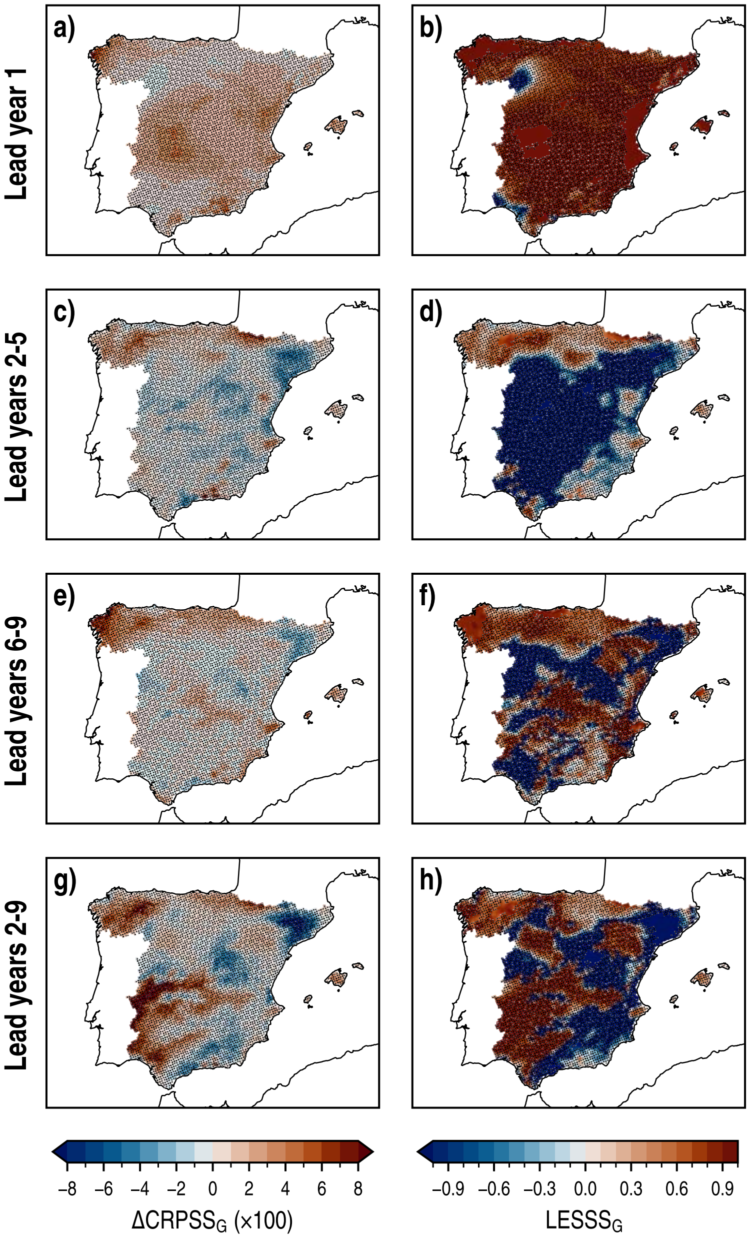

Figure 5.31: Spatial distributions of $\Delta\text{CRPSS}_\text{G}$ (left column), and LESSS_G (right column) for the WRF-DPLE multiannual mean anomalies of T_mean, with CESM-DPLE as reference, at lead years 1, 2–5, 6–9 and 2–9 (rows). The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

5.3.3. *Comparison with the WRF-LE ensemble*

In this Section, the performance of the WRF-DPLE ensemble has been compared with that of the unitilialized experiments which compose the WRF-LE ensemble. This assessment, as done for PR and the other NSAT variables, has been carried out by only using accuracy metrics, because the uninitialized experiments are not interpreted in a probabilistic manner as the initialized experiments are. In addition, since the uninitialized experiments cover only the period 1990-2005, the analysis is focused on lead years 1 and 2–5, and consider only the hindcasts initialized every year from 1990 to 1999 (10 start dates).

❦ *Accuracy analysis*

The spatial distributions of $\mathrm{MSSS_U}$, calculated with multiannual mean anomalies of T_{mean} by considering the WRF-LE as reference, are displayed in Figure 5.32 (left column). The best results are observed in lead year 1, when positive scores span the whole IP, with the exception of the eastern regions close to the Mediterranean coast. Notwithstanding, the outcomes mostly lack of statistical significance. There are some southern regions over the Baetic System and areas close to the Strait of Gibraltar which show significant positive results with values above 0.3. The maximum scores, between 0.6 and 0.7, are situated in the Baetic System. In lead years 2–5, the uninitialized experiments clearly outperform the hindcasts predictive skill. The domain is filled with negative scores, with some expections in the Baetic System, the northeast and the Cantabrian range. The negative outcomes are significant over wide areas of the domain from north to south, mainly in the eastern part, with minimum scores below -0.8. These results are derived from those obtained in terms of $\Delta\mathrm{ACC_U}$ and $\Delta\mathrm{CB_U}$(Figure 5.32, center and right columns, respectively), as stated by Eq. [3.16]. In lead year 1, the added value of WRF-DPLE to both ACC and CB contributes to producing the significant added value in terms of $\mathrm{MSSS_U}$ in the southern regions. In both cases, not significant positive differences have been found along the Cantabrian coast, the Ebro Valley and central western regions. The generalized positive $\Delta\mathrm{CB_U}$ values counteract the loss in ACC found in some central regions, but they are not high enough to do it in the easternmost locations of the domain. In the southeast, the joint action of the negative $\Delta\mathrm{ACC_U}$ and $\Delta\mathrm{CB_U}$ values leads to the lowest scores observed in the $\mathrm{MSSS_U}$ map over this region. Although some positive differences in ACC have been found along the Mediterranean coast in lead years 2–5, the lost value in terms of CB is too large to produce positive $\mathrm{MSSS_U}$. The lowest $\mathrm{MSSS_U}$ values scores observed

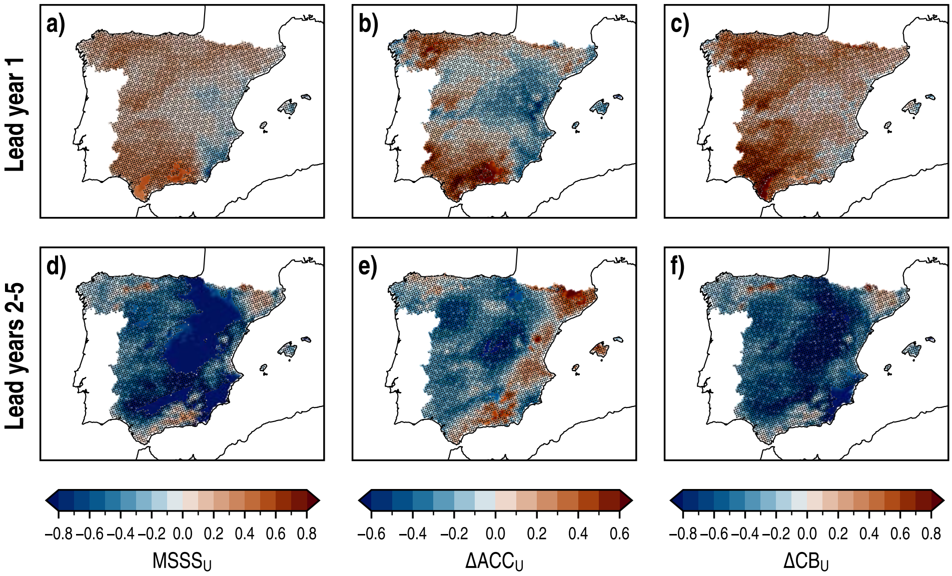

Figure 5.32: Spatial distributions of MSSS$_U$ (left column), ΔACC$_U$ (center column) and ΔCB$_U$ (right column) for the WRF-DPLE multiannual mean anomalies of T_{mean}, with WRF-LE as reference, in lead years 1, 2–5, 6–9 and 2–9 (rows). The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

in Figure 5.32d are caused by the negative results obtained for ΔACC$_G$ and ΔCB$_G$.

Reyers et al. (2019) did not find a consistent added value over the IP when comparing the performance of their downscaled hindcasts with the global (not downscaled) uninitialized experiments in terms of MSSS in lead years 1–5. Indeed, depending on the ICs and LBCs, the scores can be either generally negative or a mix of positive and negative (Figures 3a,b in Reyers et al., 2019). In both cases, significant negative results were obtained in the northeast. Their results highlight the influence of the initialization scheme of the GCM in the predictive skill of the downscaled product for T_{mean}, as happened also for PR. In the analysis of the dependence of MSSS and ΔACC on the ensemble size, they found that improvements about 0.08 and 0.02 are achieved, respectively, by increasing the number of members from 4 to 10 in their experimental framework (Figures 5a,d in Reyers et al., 2019). The added value of incrementing the number of members in this case is drastically lower than that found for PR.

At seasonal scale, the best results achieved for MSSS$_U$ have been found in MAM (Figure 5.33), with positive scores covering the whole domain and being significant in

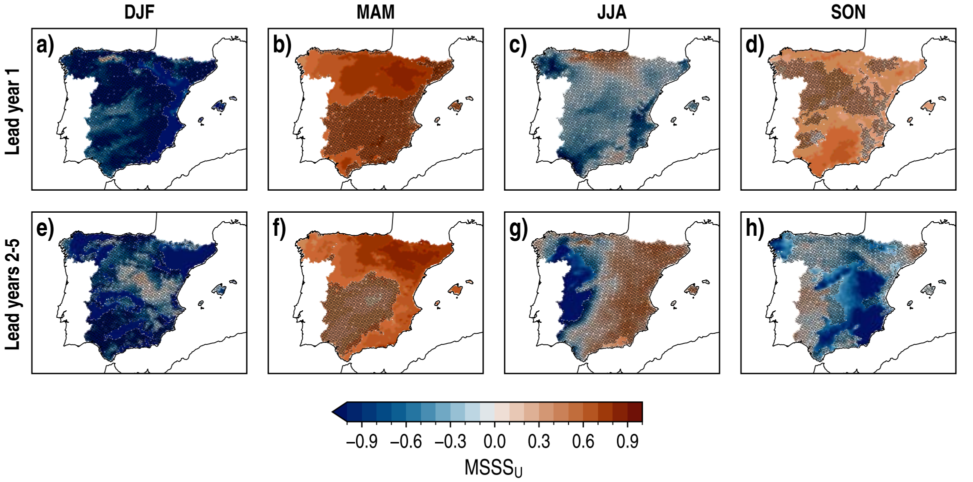

Figure 5.33: Spatial distributions of MSSS$_U$ for the WRF-DPLE multiannual mean anomalies of T_{mean}, with WRF-LE as reference, in lead years 1, 2–5, 6–9 and 2–9 (rows) in DJF, MAM, JJA and SON. The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

the northern part a lead year 1 and, additionally, also along the Mediterranean coast in lead years 2–5. At both lead times, the maximum values, between 0.8 and 0.9, are observed in the northeast. Significant positive results have also been found in SON for lead year 1 across the domain, with maximum outcomes around 0.6 in the Baetic System. These promising results are caused by the largely positive differences found in terms of ΔACC$_U$ (Figure B.54) and ΔCB$_U$ (Figure B.55). The opposite situation is depicted in DJF, the season which shows the largest lost value compared to the uninitialized simulations. In lead years 1 and 2–5, the very low MSSS$_U$ values are caused by the large degradation in terms of ACC and CB.

5.3.4. *Predictive skill for regional averages*

Finally, the assessment of the WRF-DPLE predictive skill for T_{mean} has been completed with the analysis of the performance of WRF-DPLE in the regions obtained from the regionalization done in Section 3.6. The same skill scores of the previous sections have been calculated with the spatially averaged lead time series over each region. The results are summarized in Table 5.3.

The downscaled hindcasts are generally able to reproduce the observational T_{mean} variability in all regions, as shown the positive and predominantly significant results obtained for ACC. The most robust correlations are observed in lead years 2–9, when

Table 5.3: Skill scores for the spatially averaged WRF-DPLE multiannual mean anomalies of T_{mean} in lead years 1, 2–5, 6–9 and 2–9 at annual scale. The subscripts C, G and U denote the reference data used to calculate the skill score: AEMET climatology, CESM-DPLE global hindcasts and WRF-LE uninitialized experiments, respectively. The bold formatting indicates results different from zero at the 90 % confidence level. Dashes denote data unavailability at that lead time.

Region	**Lead years**	**MSSS$_{\mathbf{C}}$**	**ACC**	**CB**	**CRPSS (×100)**	**MSSS$_{\mathbf{G(U)}}$**	**ΔACC$_{\mathbf{G(U)}}$**	**ΔCB$_{\mathbf{G(U)}}$**
SW	1	0.24	**0.49**	**-0.03**	**-0.03**	-0.07 (0.34)	-0.09 (0.32)	0.19 (0.66)
	2-5	0.44	**0.66**	-0.01	0.52	-0.10 (-0.59)	-0.05 (-0.23)	0.11 (-0.51)
	6-9	0.47	**0.69**	-0.06	-0.01	0.26 (–)	-0.02 (–)	**0.40** (–)
	2-9	0.76	**0.88**	0.14	-0.44	-0.06 (–)	-0.01 (–)	-0.00 (–)
NO	1	0.22	**0.47**	**-0.04**	0.11	-0.01 (0.20)	-0.01 (0.17)	-0.03 (0.33)
	2-5	0.34	0.59	-0.04	**-3.95**	-0.10 (-0.36)	-0.05 (-0.20)	-0.01 (-0.37)
	6-9	0.33	**0.60**	**-0.18**	**-1.22**	0.07 (–)	-0.03 (–)	0.15 (–)
	2-9	0.65	**0.81**	0.07	**-2.32**	-0.03 (–)	-0.00 (–)	-0.04 (–)
CI	1	0.17	0.42	**-0.05**	**-0.05**	0.09 (0.12)	-0.05 (-0.06)	0.31 (0.30)
	2-5	0.38	0.62	0.01	-1.09	-0.03 (-0.91)	-0.02 (-0.32)	0.09 (-0.78)
	6-9	0.34	**0.65**	**-0.27**	-2E-03	0.35 (–)	0.02 (–)	**0.37** (–)
	2-9	0.66	**0.81**	0.07	-0.51	0.11 (–)	0.02 (–)	0.08 (–)
NE	1	0.25	**0.50**	**-0.04**	**9E-04**	0.15 (0.14)	0.06 (-0.01)	0.23 (0.25)
	2-5	0.50	**0.73**	0.17	**-3.80**	-0.00 (-0.51)	0.01 (0.03)	-0.09 (-0.42)
	6-9	0.56	**0.75**	-0.10	**-0.41**	0.10 (–)	-0.00 (–)	0.16 (–)
	2-9	**0.72**	**0.87**	0.19	**-2.79**	-0.06 (–)	0.01 (–)	-0.11 (–)
CS	1	0.42	**0.65**	**-0.06**	**-0.12**	-0.02 (0.42)	-0.01 (0.42)	-0.00 (0.15)
	2-5	0.48	**0.75**	0.29	**-8.85**	-0.19 (-0.04)	-0.06 (0.25)	0.02 (-0.24)
	6-9	**0.60**	**0.78**	0.14	-1.02	-0.05 (–)	-0.00 (–)	-0.08 (–)
	2-9	**0.64**	**0.88**	0.37	**-7.41**	-0.22 (–)	-0.02 (–)	-0.05 (–)
EA	1	0.17	0.41	**-0.07**	**-0.03**	-0.02 (-0.09)	-0.08 (-0.19)	0.17 (0.11)
	2-5	0.49	0.70	0.06	**-2.94**	-0.03 (-0.72)	-0.01 (0.10)	-0.04 (-0.57)
	6-9	0.50	**0.74**	-0.22	6E-04	0.26 (–)	-0.02 (–)	0.28 (–)
	2-9	0.73	**0.85**	0.04	-0.24	0.00 (–)	-0.01 (–)	0.09 (–)
MT	1	0.24	**0.49**	**-0.05**	-0.03	-0.05 (0.31)	-0.04 (0.22)	0.02 (0.49)
	2-5	0.41	0.64	0.05	-0.83	-0.12 (-0.20)	-0.05 (-0.02)	-0.01 (-0.20)
	6-9	0.52	**0.72**	-0.03	**-0.11**	-0.07 (–)	-0.04 (–)	0.12 (–)
	2-9	0.71	**0.87**	0.21	-0.54	-0.15 (–)	-0.00 (–)	-0.12 (–)
WI	1	0.14	0.38	**-0.06**	-0.14	0.01 (0.21)	-0.05 (0.06)	0.17 (0.34)
	2-5	0.29	0.55	**-0.13**	-0.51	-0.02 (-0.47)	-0.03 (-0.33)	0.06 (-0.48)
	6-9	0.03	0.45	**-0.41**	-0.15	0.22 (–)	-0.04 (–)	0.28 (–)
	2-9	0.60	**0.78**	-0.07	-0.22	0.18 (–)	0.01 (–)	0.21 (–)

the scores show the highest values in each region and they are also statistically significant over the whole domain. The best results in terms of ACC have been obtained in the SW, NE and CS regions, where significant positive correlations have been found at all lead times. Although the absolute value of CB is typically lower in T_{mean} hindcasts than in the cases of PR (Table 4.1) or T_{max} (Table 5.1), it is higher than that commonly found for T_{min} (Table 5.2). The results obtained for CB in the analysis of T_{mean} still show significant values at some lead times in all regions, hindering the achievement of significant positive outcomes for $\mathrm{MSSS_C}$. Although the $\mathrm{MSSS_C}$ values are positive at all lead times in all regions, they are only significant for lead years 2–9 in the NE region ($\mathrm{MSSS_C}$=0.72) and for lead years 6–9 and 2–9 in the CS region ($\mathrm{MSSS_C}$=0.60 and $\mathrm{MSSS_C}$=0.64, respectively). Likewise for the skill scores analyzed in the previous sections, the addition of more members to the ensemble would positively contribute to enhancing the predictive skill and, therefore, it could help to increase the robustness of $\mathrm{MSSS_C}$, among other skill scores. As observed for the previous variables, the comparison with the global hindcasts and the uninitialized experiments does not generally highlight any significant improvement or degradation of the predictive skill. From the point of view of reliability, all regions show that the average ensemble spread $\overline{\sigma_Y^2}$ is suitable to quantify the forecast uncertainty at least at one lead time window because of the not significant CRPSS results, excepting for the NE region, where CRPSS is significant at all lead times. In this sense, the best performance has been obtained in the WI region, where the WRF-DPLE hindcasts are reliable at all lead times. Other regions such as the MT, CI and SW regions also show good results in these terms, with most lead times showing not significant CRPSS results.

The lead time series for the WRF-DPLE ensemble mean and AEMET in the CS region have been depicted in Figure 5.34. As the significant positive results obtained for ACC indicate, WRF-DPLE shows skill to reproduce to some extent the observational variability at all lead times. Although the ACC results benefit from the positive trends of T_{mean} in WRF-DPLE and AEMET, there is also relative skill to reproduce the maximums and minimums observed in the AEMET time series. In lead years 6–9, for example, the WRF-DPLE is able to partially replicate in a general sense the increases and decreases described by the AEMET T_{mean} in this region. Although the ensemble mean is not fully accurate in capturing the magnitude of the signal, it is enough to get a statistically significant $\mathrm{MSSS_C}$ in lead years 6–9 and 2–9, as mentioned above. There is an overestimation of the AEMET T_{mean} anomaly at the first years of the control period, mainly in lead years 2–5, 6–9 and 2–9, as also happened in the cases

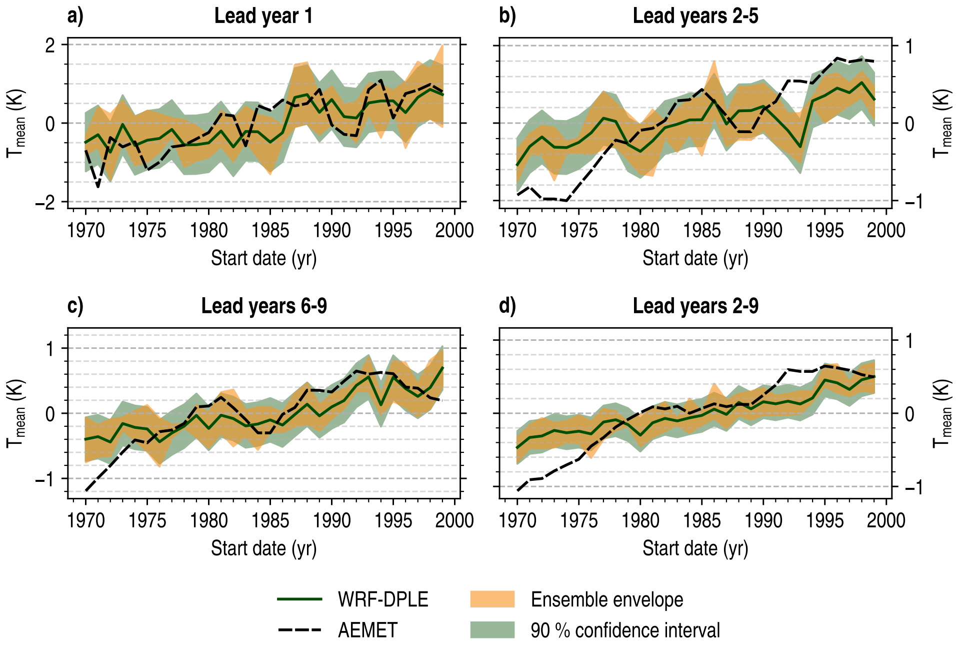

Figure 5.34: Time series of the spatially averaged multiannual mean anomalies of T_{mean} in the CS region for lead years 1, 2–5, 6–9 and 2–9 at annual scale. Solid green lines identify the WRF-DPLE ensemble mean, whereas dashed black lines correspond to AEMET. Shaded green surfaces indicate the 90 % confidence interval for a WRF-DPLE single member, calculated from the average ensemble spread (Eq. [3.32]). Shaded yellow surfaces show the ensemble envelope which encloses the trajectories followed by the members composing the WRF-DPLE ensemble.

of T_{max} (Section 5.1.4) and T_{min} (Section 5.2.4). In addition, the underestimation of the T_{min} anomaly in this region during the last start years of the control period (Figure 5.22) is also observed here for T_{mean} in lead years 2–5 and 2–9, although it is slightly less accentuated. In the same line as for T_{min}, this limited ability to represent the anomalies at the beginning and end of the control period is related to the existence of statistically significant differences between the WRF-DPLE and AEMET trends (see Table A.2 in Appendix A.2 to consult the results and Section 3.2 for the methodology applied to compute the trends). In this case, the trends of the difference between the WRF-DPLE and AEMET time series are about -0.41 and -0.28 K/decade in lead years 2–5 and 2–9, respectively, indicating that the actual anomaly trend is underestimated by the WRF-DPLE ensemble mean. These differences have a negative impact on the predictive skill in general. In particular, how they affect the reliability of the hindcasts can be examined by observing Figure 5.34. The AEMET anomalies often fall outside the confidence intervals mainly at the beginning, but also at end of the control period at these lead times. For this reason, the CRPSS results have high absolute values just at these lead times in the CS region. While the overestimation during the early start dates is frequent in all regions, the underestimation at the end is mainly present in the CS region. Some additional examples have been depicted in Figures B.56 and B.57, available in Appendix B.2.3. Likewise for T_{max} and T_{min}, the confidence intervals of the T_{mean} lead time series, relative to the magnitude of the signal, are narrower than those generally observed for PR because the signal-to-noise paradox is weaker for temperature variables, as concluded from Figures 4.5, 5.3, 5.14 and 5.26.

5.4. Analysis of near-surface air temperature trends in CESM-DPLE

In the same line of the analysis done in Section 4.5 for PR, this Section is devoted to explore the ability of CESM-DPLE to reproduce some characteristics of special interest observed in real climate fields. In this case, an assessment of the skill of CESM-DPLE to reproduce the trends of T_{mean} in lead years 2–9 has been conducted. This lead time has been considered in order to remove the interannual variability and fully focus on the decadal scale. Since CESM-DPLE provides the ICs and LBCs for T_{mean}, among other variables, to generate the downscaled WRF-DPLE hindcasts, the ability to predict the evolution of T_{mean} influences on the predictive skill achieved by the WRF outputs. In particular, the skill to reproduce the T_{mean} trends is essential in the generation of accurate and reliable downscaled DCPs, as these trends have a large contribution to the predictive skill through ACC. The methodology applied to compute the trends and assess their statistical significance is described in Section 3.2. As in Section 4.5, ERA5 has been used as the reference dataset. ERA5 not only provides T_{mean} information over the sea, commonly absent in observational products, but also it does with a higher spatial resolution than CESM-DPLE, allowing to work on the native model resolution by linearly interpolating the ERA5 T_{mean} onto the CESM-DPLE grid. The trends have been calculated for the 4-member and the 10-member CESM-DPLE ensemble means (ENS4 and ENS10, respectively) to evaluate the impact of the ensemble size in the results.

The results obtained at annual scale have been depicted in Figure 5.35. The ERA5 multiannual lead time series of T_{mean} show a positive trend over the whole EURO-CORDEX domain, generally showing statistical significance. The highest values have been found over the Arctic Ocean and Iceland, with significant outcomes above 0.8 K/decade. The IP shows slightly lower but still significant positive results between 0.6 and 0.7 K/decade, similarly to the northern Africa, the Middle East, the coastal regions of Scandinavia and the Alps. The lowest trends can be generally

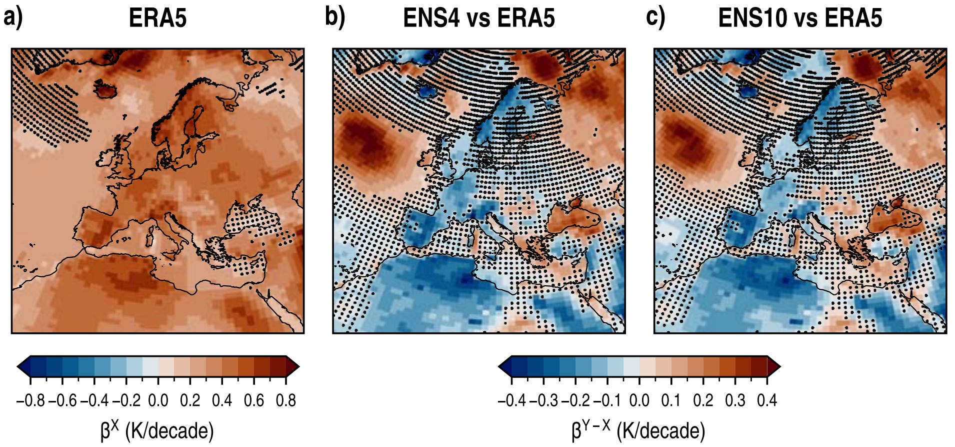

Figure 5.35: Spatial distributions of **a**) the trend of the ERA5 multiannual lead time series of T_{mean}, **b**) the trend of the difference between the 4-member CESM-DPLE ensemble mean (ENS4) and ERA5 lead time series of T_{mean}, and **c**) as **b**) but for the 10-member CESM-DPLE ensemble mean (ENS10). The lead time series have been calculated for the lead years 2–9 in the control period at annual scale. The ERA5 trend is represented by β^X, whereas β^{Y-X} denotes the trend in the difference of CESM-DPLE and ERA5 time series. The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90 % confidence level.

observed over the sea. Indeed, not significant results have been found mainly in the North Atlantic, the easternmost part of the Mediterranean Sea and the Black Sea and surrounding regions. Some areas on the northern Eurasia and Greenland also show not statistically significant results. The spatial distributions of the trend of the difference between the model and reanalysis time series are almost identical for both CESM-DPLE subensembles. This not only occurs at annual scale, but also in the seasonal results (Figure 5.36). This finding is consistent with what has already been discussed in Sections 5.1.1, 5.2.1 and 5.3.1, and it is also supported by the results found by Reyers et al. (2019). For temperature variables, the contribution of increasing the ensemble size to the predictive skill is expected to be positive but more modest than for other variables being affected by a stronger signal-to-noise paradox. Significant negative differences have been found mostly over land, especially over Greenland, Iceland and regions close to the Mediterranean basin, with values below -0.3 K/decade. On the other hand, the highest significant positive results have been found over the North Atlantic, the Arctic Ocean and the northern Eurasia. The hindcasts have shown to be very skilful in reproducing the ERA5 trends mainly over part of the central, eastern and northern Europe, as well as over large sea areas in the North Atlantic, the Mediterranean Sea and part of the Arctic Ocean. Differences between ensemble sizes are hardly identifiable, but they still exist. The trend of the difference series is slightly smaller for ENS10 over the cluster of significant positive result in the North Atlantic Ocean, and some additional not significant results have been found to the east of Iceland and spurious locations over land compared to ENS4. On the other hand, the area showing statistically significant results to the north of Scandinavia is larger for ENS10 than for ENS4.

In the particular case of the IP, the statistically significant underestimation of the actual trend observed in most part of the region supports the results discussed in Sections 5.1.4, 5.2.4 and 5.3.4 and summarized in Table A.2 (Appendix A.2), which showed that the observed NSAT trends are underestimated by the downscaled hindcasts over most part of domain at several lead times, including lead years 2–9, although not always with statistical significance. This underestimation seems to have been partially transferred by CESM-DPLE to WRF-DPLE hindcasts during the DD simulations.

The results obtained for the trend analysis at seasonal scale have been depicted in Figure 5.36. The positive trend observed in ERA5 T_{mean} is still present for most part of the EURO-CORDEX domain, although its magnitude varies depending on the season. While the highest significant positive results are shown for Scandinavia and the Arctic Ocean in DJF, they are generally observed over regions close to the Mediterranean basin in MAM and JJA. In SON, however, they are observed again in the north of the EURO-CORDEX domain and in the northern regions of Africa. The trend is positive and significant for most part of the IP in MAM and JJA, with values commonly between 0.6 and 1.2 K/decade. It is lower but mostly still significant in DJF, when the results mainly range from 0.2 to 0.6 K/decade. Only the northeastern regions of the IP show not significant results in this season. The situation is different in SON, when only the regions in the eastern half of the IP show statistically significant results, with values between 0.2 and 0.4 K/decade. As happened at annual scale, the spatial distributions of the results obtained for the trend of the difference between CESM-DPLE and ERA5 series are almost identical for both ensemble sizes. Large trend differences are commonly observed in the aforementioned regions which show the most accentuated trends for ERA5 T_{mean}. However, there are also significant overestimations of the trend in regions which do not show a significant trend, or in regions exhibiting a significant but more moderate trend. These regions are the North Atlantic (in all seasons), the Black Sea (mainly in DJF and MAM) and in the northeast of the domain (in DJF and MAM). The differences between ensemble sizes are generally very small and tend to benefit ENS10 over ENS4. See, for example, the

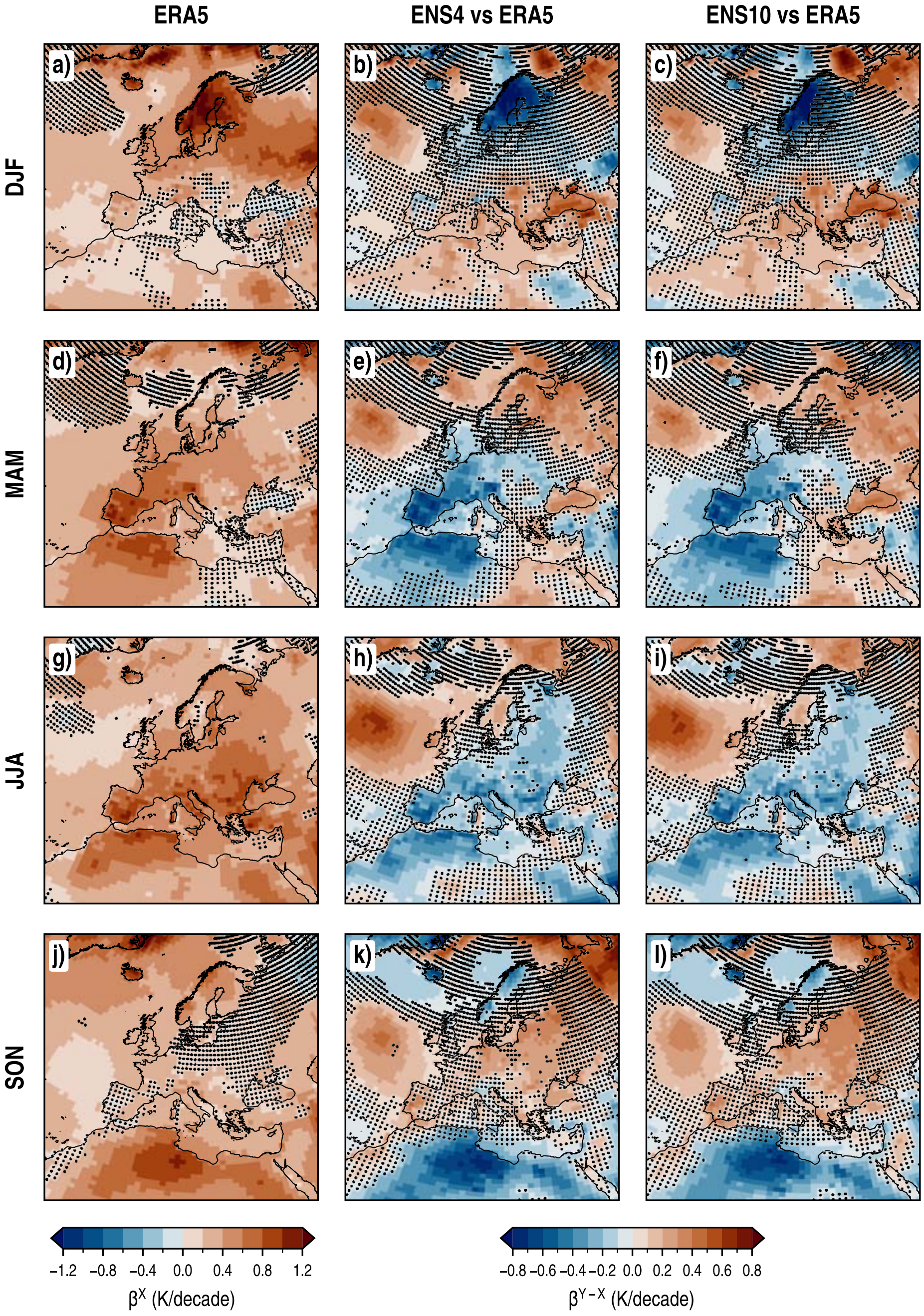

Figure 5.36: As Figure 5.35 but for DJF, MAM, JJA and SON (rows).

slight reduction of the trend of the difference series in the positive cluster over the North Atlantic for the four seasons or over the northeastern Europe in MAM and JJA. However, there are a few cases where the error in ENS10 slightly increases compared to ENS4, such as in Russia during MAM.

In the IP, the largest differences between ERA5 and CESM-DPLE have been found in MAM and JJA, with the DCPs showing a strong underestimation of the reanalysis trends. Again, this underestimation has been partially transferred to the WRF-DPLE hindcasts, as the results obtained for the WRF-DPLE spatially averaged lead time series reveal in Tables A.4 and A.5 (Appendix B.2.1) for MAM and JJA, respectively. For these downscaled lead time series, the trends are underestimated particularly in MAM, when significant trends of the difference series have been found for all NSAT fields at almost all lead times for the whole domain. Nevertheless, the best seasonal results for the WRF-DPLE hindcasts in terms of $\mathrm{MSSS_C}$ and ACC were precisely found in MAM and JJA for the three NSAT fields (see, e.g., Figures 5.2, 5.13 and 5.25 for ACC in T_{max}, T_{min} and T_{mean}, respectively). Despite the large trends of the difference series, the WRF-DPLE NSAT trends in these two seasons (Tables A.4 and A.5 in Appendix B.2.1) are still commonly larger than in DJF and SON (Tables A.3 and A.6, respectively, in Appendix B.2.1), contributing to enhancing the ACC results of the downscaled hindcasts and, as consequence, those obtained for $\mathrm{MSSS_C}$ in these seasons.

In conclusion, the analysis done in this Section has revealed the existence of errors in the representation of the T_{mean} trends by CESM-DPLE, especially at annual scale, in MAM and JJA over the IP. These errors have been partially propagated to the downscaled WRF-DPLE hindcasts during the DD simulations, hindering the achievement of better results for the skill scores analyzed in previous sections. Given the role that the temperature trend plays in predictive skill through the ACC, especially at decadal scale, all improvements which can be made in the representation of these trends in CESM-DPLE would help to improve their representation also in the downscaled product and, consequently, would increase the predictive skill of WRF-DPLE hindcasts for temperature. A possible approach to address this issue could be the use of a drift correction method which accounts for trends in the adjustment of the drift of the global DCPs. Chapter 8 has been devoted to study in more detail this topic.

5.5. Concluding remarks

This Chapter has been dedicated to the analysis of the WRF-DPLE predictive skill for T_{max}, T_{min} and T_{mean}. Several accuracy and reliability metrics have been calculated to assess not only the quality of the downscaled hindcasts, but also their added value over the CESM-DPLE subensemble, which provides WRF with the ICs and LBCs for the DD simulations, and over WRF-LE, the uninitialized downscaled experiments. The spatial distributions of these metrics for several lead times have been examined, as well as those calculated for the lead time series averaged over the regions obtained from the regionalization described in Section 3.6. The main findings are summarized in the following:

- **The signal-to-noise paradox is present in the WRF-DPLE hindcasts of NSAT, although it is weaker than for PR.** The addition of new members to the donwscaled ensemble would contribute to improving the predictive skill, especially in terms of ACC, by removing the unpredictable background noise to better capture the climate signal. However, the benefits of increasing the ensemble size may not be as pronounced as for PR, as suggested the results obtained for the RPC and shown in previous studies.

- **At annual scale, all NSAT variables show predominant positive correlations over the whole domain, although the statistical significance depends on the lead time.** The WRF-DPLE hindcasts for T_{max} show the lowest ACC and the highest RMSE values among the three NSAT variables at annual scale. Nevertheless, the ACC results are predominantly positive, with statistical significance mainly in lead years 6–9 and, especially, 2–9. On the other hand, T_{min} is the best represented variable. It always shows positive results for ACC along all lead times with a generalized statistical significance, obtaining the highest correlations also in lead years 2–9. The results achieved for T_{mean} represent a midpoint between those for T_{max} and T_{min}. The regions showing the best results are often placed in the northeast and in the southern half of the IP.

- **At seasonal scale, the highest statistically significant correlations are observed in MAM and JJA.** In both seasons, the significant positive results cover large areas of the domain in lead years 2–5, 6–9 and 2–9 for T_{min} and T_{mean}. Similar outcomes have been found for T_{max}, but generally with smaller areas showing statistical significance. In DJF, not significant negative correlations are predominant in T_{min}, while they are generally positive for T_{max} (with the exception of lead year 1), for which significant results are found in lead years 2–5 and 2–9 mainly in the southern half of the IP, the northeast and the northwest. In SON, the situation is the opposite, because positive correlations (but mainly not significant) have been found for T_{min} and fundamentally negative values are observed for T_{max}. The results achieved for T_{mean} in DJF and SON show positive outcomes for lead years 2–5 and 2–9 mainly in southern and northeastern regions, with some lead times showing also promising results for other regions in the north and the eastern half of the IP.

- **The results obtained for MSSS$_{\mathrm{C}}$ at annual scale are generally positive for T_{min} (which shows the most promising outcomes), T_{mean} and, to a lesser degree, for T_{max}.** The regions with high correlations typically show positive MSSS$_{\mathrm{C}}$ values too, as expected from the relationship between both metrics. For T_{min} and T_{mean}, there is a lack of statistical significance at lead years 1 and 2–5. Results are more robust in lead years 6–9 and, especially, 2–9 for these variables, when the best represented regions are commonly those in the southeast, the northeast and some in the northwest. The situation for T_{max} is slightly different. Although the areas covered by positive results are larger than those covered by negative ones, there are some regions to the south of the Northern Subplateau, in Sierra Morena and close to the Strait of Gibraltar with significant negative MSSS$_{\mathrm{C}}$ values. These results are caused by very strong negative CB results and low correlations in these regions. The CB outcomes are generally negative for T_{max}, favored by the low correlations. For T_{min} and T_{mean}, the absolute CB values are smaller than for T_{max}.

- **The best results for MSSS$_{\mathrm{C}}$ at seasonal scale have been found in MAM and JJA.** The spatial distributions obtained for the three NSAT variables resemble to those obtained for ACC at seasonal scale. The CB is commonly negative and significant for almost all seasons for T_{max}, with the exception of MAM. For T_{min}, the same situation is presented in DJF, but not significant results are dominant in MAM and JJA at all lead times, with the exception of lead year 1. The best CB results have been also found in MAM and JJA for T_{mean}, whereas they are typically negative and significant in DJF and SON.

- **The downscaled hindcasts for T_{max} and T_{mean} are generally reliable in the IP over almost the whole domain at annual scale, especially in lead years 2–5, 6–9 and 2–9, whereas they are reliable mainly for T_{min} in lead year 1**

in the northern regions of the IP. The reliability is determined by the not significant results obtained for CRPSS. The CRPSS results are highly influenced by those obtained for LESS. At annual scale, the significant LESS results show an underdispersion in the hindcasts for the three NSAT variables. Although this underdispersion is generally significant for T_{min}, there are not significant LESS results in several regions mainly for lead years 6–9 and 2–9 in the case of T_{mean} and, especially, for T_{max}. The hindcasts are generally less reliable in the Baetic System, in the northwestern regions and also in the northeast. The season with the largest areas showing hindcast reliability is JJA, particularly for T_{max} and T_{mean} in lead years 2–5, 6–9 and 2–9. The worst results in these terms have been obtained for the three NSAT variables in SON.

- **The largest areas showing added value of WRF-DPLE to the predictive skill over CESM-DPLE at annual scale are observed in lead years 6–9 for T_{max} and T_{mean}, but in lead year 1 for T_{min}.** In lead years 6–9, the MSSS$_{\text{G}}$ positive values cover most part of the domain for T_{max}, T_{mean} and, to a lesser degree, for T_{min}. These results are statistically significant for most part of the IP in the case of T_{max}, in some inner regions close to the Central System and to the south for T_{mean}, and mostly over the northern half of the domain for T_{min}. Very robust results indicating added value in terms of MSSS$_{\text{G}}$ have also been obtained in lead year 1 for T_{min} and in lead years 6–9 for T_{max}. In both cases, the statistically significant positive results generally cover large areas of the domain. A more limited added value can be still found at other lead times over specific regions for the three variables, particularly along the Ebro Valley or in the southwestern sector of the IP. The added value in terms of MSSS$_{\text{G}}$ is mainly motivated by the results obtained for ΔCB$_{\text{G}}$ in all cases, since the positive contribution of ΔACC$_{\text{G}}$ is fundamentally constrained to lead year 1 over large areas of the domain and lead years 2–5 and 2–9 in the Ebro Valley for T_{min}.

- **At seasonal scale, the added value of WRF-DPLE to predictive skill over CESM-DPLE tends to increase when the performance of the global hindcasts is poor.** For example, the highest MSSS$_{\text{G}}$ results for T_{max} are observed in lead years 6–9 and 2–9 in SON, when significant positive values are widespread along the IP. A similar situation is presented for T_{mean}. For T_{min}, positive MSSS$_{\text{G}}$ outcomes are predominant at all lead times in DJF, although not always showing statistical significance. There are also positive and significant results in MAM and JJA, especially for T_{min} and T_{max}, although the regions involved depend on the variable and the lead time. The surroundings of the Ebro Valley are commonly among them. In the case of T_{mean}, the significant added value is constrained to lead years 6–9 and 2–9 in JJA for central inner regions and the Ebro Valley.

- **There are not significant differences between the reliability of the WRF-DPLE and CESM-DPLE hindcasts.** The absolute values of $\Delta\mathrm{CRPSS}_{\mathrm{G}}$ are not large enough to show statistical significance. This situation occurs at both annual and seasonal scales. The sign and magnitude of $\Delta\mathrm{CRPSS}_{\mathrm{G}}$ is determined by the results achieved for $\mathrm{LESSS}_{\mathrm{G}}$, which quantify the extent to which the ensemble underdispersion or overdispersion is corrected or deteriorated in WRF-DPLE compared to CESM-DPLE. The spatial distributions of $\mathrm{LESSS}_{\mathrm{G}}$ present some structure which depends on the variable, lead time and time scale. Positive results have been obtained in these terms in lead year 1 for T_{min} and T_{mean} at annual scale over almost the whole domain. In the case of T_{mean}, these results are maintained along all lead times in the northern regions. Also at annual scale, the surroundings of the Ebro Valley are among the regions showing the best results at all lead times.

- **Scores showing the added value of WRF-DPLE to the predictive skill over WRF-LE at annual scale are mainly observed in lead year 1 for the three NSAT variables**. Nevertheless, these results do not generally present statistical significance. For T_{max}, positive $\mathrm{MSSS}_{\mathrm{U}}$ values are observed over the whole domain, with some locations showing statistical significance over Sierra Morena and the Baetic System. These results are caused by positive (but generally not significant) $\Delta\mathrm{ACC}_{\mathrm{U}}$ outcomes in the eastern, southern and northern sectors of the IP, and generalized positive outcomes for CB over the whole domain. A similar situation is observed for T_{mean}. The positive results for these three metrics, in the case of T_{min} for lead year 1, are observed in the southwestern sector of the IP and in some regions in the north. In lead years 2–5, the results are predominantly negative for the three NSAT variables (with some exceptions for T_{max} and T_{mean}) because of the confluence of a deterioration of the performance in terms of ACC and CB in large areas of the domain. Note that these results should be taken with caution since only 10 start dates (from 1990 to 1999) have been considered in this evaluation, so they are affected by a large sampling bias.

- **At seasonal scale, the largest added value of WRF-DPLE to the predictive**

skill over WRF-LE is observed in MAM. Significant positive results for MSSS_U have been found in the northern and eastern regions, in lead years 1 and 2-5, for T_max and T_mean. In the case of T_min, the area covered by the significant results is larger, even spanning almost the whole the domain in lead years 2–5. General positive and significant MSSS_U results have also been obtained in lead year 1 for T_min and, to a lesser degree, for T_mean in SON, although they are only present in some northern, northeastern and southern locations for T_max. Promising results have been found also in the eastern sector for lead years 2–5 in JJA. The worst results have been obtained in DJF. As at annual scale, these results are determined by those achieved for $\Delta\text{ACC}_\text{U}$ and ΔCB_U.

- **A generalized overestimation of the anomaly at the beginning of the control period has been found in the analysis of the regional lead time series.** This overestimation is mainly observed at lead years 2–5, 6–9 and 2–9 in several regions for the three NSAT variables. It contributes to enhancing the differences between the trends of the WRF-DPLE and AEMET lead time series, frequently observed for all NSAT variables along the regions, although they are often not statistically significant. These errors in the representation of the trends have been partially transferred to WRF-DPLE by the CESM-DPLE hindcasts during the DD process, as they are also present in the global product. With respect to the analysis of the skill scores for the spatially averaged lead time series, the results are consistent with those obtained from the grid-scale analysis for the three NSAT variables, as expected.

6

Analysis of sensitivity to extreme initial conditions of soil moisture in WRF simulations

This Chapter is dedicated to the analysis of the sensitivity of WRF simulations to extreme ICs of soil moisture. This analysis has allowed a measurement of the spin-up time for soil moisture and an evaluation of the impact of these ICs on PR and NSAT variables. The results obtained in this Chapter have led to consider the initialization of soil in the DD simulations of the DCPs for the decade 2015–2025, presented in Chapter 7.

6.1. Spin-up time and soil initialization in dynamical downscaling simulations

The concept of spin-up time and its implications in the representation of climate fields by a RCM were briefly summarized in Section 3.7. In a DD simulation, the model climatology depends on the interaction between two main factors: the LBCs provided by a GCM (or reanalysis) and the internal RCM physics and dynamics (Giorgi and Mearns, 1999). LBCs and RCM constantly interact to generate fine-scale features from coarser-resolution fields, but this generation is not instantaneous. The information provided by the LBCs starts to spread across the RCM domain from the moment of the initialization. During a certain interval of time, the bias of the RCM tends to change and eventually oscillate around an asymptotic value (Figure 6.1). The time when this asymptotic stage is reached establishes the spin-up time needed by the RCM to generate a dynamical equilibrium between the LBCs and itself. With this dynamical equilibrium, the RCM is able to represent the physical processes generated by the joint action between both LBCs and internal RCM physics with a relatively constant level of skill, while the RCM bias is constrained to a stationary value (Giorgi,

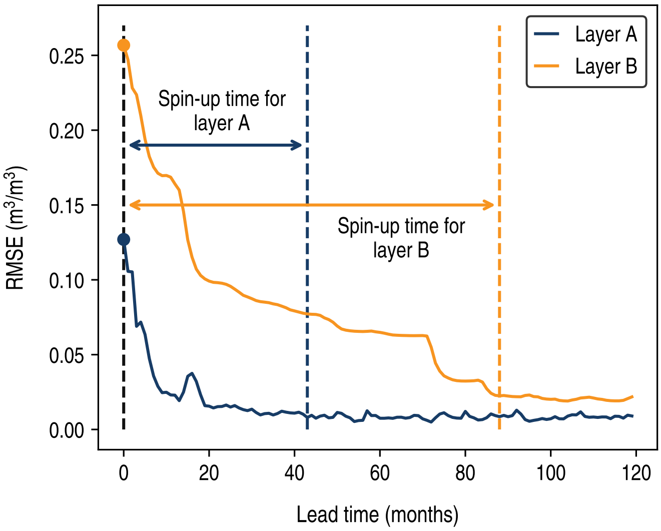

Figure 6.1: Schematic example of the spin-up stage for soil moisture. The RCM has been initialized starting from very dry ICs of soil moisture. The RMSE for soil moisture (solid lines) in a superficial layer A (blue) and a deep layer B (orange) changes until reaching an asymptotic state, whose beginning constitutes the end of the spin-up period (blue and yellow dashed lines).

2019; Giorgi and Mearns, 1999). Therefore, the output fields produced during this spin-up period should not be considered in the analysis (Laprise, 2008).

Jerez et al. (2020) provide a comprehensive list of the multiple factors which determine the length of the spin-up period. Firstly, the output field to be examined highly influences on the time required by the RCM to reach the dynamical equilibrium. The variability time scales of the components of the climate system can be very different from each other, so the length of the spin up time periods also is. While the atmospheric fields may often need lengths spanning from a few days to weeks (Gómez and Miguez-Macho, 2017; Jerez et al., 2020) the longer response of the soil variables may lead to much longer spin-up periods, spanning several years (Khodayar et al., 2015). Secondly, the configuration of a RCM influences on its internally generated variability, so the spin-up time may vary depending not only on the RCM itself, but also on the selection of the parametrization scheme (Hu et al., 2023). A third factor is the domain size, because the distance to the lateral boundary also influences on the ability of the RCM to generate fine-scale features, which ultimately depends on the GCM resolution, regardless of the RCM resolution (Matte et al., 2017). This aspect is also known as the spin-up distance (Laprise, 2008; Matte et al., 2017). Jerez et al. (2020) mention the discrepancies between the RCM and GCM physics (e.g., Turco et al., 2013) as another factor which may impact on the length of the spin-up period through the ICs supplied to the RCM (Jacob and Podzun, 1997). The meteorological situation or the presence of extreme conditions, such as very wet or dry soil ICs, can also affect the spin-up requirements (Seck et al., 2015; Yang et al., 2011).

The importance of feedbacks and general interaction between atmosphere and land surface in determining how climate evolves (Jaeger and Seneviratne, 2011; Seneviratne et al., 2010) highlights the relevance of the spin-up time for studies in which these coupling processes are present. The land surface-atmosphere coupling is highly influenced by the role of soil moisture in the water cycle and in the partition of the incoming radiative energy into sensible and latent heat fluxes. The impacts of soil moisture on climate processes are particularly meaningful in regions characterized by a limitation in soil moisture content (Seneviratne et al., 2010), such as the IP (García-Valdecasas et al., 2020b). Extreme soil moisture conditions may influence on the trends and occurrence of climate extremes of temperature and, to a lesser extent, precipitation in Europe. Jaeger and Seneviratne (2011) showed that wetter soil conditions would generally lead to lower $T_{\max}$ in arid areas, whereas drier soils would lead to higher values in humid regions. The same study also found a positive feedback between precipitation and soil moisture in terms of the frequency of wet days. The climate persistence (or memory) induced by soil moisture, which is a reservoir of water and energy, acts also as a driver of the land surface-atmosphere interactions (Seneviratne et al., 2010) and gives this field a potentially important role in applications which require accurate initial states to make actual predictions of the climate evolution, such as DCPs (Bellucci et al., 2015). In this line, the initialization of RCMs by using soil ICs internally consistent with the RCM has shown to provide some added value in the representation of variables such as temperature or precipitation (Kothe et al., 2016), and may be particularly useful in the context of DD in a soil moisture-limited region such as the IP.

6.2. Noah Land Surface Model and soil properties

Among the all options available in WRF to represent the physical processes occurring in land surface and their interactions with the atmosphere, the WRF configuration used in this Thesis considers the Noah Land Surface Model (Noah LSM; Chen and Dudhia, 2001; Ek et al., 2003; Wang et al., 2010). In the WRF framework, LSMs take information from atmospheric variables, radiative forcing, internal land fields and land surface properties to generate heat and moisture fluxes at a grid cell level over land and sea ice. Although the LSMs do not provide trends, they update the land state defined by variables such as soil temperature and moisture, skin temperature, snow cover or vegetation properties at each model time step. The exchange of information between adjacent grid cells is done along the vertical column, without any communication over the horizontal level (Skamarock et al., 2008). Noah LSM uses four soil layers with depths 0–10 cm, 10–40 cm, 40–100 cm and 100–200 cm, and only a canopy layer. The root zone spans the three layers above the 100 cm depth level, whereas the deepest fourth layer works like a reservoir with drains water at the bottom by the action of gravity. The prognostic variables are the soil moisture and temperature profiles, canopy moisture and snow storage (Chen and Dudhia, 2001). Noah LSM also includes a evapotranspiration scheme, runoff and accounts for soil textures, land cover and vegetation properties (Skamarock et al., 2008).

The physical properties of the soil in each grid cell are established by the land cover and soil textures. These properties determine aspects such as how much water can be stored in soil, how fast the thermal energy is transferred through soil layers or the portion of solar energy which is reflected or absorbed by the surface. Thus, land cover and soil textures are decisive for the LSM in the calculation of the heat and moisture fluxes at the surface (Wang et al., 2010). The WRF set up for this Thesis includes the modified IGBP MODIS 20-category vegetation classification as the land cover dataset (Figure 6.2; Friedl and Land Team/EMC/NCEP, 2008; Friedl et al., 2010) and the hybrid STATSGO/FAO 16-classes categories (FAO/UNESCO, 1978; FAO/USDA, 2002; Miller and White, 1998) as soil textures. While the land cover classes determine vegetation/surface-related properties, such as the green vegetation fraction, albedo or emissivity, the soil textures are more involved in characteristics related to soil moisture, such as the field capacity, wilting point or the soil thermal conductivity/diffusivity (see "Vegetation parameters" and "Soil parameters" sections in "Unified Noah LSM", n.d.).

The soil textures have been used to define the soil moisture ICs used to conduct the sensitivity experiments analyzed in this Chapter. The hybrid STATSGO/FAO soil textures are determined by the percentage of clay, silt and sand (Figure 6.3a; Soil Survey Division Staff, 1993) in the composition of the soil categories defined by FAO/UNESCO (1978) over the whole world excepting the conterminous United States, where the soil categories defined by Miller and White (1998) are used instead. Clay, silt and sand are particles of mineral material which can be categorized in terms of their sizes. The sand type comprises particles with sizes from 0.05 mm to 2 mm, the size of silt particles ranges from 0.002 mm to 0.05 mm and clay particles have sizes smaller than 0.002 mm (Soil Survey Division Staff, 1993). The hybrid STATSGO/FAO

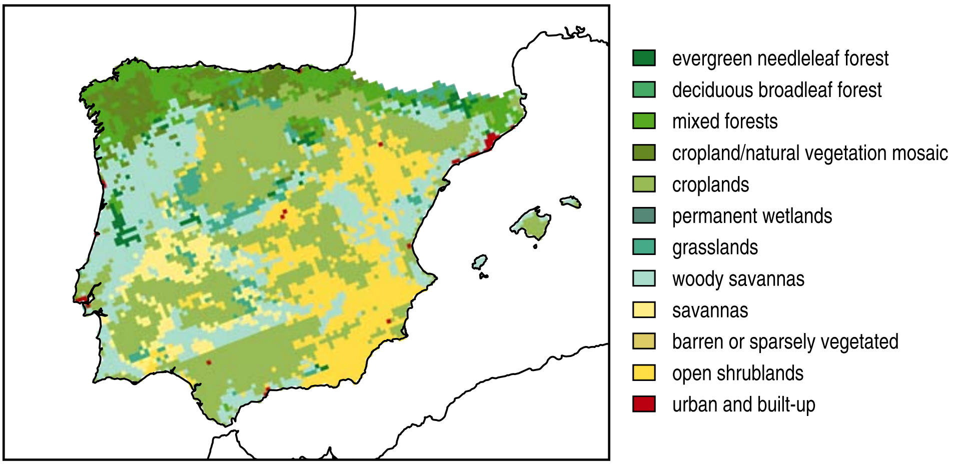
evergreen needleleaf forest
deciduous broadleaf forest
mixed forests
cropland/natural vegetation mosaic
croplands
permanent wetlands
grasslands
woody savannas
savannas
barren or sparsely vegetated
open shrublands
urban and built-up

Figure 6.2: Dominant land cover classes used by Noah LSM in WRF simulations. Data retrieved from the modified IGBP MODIS 20-category vegetation classification (Friedl and Land Team/EMC/NCEP, 2008; Friedl et al., 2010).

dataset provides the spatial distribution of the dominant soil textures on a top and a bottom layers. However, only the top texture layer is considered by Noah LSM (Figure 6.3b), which is globally applied to the four soil layers. The domain is mainly covered by loamy soil materials (loam, sandy loam, sandy clay loam and clay loam; Soil Survey Division Staff, 1993), with loam spanning most part. There are a few regions with the clay texture, mainly placed in the south, and other, narrower, with sand as the dominant texture.

6.3. Experimental design

The experiments conducted in this Chapter aim at analysing the sensitivity of WRF simulations to extreme ICs conditions of soil moisture. The experimental framework can be divided into three phases:

a) calculation of the soil moisture ICs;
b) initialization of DD simulations with the soil moisture ICs previously calculated;
c) analysis of the sensitivity of the simulations to these ICs in terms of the spin-up time required by soil moisture, PR and NSAT.

These simulations were conducted by using the same WRF configuration described in Section 3.1, but with ERA-Interim reanalysis data providing the RCM with the information for the ICs and LBCs of all variables with the exception of soil moisture.

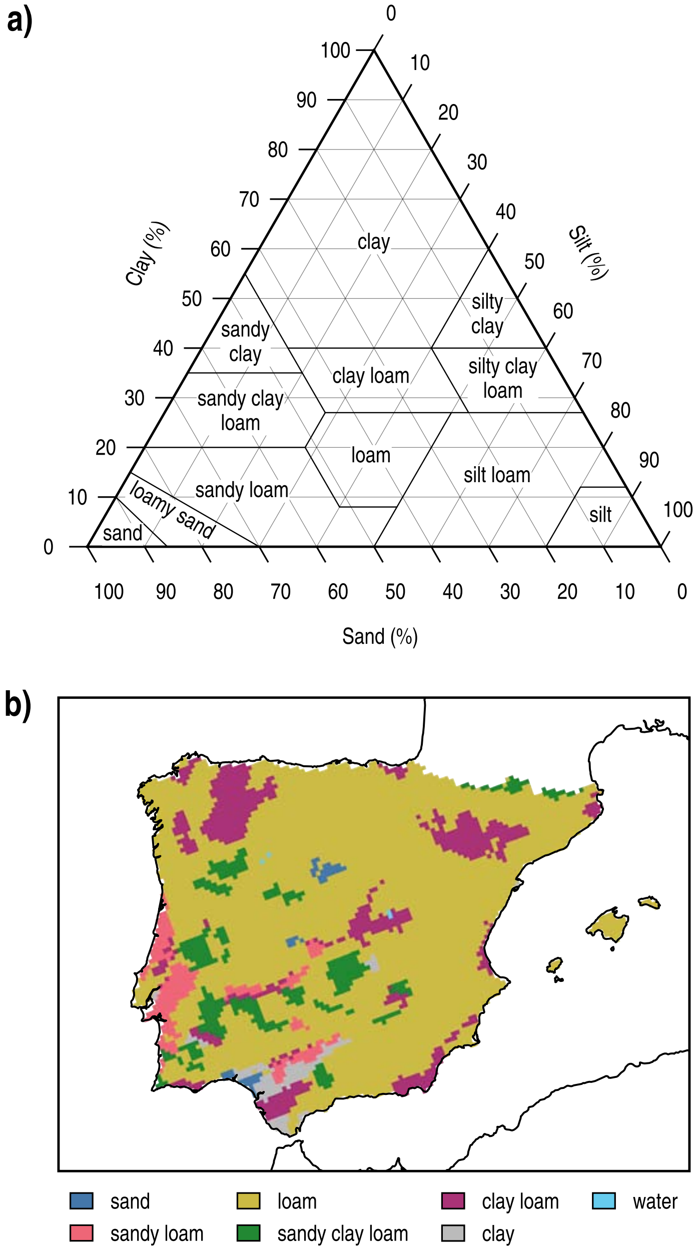

Figure 6.3: Dominant soil textures used by Noah LSM in WRF simulations. a) Chart which shows the composition of the texture classes. Composition information retrieved from Soil Survey Division Staff (1993). b) Soil texture distribution in the IP and Balearic Islands. Data retrieved from FAO/USDA (2002).

The ICs of soil moisture were set to represent three different soil conditions in terms of the moisture content: a wet soil, a dry soil and a very dry soil. With wet and very dry ICs, the response of the WRF simulations to the initialization from extreme soil moisture conditions can be evaluated, and the influence of the level of dryness can be addressed by incorporating the dry ICs. Since this analysis is oriented to provide insights on the impact of soil initialization on the prediction skill of dynamically downscaled DCPs, the period under study encompass 10 years. In addition, the simulations have been initialized in two different dates to also test the impact of the moment of the initialization on the simulated climate: January (boreal winter) and July (boreal summer). Therefore, there are two different periods considered here, one ranging from 1990-01-01 to 1999-12-31 and another ranging from 1990-07-01 to 2000-06-30. A control simulation in which ERA-Interim supplied all initial and boundary information was also conducted, starting in 1982-01-01 to account for a 8-year spin-up period. The length of the spin-up period was chosen considering the results obtained by Khodayar et al. (2015), who showed that soil moisture may need a maximum of 7.5 years for the deepest soil layers of their RCM, down to a depth of 15 m, in DD simulations over Europe. Although this spin-up period length is required only in the Scandinavian Peninsula and the authors show that a length around 80 months (~6.6 years) may be enough in the IP, their experimental framework is different from that considered in this Thesis: their analysis is done for the IP on average, whereas in this Thesis it has been done at grid-point level, different RCMs are used, with also different LSMs, parametrization schemes, ICs, etc. Therefore, the spin-up requirements may also differ. For this reason, a spin-up period of 8 years was considered in the control simulation to guarantee a fully equilibrated soil at the beginning of the experimental periods.

The soil moisture ICs were calculated by combining part of the physical properties which characterize the dominant soil texture classes in the domain, summarized in Table 6.1, with the soil moisture index (SMI; Betts, 2004; Seneviratne et al., 2010) given by

$$\mathrm{SMI} = \frac{\theta - \theta_{\mathrm{WP}}}{\theta_{\mathrm{FC}} - \theta_{\mathrm{WP}}}\,, \qquad [6.1]$$

where θ is the soil moisture, θ_{FC} identifies the field capacity and θ_{WP} denotes the wilting point. The field capacity θ_{FC} is commonly defined as the amount of water held in soil after the excess water has drained away by action of the gravity. On the other hand, the wilting point θ_{WP} represents the minimum moisture content below which the plant activity abruptly decreases (Hillel, 1998). The SMI is a measure of

Table 6.1: Soil moisture values for the wilting point (θ_{WP}) and field capacity (θ_{FC}) which correspond to each soil texture class in the IP.

Texture class	θ_{WP} (m^3/m^3)	θ_{FC} (m^3/m^3)
Sand	0.010	0.192
Sandy loam	0.028	0.283
Loam	0.066	0.329
Sandy clay loam	0.069	0.315
Clay loam	0.103	0.382
Clay	0.138	0.412
Water	0.0	0.0

the moisture content available in the soil and generally ranges from 0 ($\theta = \theta_{WP}$) to 1 ($\theta = \theta_{FC}$). In riverbank areas or after extreme precipitation episodes, the soil saturation ($\theta > \theta_{FC}$) may lead to SMI values higher than 1 (Seneviratne et al., 2010).

To obtain the soil moisture ICs for the wet soil, θ_W, the SMI was set to 1 and, consequently, the soil mositure ICs came from $\theta_W = \theta_{FC}$. On the contrary, the very dry ICs, θ_{VD}, were obtained from the soil moisture values corresponding to SMI = 0, i.e., $\theta_{VD} = \theta_{WP}$. Since SMI = 0.5 constitutes the midpoint between full wetness and dryness, the ICs for the dry soil, θ_D, have been obtained from setting SMI = 0.25, so $\theta_D = (\theta_{FC} + 3\theta_{WP})/4$. The spatial distributions of the soil moisture ICs are available in Figure 6.4. The spin-up times analyzed in Section 6.4 for soil moisture, PR and NSAT variables have been obtained by following a similar approach to that described in

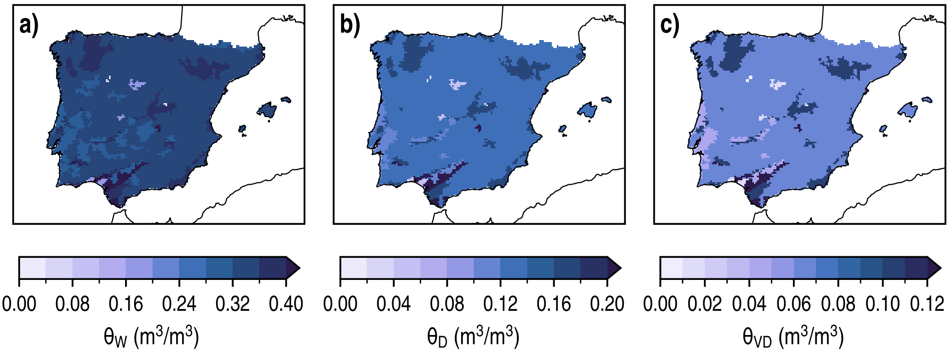

Figure 6.4: ICs of soil moisture (θ) in the sensitivity experiments for the **a**) wet, **b**) dry and **c**) very dry soils. These ICs have been obtained by combining the SMI in Eq. [6.1] with the physical properties of each soil texture in Table 6.1.

Khodayar et al. (2015), which can be split into three parts. Firstly, the absolute error between the monthly time series of the experimental and control simulations has been calculated at each grid point and time step, for each variable and soil layer (in the case of soil moisture). Secondly, the median value of this absolute error during the last 5 years of the experimental period has been calculated. Finally, the spin-up time for each grid cell, soil layer and variable, needed to reach the dynamical equilibrium, has been defined as the time required by the absolute error time series to cross below the median value for the first time.

6.4. Results and discussion

6.4.1. *Analysis of the spin-up time of soil moisture*

This Section is devoted to assess the spin-up time of soil moisture from the sensitivity experiments conducted with WRF. The results obtained from the experiments initialized on 1990-01-01 and 1990-07-01 are depicted in Figures 6.5 and 6.6, respectively. The spatial distributions show the time required to get a dynamical equilibrium state in terms of soil moisture, depending on the soil layer, the soil moisture ICs and the initialization date. The results have been only shown for the IP, excluding the Balearic Islands because a mistake in setting the land mask for the DD simulations attributed initial values equal to 0 for soil moisture in all sensitivity experiments, leading to inconsistent results in this region.

The spin-up time required by soil moisture to guarantee the dynamical equilibrium in all layers after starting from extreme ICs in the IP is 8 years, determined by the maximum value observed in both Figures 6.5 and 6.6. This value is larger than that obtained by Khodayar et al. (2015). As mentioned Section 6.3, the difference between both results can be attributed to the differences between the experimental frameworks. Depending on the specific scopus of the study, shorter spin-up periods may be considered, since the impact of a not equilibrated soil in atmospheric variables, such as PR or NSAT, is not as determinant as it would be for the analysis of hydrological fields. For NSAT and PR, a spin-up time around 1 week may be enough in some cases, as showed by Jerez et al. (2020). The spin-up requirements for these fields, PR and NSAT, in our experimental framework are analyzed in Section 6.4.2.

The impact of the soil layer depth on the length of the spin-up period is clear in Figures 6.5 and 6.6, especially when comparing the first three layers with the fourth one. The soil moisture in the most superficial layers is subjected to a higher variability because the influence of the interaction between the atmosphere and

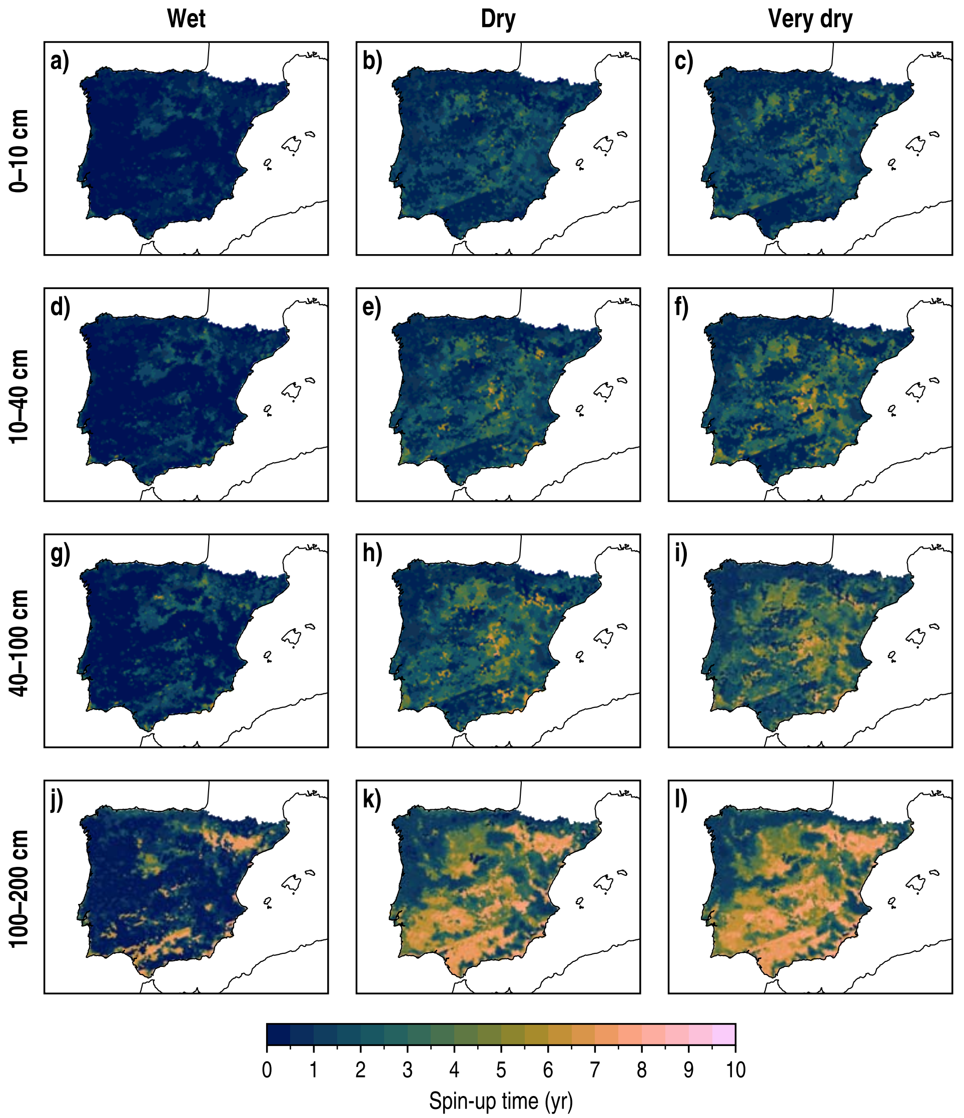

Figure 6.5: Spin-up time for soil moisture depending on the soil layer (rows) and the soil moisture ICs (columns) for the sensitivity experiments initialized on 1990-01-01.

them on determining the evolution of the soil variables is more immediate than in deeper layers (Jerez et al., 2020). The water coming from precipitation needs more time to reach the deepest soil layers, making the soil state set by the ICs more persistent and consequently increasing the spin-up time (Khodayar et al., 2015). For example, the differences among soil layers in results depicted over the Southern

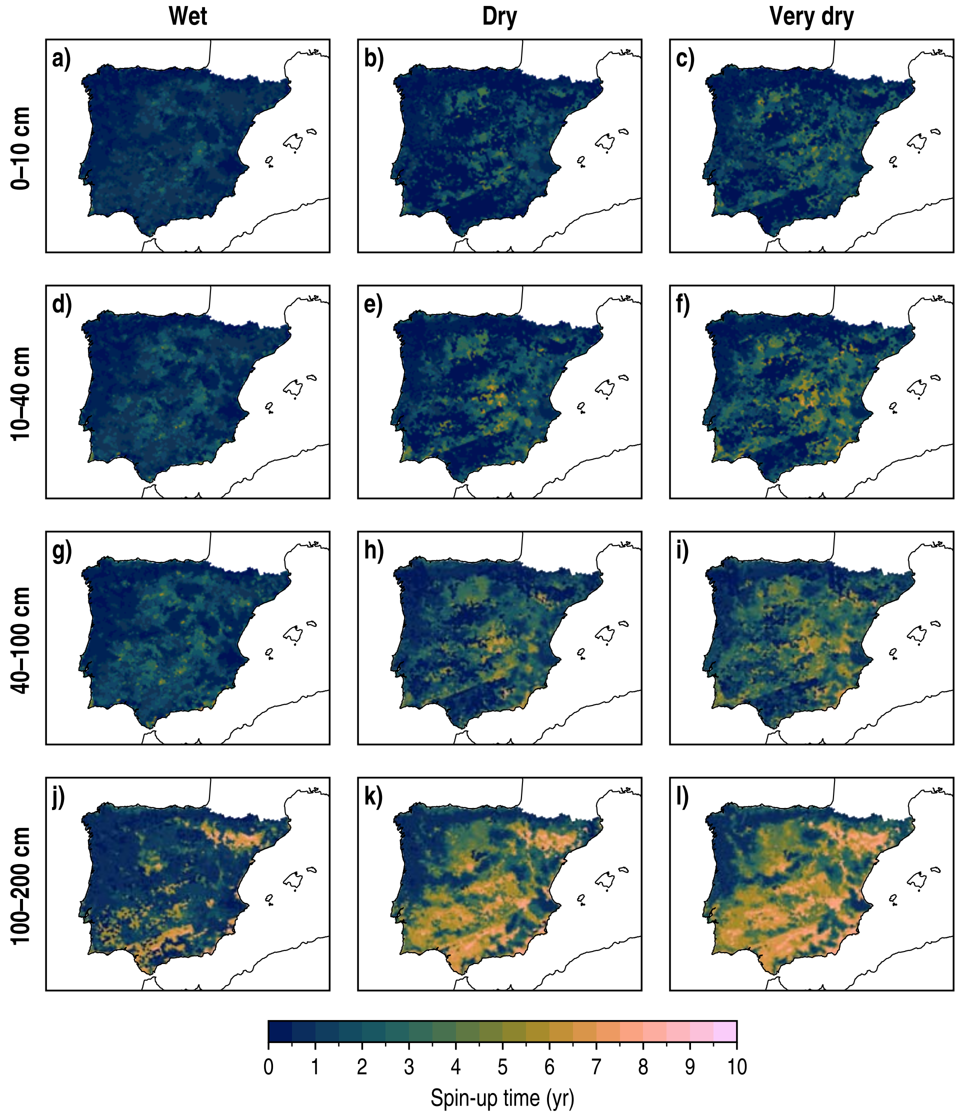

Figure 6.6: As Figure 6.5 but for the sensitivity experiments initialized on 1990-07-01.

Subplateau are high enough to exhibit appreciable changes in the spin-up times in the dry experiments initialized in January (Figures 6.5b, 6.5e, 6.5h and 6.5k). While the longest spin-up period is around 5 years in the first layer, the maximum length increases to approximately 7 years in the second layer, to about 7.5 years is the third one and up to 8 years in the deepest layer.

Another aspect which influences on the spin-up time is the soil moisture content defined by the ICs. The soil memory generally increases with the decrease of the amount of water stored by the soil, showing longer spin-up periods the regions in experiments with drier initial values of soil moisture. The soil capacity to hold water is higher in drier conditions (Seneviratne et al., 2010), so it is expected that the absence of soil water slows down the transport of the water from precipitation through the soil layers (Khodayar et al., 2015). Therefore, the dynamical equilibrium state is normally reached later in the experiments with very dry ICs compared to those with wet ICs. The experiments with dry ICs are in a midpoint between the wettest and the driest ones.

The effects of the ICs on the spin-up time exhibit some seasonality, as the impact of the initial moisture content on the evolution of soil moisture is modulated by the experiment start dates. The wet experiments initialized in January show shorter spin-up times than those initialized in July. For instance, looking at the first soil layer in the western part of the IP in Figures 6.5a and 6.6a, the spin-up time slightly increases from 0–0.5 years to 0.5–1 years, respectively. This behaviour is observed in most part of the domain through all soil layers. On the other hand, the dry and very dry experiments initialized in July typically need shorter spin-up periods than the experiments initialized in January. As an illustrative example, look at the Guadalquivir Valley in Figures 6.5l and 6.6l, where the spin-up time generally turns from 6.5–7 years to 7–7.5 years, respectively. This seasonality may be related to the magnitude of the perturbation of soil moisture ICs with respect to the values which correspond to the equilibrated soil state at the same moment. Figures 6.7 and 6.8 show the differences between the ICs of the experimental soil moisture, θ_{EXP}, with EXP $= \{\text{W}, \text{D}, \text{VD}\}$, and the control soil moisture, θ_{CTL}, on 1990-01-01 and 1990-07-01, respectively. In wet experiments, the differences $\theta_{\text{EXP}} - \theta_{\text{CTL}}$ are more pronounced for the simulations initialized in July. Therefore, the shock produced by initializing with wet conditions at this date leads to a longer time to reach an equilibrated state. Indeed, while $\theta_{\text{EXP}} - \theta_{\text{CTL}}$ are undoubtedly positive for experiments initialized in July (excepting in a few locations with water as soil texture), negative differences appear in January-initialized experiments, mostly in the deepest layer, indicating that the perturbed ICs are slightly drier than the control soil. The opposite situation occurs when examining the dry and very dry experiments. In these cases, the deviations from the control states at the initialization dates are higher in the experiments initialized in January, explaining the generally longer spin-up periods needed compared to the July counterparts. The differences between January and

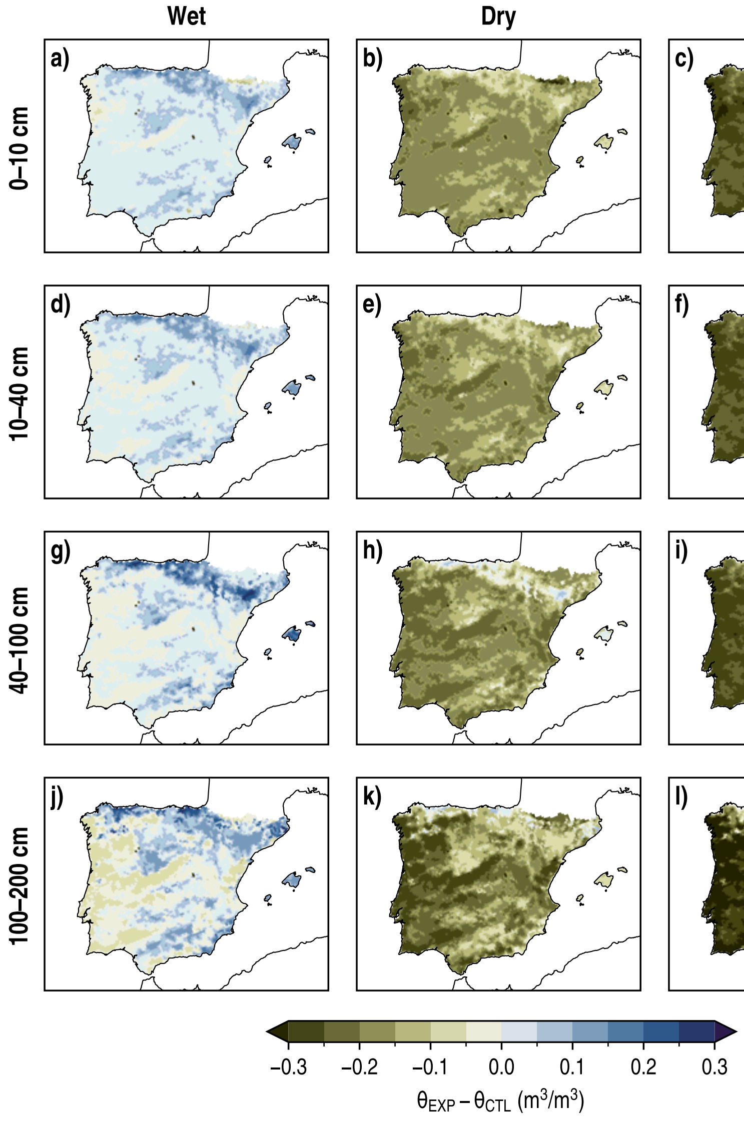

Figure 6.7: Differences between the ICs of soil moisture for the sensitivity experiments (columns) and the control simulation in each soil layer (rows). The sensitivity experiments were initialized on 1990-01-01.

July control soil moisture are partly caused by the differences in the meteorological conditions of the preceding months (Figures 6.9a to 6.9d). In the IP, as a soil moisture-limited region, the occurrence of precipitation events and low temperatures in winter positively influences on moisture content in soil, whereas summer low precipitation

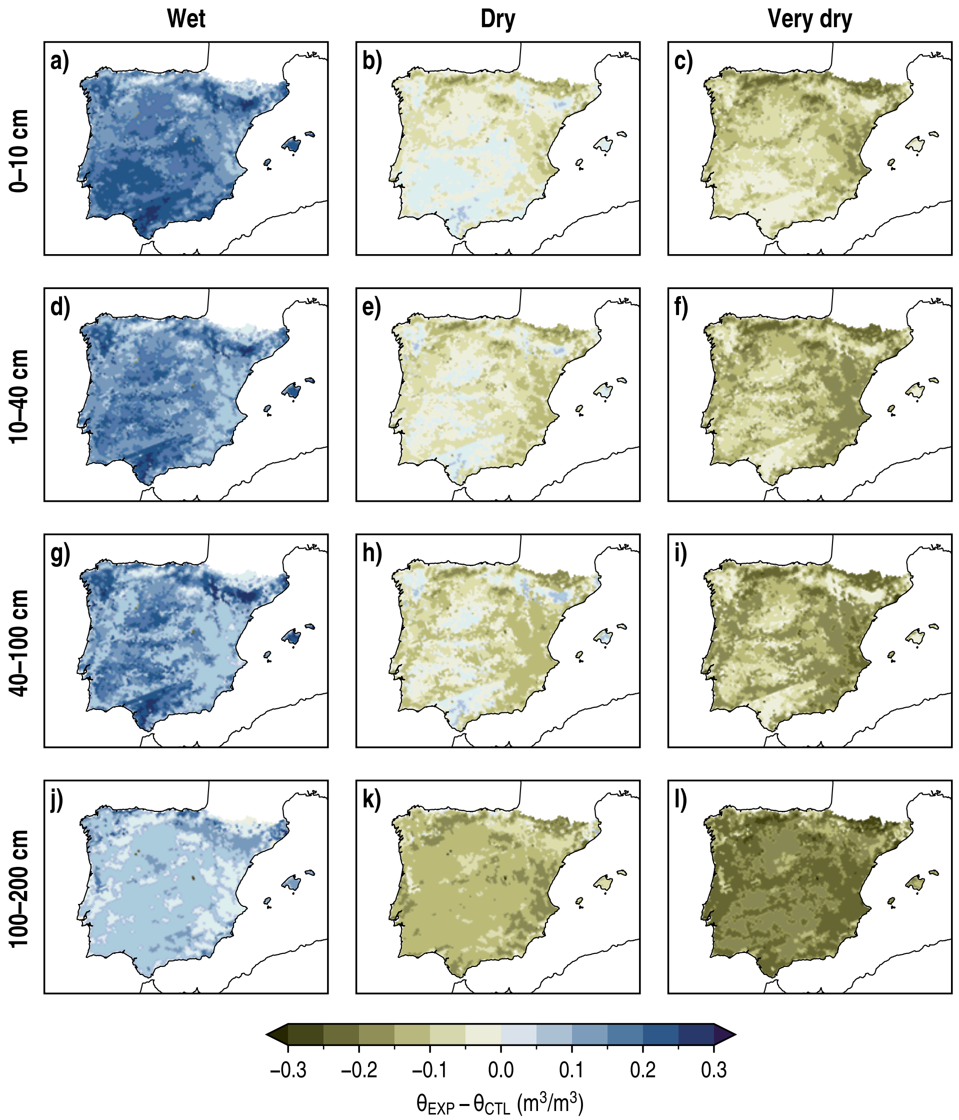

Figure 6.8: As Figure 6.7 but for the sensitivity experiments initialized on 1990-07-01.

rates and high temperatures strongly contribute to decreasing soil moisture through evapotranspiration (Seneviratne et al., 2010).

The length of the spin-up period also shows some spatial variability, which might be attributed to some extent to the spatial distributions of PR, T_{mean} and soil textures. The longest spin-up periods are mainly observed over the Guadalquivir and Ebro

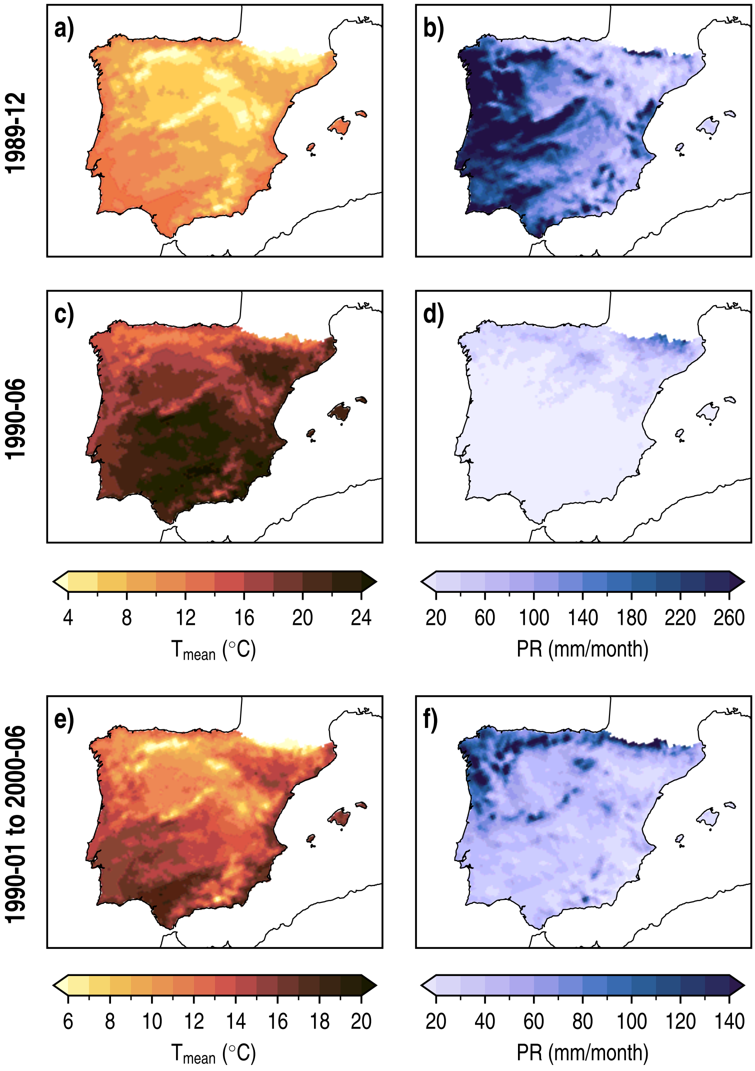

Figure 6.9: Spatial distributions of monthly control T_{mean} (left column) and PR (right column). Top and middle rows correspond to the monthly fields in the previous months to the initialization dates of the sensitivity experiments. Panels in the bottom row show the average of the monthly fields over the control period, which starts in 1990-01-01 (start date of the experiments initialized in January) and ends in 2000-06-30 (end date of the experiments initialized in July).

Valleys for all sensitivity experiments, where clay and clay loam are among the dominant soil textures (Figure 6.3a). Clay soils are characterized by having a low hydraulic conductivity (Chen and Dudhia, 2001), a magnitude proportional to the amount of water which is transported through them along time (Soil Survey Division

Staff, 1993). Thus, soils with high composition of clay, such as clay and clay loam textures (Figure 6.3b), are expected to have a higher climate persistence because more time than for other soil types might be required to observe changes in soil moisture content. This phenomenon could be amplified if these soils are additionally placed in regions characterized by atmospheric conditions which lead to limitations in soil moisture content (relatively low precipitation rates and high temperatures), just as it happens in the cases of Ebro and Guadalquivir Valleys on average over the whole control period (see Figures 6.9e and 6.9f). The same combination of atmospheric conditions and soil textures may also explain the long spin-up periods observed in dry and very dry experiments in the fourth layer over mostly the southern half of the IP. Nevertheless, since hydraulic conductivity is proportional to soil moisture content (Chen and Dudhia, 2001), different precipitation/temperature regimes may lead to different spin-up times in soils sharing the same texture class. For this reason, looking at Figures 6.5j to 6.5l, for example, a spin-up period around 7 years is needed in Ebro Valley, but only 1 year or less may be required in part of the northwestern area, characterized by higher precipitation rates and lower temperatures (Figures 6.9e and 6.9f), although the soil texture is the same in both regions (Figure 6.3a).

There is also a curious phenomenon which occurs over the Guadalquivir and Ebro Valleys, alongside some locations in the Northern Subplateau or in the southwest of the IP. In dry and very dry experiments, these regions are among those which show the longest spin-up periods for the deepest layer, but they do not for the upper three layers (Figures 6.5 and 6.6). On the contrary, in these layers, the spin-up times are shorter than for the surrounding regions. This marked contrast between the upper three and the fourth layers is observed regardless the initialization date, but it is not present in wet experiments. The land cover may be partly responsible of these results; indeed, the shape of the land cover structure in Figure 6.2 somewhat resembles the patterns observed in Figures 6.5 and 6.6 for dry and very dry experiments in the first three layers, especially in the Guadalquivir Valley. The areas which show this behaviour are covered by croplands. This land class is characterized by a root zone spanning the upper three soil layers (excluding the deepest layer) and a very low value of the minimum stomatal resistance, among other features (see the "Vegetation parameters" section in "Unified Noah LSM", n.d.). According to Seneviratne et al. (2010), the stomatal resistance plays a crucial role in the evapotranspiration process represented by LSMs. The stomatal resistance is a measure of the resistance to water transport offered by stomata, which are pores on the vegetation epidermis that regulate the exchange of water and CO_2 with atmosphere. A low (high) value of stomatal resistance positively (negatively) affects the magnitude of the water evapotranspirated from vegetation and, consequently, has an impact on the soil moisture content (Chen and Dudhia, 2001). In regions habituated to exhibit minor resistance to this type of evapotranspiration (e.g., regions covered by croplands), thus, the soil state in dry and very dry scenarios generally tends to be closer to the control state than in the case of regions where the stomatal resistance is higher. In consequence, the former would need less time than the latter to reach the dynamical equilibrium. This phenomenon is observed only in the upper three layers because the root zone of croplands is constrained to them. In the fourth layer, where there is no water to be taken by vegetation, the low hydraulic conductivity of clay soils produce the long spin-up periods observed in Figures 6.5k, 6.5l, 6.6k and 6.6l.

6.4.2. *Analysis of the spin-up time of precipitation and near-surface air temperature*

The effect of the soil moisture ICs on the spin-up of PR, T_{max}, T_{min} and T_{mean} is displayed in Figures 6.10 and 6.11 for simulations started in January and July, respectively.

The variable which shows the shorter spin-up periods is PR. For experiments initialized in January (Figures 6.10a to 6.10c), PR generally needs a period below 2 months to reach the dynamical equilibrium. There are a few locations which require a longer spin-up time, mostly situated in the northwestern sector of the IP, but they conform a sparse minority and their spin-up times are commonly below 1 year. Although these regions are slightly more present in the very dry experiment, there are not big differences between the three scenarios, suggesting that the role of the model internal variability prevails over the imposed ICs of soil moisture in determining the evolution of PR. For this variable, the impact of the soil moisture conditions is more evident in experiments initialized in July (Figures 6.11a to 6.11c), when the wet experiment seems to be the most affected. The spin-up time is usually longer for the wet scenario than in the dry and very dry experiments, as occurred for soil moisture in Figures 6.6a to 6.6c, because the deviation from the control state for wet ICs is larger in July than in January. Large areas of the domain, mainly placed in the northern half of the IP, show spin-up periods from 3 to 10 months for wet ICs (Figure 6.11a). The area showing similar results is narrower for dry and very dry ICs (Figure 6.11b, respectively). In all cases, the lowest values of the spin-up time have been found in the southern regions.

The most affected variable by the shock associated to the soil moisture ICs is T_{max}, since it shows the longest spin-up times among the analyzed fields. For experiments

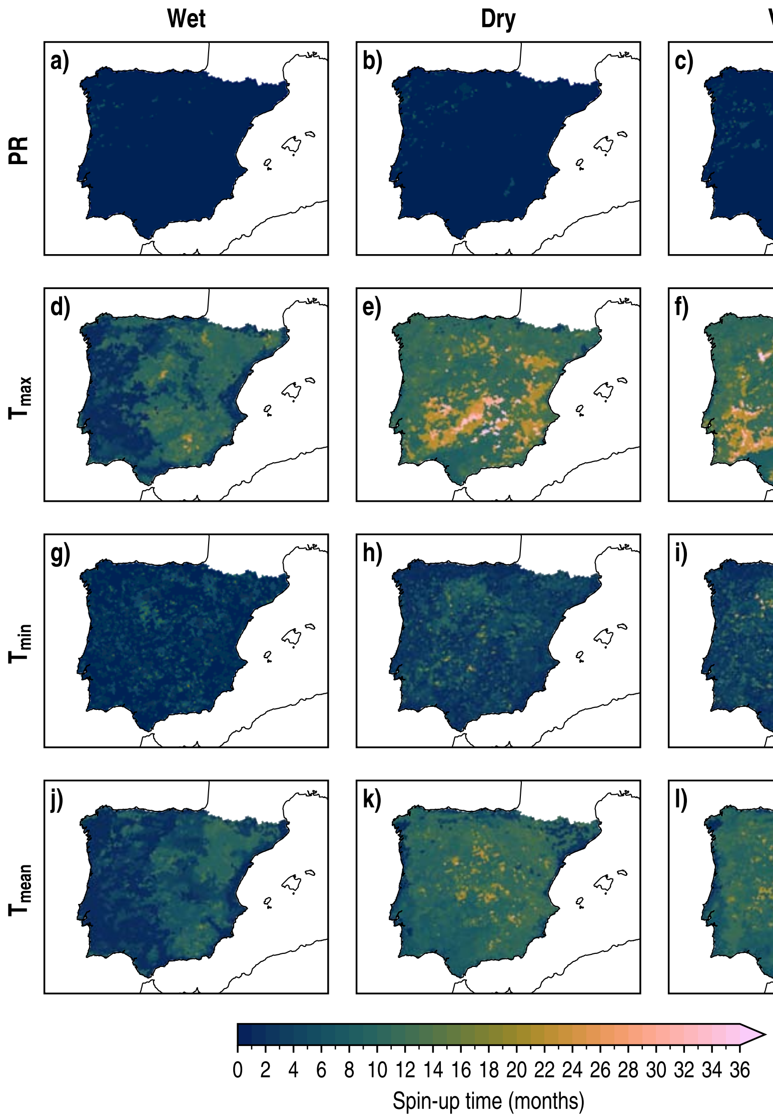

Figure 6.10: Spin-up time for PR, T_{max}, T_{min} and T_{mean} (rows) depending on the soil moisture initial conditions (columns) for the sensitivity experiments initialized on 1990-01-01.

initialized in January (Figures 6.10d to 6.10f), the longest periods have been found for the dry and very dry experiments, with spin-up times mainly above 10 months over the whole domain. Higher values can be observed in these scenarios over eastern regions, some along the Mediterranean coast, those locations between the Guadalquivir Valley and the Central System and some northern areas. In these zones,

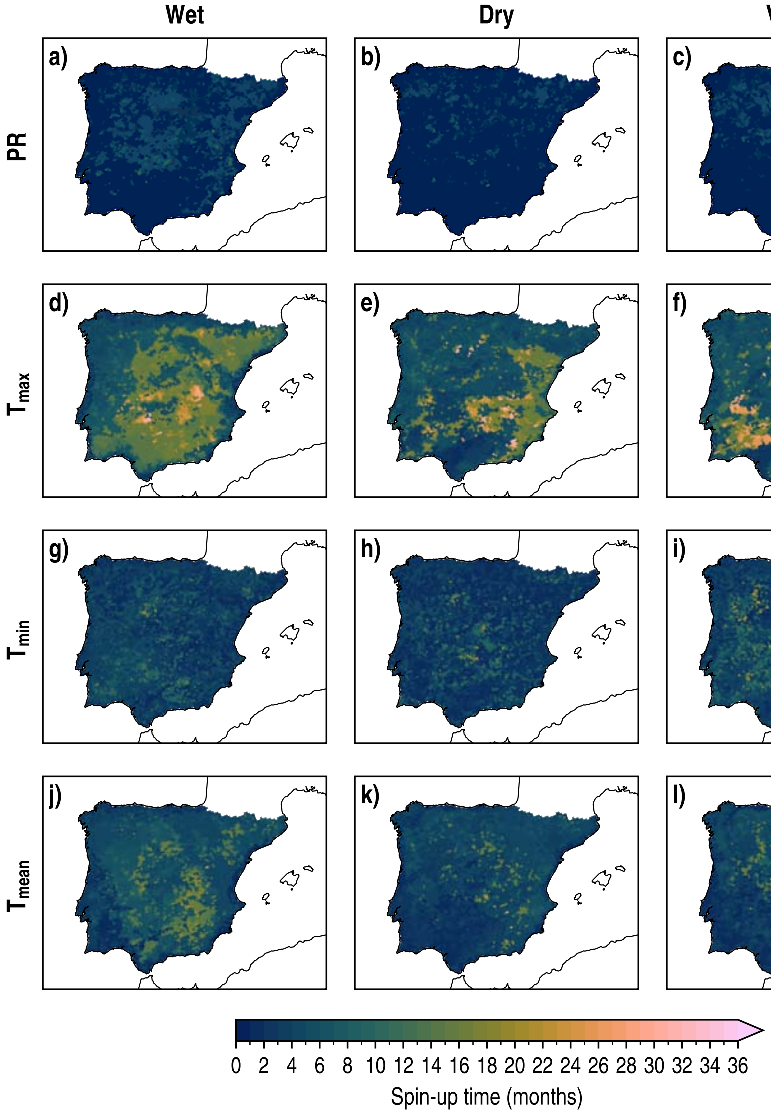

Figure 6.11: As Figure 6.10 but for the sensitivity experiments initialized on 1990-07-01.

the length of the period is typically between 20 and 30 months, although some sparse locations can reach values even above 36 months up to a maximum of 45 months in the most extreme case. The very dry experiment generally shows periods longer than the dry simulation. The spatial distributions of spin-up time show a different structure when the experiments are initialized in July (Figures 6.11d to 6.11f). In dry and very dry scenarios, the spin-up time decreases compared to the other initialization date. In contrast, it generally increases in the wet experiment, when differences even above 10 months can be found between both initialization dates. The similarities between the spin-up time spatial distributions for T_{max} in Figures 6.10 and 6.11 and soil moisture in Figures 6.5 and 6.6 in the upper three layers reflect the strong relationship existing between both variables in the IP, showing the existence of land surface-atmosphere coupling processes involving these two variables, as already documented in previous studies (see, e.g., García-Valdecasas et al., 2020b; Jaeger and Seneviratne, 2011; Lorenz et al., 2012; Seneviratne et al., 2010).

The results obtained for T_{mean} and, especially, T_{min} show that these variables are not as influenced by the extreme soil moisture ICs as T_{max}. This finding is consistent with the study made by Vidale et al. (2007), who showed that drier soil conditions lead to a broadening of the diurnal temperature cycle as consequence of the increase of T_{max}, with a lower contribution of T_{min}. The spatial distributions of T_{min} show noisy patterns regardless of the ICs and the initialization date, and the spin-up times are much shorter than for T_{max}. In January, as usual, the longest spin-up periods are observed in the dry and very dry scenarios. Although some scattered locations might present spin-up times around 18 months and above in both experiments, the results are commonly below 12 months, approximately. The regions showing these values are mainly situated over the Northern Subplateau in both experiments. On the other hand, periods with lengths typically below 7 months are found in the wet experiment, with a considerable portion of the domain needing less than 2 months of spin-up time. The experiments initialized in July show similar characteristics: noisy patterns and spin-up times normally below 12 months. In this case, dry experiments show slightly longer spin-up times than those initialized in January, whereas it depends on the region for the very dry simulations. On the other hand, the wet experiments clearly exhibits longer spin-up times for the initialization in July. After PR, T_{min} is the variable least affected by soil moisture ICs. Some spatial noise is observed also in T_{mean} panels for initialization in January (Figures 6.10j to 6.10l), but with longer spin-up times compared to T_{min}. In dry and very dry scenarios, values above 9 months are generalized over the whole domain, with maximum lengths starting from 24 up to 32 months in a few locations. In the wet experiment, there is a clear difference between the eastern and the western sectors of the IP, with the former showing spin-up times between 5 and 10 months and the latter with values generally below 2 months. For the initialization in July, the longest spin-up times are commonly found once again in the wet experiment, with maximum values which usually do not surpass 17 months mainly situated in the Southern Subplateau and some locations in the Northern Subplateau, the northeast and some southern locations.

6.5. Concluding remarks

The simulations analyzed in this Chapter have contributed to stressing the importance of land surface-atmosphere interactions in the evolving climate. The simulations initialized with extreme soil moisture conditions have shown the effects that the soil water content can have on determining the evolution of atmospheric fields such as PR or NSAT. These findings have some important implications in the context of dynamically downscaled DCPs. The moisture content in soil at the moment of initialization clearly influences on the time needed by soil moisture to reach a dynamical equilibrium and, consequently, the time needed by those atmospheric fields which directly interact with soil. While PR might reach this equilibrium relatively quickly (in less than 10 months in most part of the IP), NSAT fields usually need longer spin-up times over the whole domain under extreme soil moisture conditions, especially in the case of T_{max}. During this period, the RCM is not able to represent the evolution of these variables with a constant level of skill, since the bias fluctuates until it reachs the stationary state which determines the end of the spin-up period. This affects directly the simulations analyzed in Chapters 4 and 5, since no spin-up time was considered because it would have implied losing the first years of the hindcasts, making impossible assessing the predictive skill in these early years of the decade. In consequence, the predictive skill of the variables analyzed in those Chapters might be deteriorated to some extent by the presence of spin-up-related biases, at least during the first years of the simulations. Although PR may not experience significant improvements by starting the simulations from an already dynamically equilibrated soil state, it may potentially benefit the predictive skill for NSAT, especially in the cases of T_{mean} and T_{max}, for lead years 1, 2–5 and even 2-9, presumably. Given this situation, a dynamically equilibrated soil state, taken from a WRF simulation with ICs and LBCs provided by ERA-Interim, was been used to initialize the dynamically downscaled DCPs for the decade 2015–2025, presented in the following Chapter 7, in order to maximize as much as possible the predictive skill of the analyzed variables.

7

Decadal climate predictions for the period 2015–2025

This Chapter is devoted to examine the results obtained by the WRF-DPLE DCPs for PR, T_{max}, T_{min} and T_{mean} in the decade 2015–2025. As stated in Section 3.8, the whole 10-member CESM-DPLE subensemble available for DD simulations has been downscaled in this case, so the dependence of the results on the ensemble size has also been evaluated. Given the results obtained in Chapter 6, the initialization from a dynamically equilibrated soil state has been considered in the DD simulations for the decade 2015-2025 with the aim of enhancing as much as possible their predictive skill. More details about the soil initialization are available in Section 3.7.

To distinguish between the 4-member and 10-member WRF-DPLE ensembles, they will be referred to as WRF-DPLE$_4$ and WRF-DPLE$_{10}$, respectively, in the course of this Chapter. WRF-DPLE$_{10}$ has been recalibrated by following the same procedure previously applied for WRF-DPLE$_4$, which was described in Section 3.5.

7.1. Precipitation

7.1.1. *Analysis of the WRF-DPLE$_4$ predictions*

The spatial distributions of the multiannual mean anomalies of PR, half of the width of the confidence intervals at the 90 % level associated to a single WRF-DPLE$_4$ member ($\pm\Delta\mathrm{PR}_{90}$) and the relative anomaly error (E_{R}; Eq. [3.12]) at annual scale for WRF-DPLE$_4$ have been depicted in Figure 7.1. The anomalies have been computed by subtracting to the full fields the same lead time-dependent climatology calculated for the control period with the hindcasts analyzed in the previous chapters (see Eqs. [3.3] and [3.4]). The confidence intervals have been calculated also with those hindcasts by considering that the single members follow a Gaussian distribution

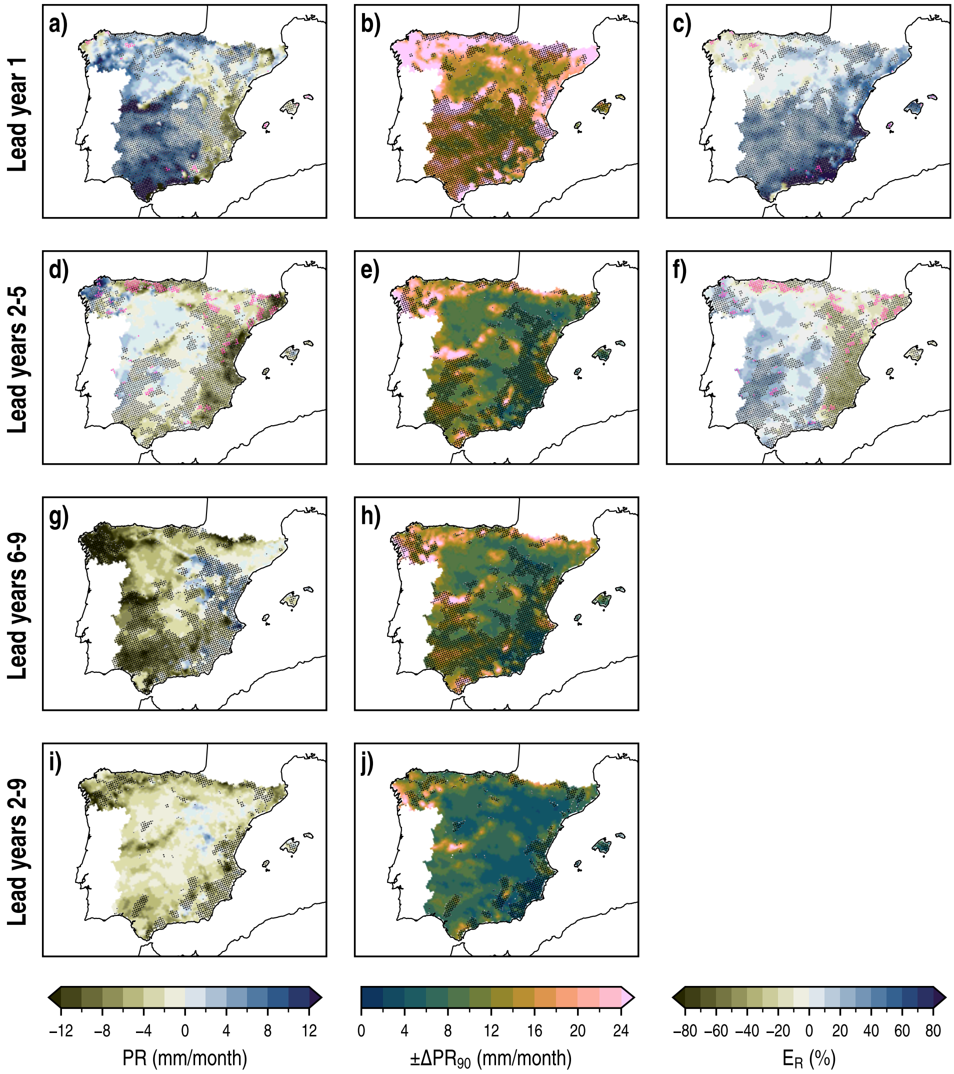

Figure 7.1: Spatial distributions of the WRF-DPLE$_4$ multiannual mean anomalies of PR (left column), half the width of the 90 % confidence intervals for a single WRF-DPLE$_4$ member (±ΔPR$_{90}$, center column) and the relative anomaly errors (E_R, right column), with AEMET as the observational dataset, at annual scale for several lead times (rows). The absence (presence) of black dots denote the locations where the forecast uncertainty is (not) represented by the confidence intervals. In PR and E_R maps, pink triangles identify the locations where the predictions are reliable but the confidence intervals do not contain the AEMET anomalies.

of mean equal to the ensemble mean and variance equal to the average ensemble variance $\overline{\sigma_Y^2}$ (Eq. [3.32]). The same considerations were assumed in the previous analyses of the reliability of the hindcasts. The spatial distributions of E_{R} have not been shown in lead years 6–9 and 2–9 because there are not available observations in the AEMET dataset from 2023 onwards.

As shown in Figure 4.7, the predictions are reliable mainly in the northern sector of the IP and part of the Mediterranean coast for lead year 1. At this lead time, the anomalies depicted in Figure 7.1a are commonly positive over regions in the northern sector, where values above 6 mm/month have been found in the central north and the northwest, with the latter showing maximum values above 12 mm/month in some locations. Some negative anomalies are shown in the northwestern coast, over the Central System, the Mediterranean coast, the northeast sector and some regions in the south of the IP. The strongest negative values are observed in the eastern flank, with anomalies below -6 mm/month, and in the southernmost regions close to the Strait of Gibraltar, with anomalies even below -12 mm/month. In general, higher positive anomalies than in northern areas have been found in the southern half of the IP, but the predictions are commonly not reliable in this part of the domain. There are also a few locations in the northwest and south whose AEMET anomalies fall outside the confidence intervals. At this lead time, the lowest relative errors in absolute value, which are below 20 %, have been found over the Northern Subplateau and surrounding regions (Figure 7.1c). Higher errors can be found in other locations with reliable predictions, such as the northeast of the IP, with values generally higher than 20 %, or the southernmost regions, where the E_{R} outcomes surpass 80 %. As the positive E_{R} values indicate, the anomalies are generally overestimated in lead year 1.

In lead years 2–5, the positive anomalies show smaller values than in lead year 1, in some cases turning into negative results, and the negative anomalies are slightly more intense (Figure 7.1d). In this case, the predictions are reliable over vast areas of the domain, excluding some regions in the eastern flank, in the southwestern sector and along the Atlantic and Cantabrian coasts. The highest reliable positive anomalies are still observed in the northwestern regions, with values commonly above 6 mm/month, whereas the most robust negative results are found in the Central System and northern regions, with minimum values mainly between -8 and -6 mm/month. At this lead time, more locations than in lead year 1 exhibit AEMET anomalies which fall outside the confidence intervals. These locations are generally placed in the north and northeast of the IP, although other, scattered in the eastern and southern sectors, are also observed. The results obtained for E_{R} generally show the same sign of their associated anomaly, indicating that the absolute value of the actual anomaly is commonly overestimated. In those regions where predictions are reliable, the lowest errors are usually found in central southern locations and in the north part of the domain. The northwest and part of the Central System show negative errors, whereas positive errors are frequently found elsewhere. The highest positive errors have been found in the southern part of the Northern Subplateau, with values up to 40 %.

The anomalies generally turn into negative values at the end of the decade. In lead years 6–9, most part of the regions where predictions are reliable show negative results, excepting some areas in the surrounding regions of the Iberian System and in the northeastern coast (Figure 7.1g). The most intense negative results are below -12 mm/month in the Central System, the northwest of the IP and to the south of the Pyrenees. On the other hand, the positive anomalies are between 4 mm/month and 12 mm/month close to the Iberian System, whereas they reach minimum values below 0.2 mm/month in the northeastern regions. A similar situation is observed in lead years 2–9, but with more moderate anomalies (Figure 7.1i). In this case, most part of the domain exhibits reliable results, with positive outcomes mainly covering the Iberian System and, on the contrary, negative values almost spanning the rest of the domain. The strongest reliable negative anomalies are again observed in the Central System, in locations close to the Pyrenees and in the northwestern regions, with values even below -12 mm/month for the latter. At all lead times, the confidence intervals defined by $\pm\Delta PR_{90}$ are much larger than the magnitude of the anomaly and, therefore, they usually encompass anomaly values with opposite signs. This fact is a consequence of the impact that the signal-to-noise paradox has on the DCPs for PR, which was previously discussed in Section 4.1.

The results for the WRF-DPLE$_4$ multiannual mean anomalies of PR at seasonal scale are depicted in Figure 7.2. The most intense anomalies for each region have generally been found in lead year 1. The highest absolute values are mainly observed in DJF at this lead time, although many regions lack of reliable predictions. The maximum values at this time are above 30 mm/month in the northern regions where the predictions are reliable. More moderate positive results are observed in the northeastern sector, whereas negative anomalies have been found in the southeast of the domain. The anomalies in DJF are predominantly positive in the first half of the decade, although they turn into negative for large areas of the domain in lead years 6–9 (Figure 7.2i). However, the area of negative anomalies found in the southern half of the IP at this lead time is mainly placed over regions where the predictions are not reliable. Reliable anomalies below -20 mm/month can be found in locations close to the Guadalquivir Valley and, with a more moderate magnitude, in the Northern

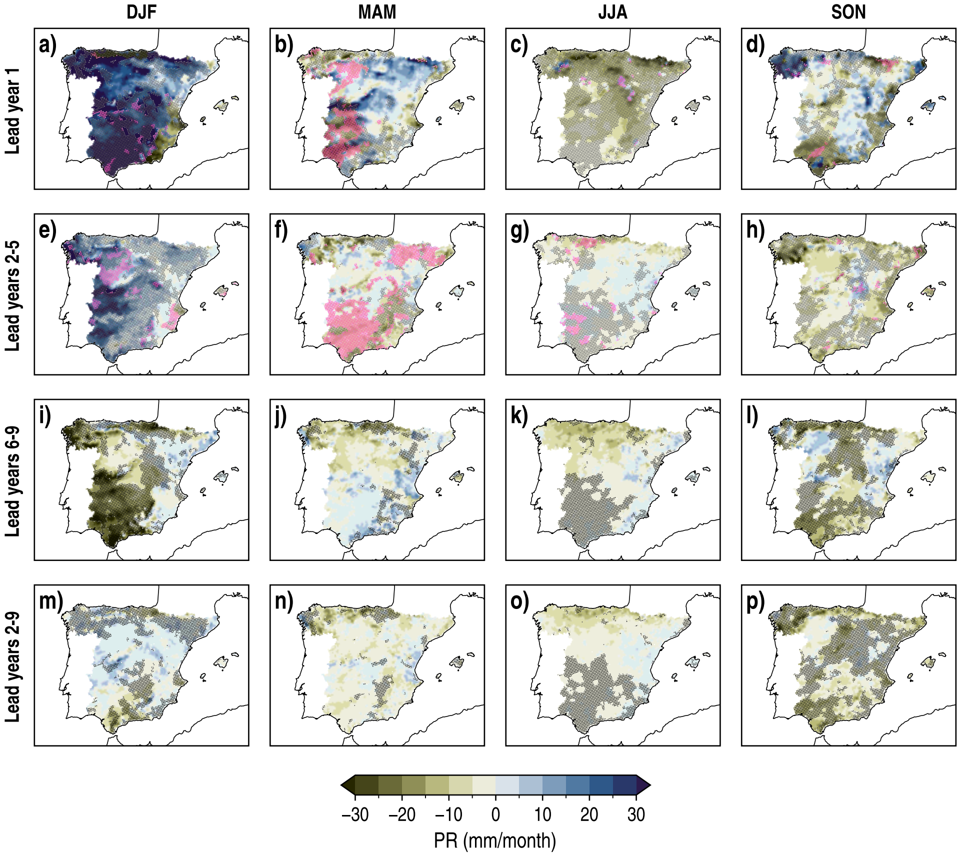

Figure 7.2: Spatial distributions of the WRF-DPLE$_4$ multiannual mean anomalies of PR for lead years 1, 2–5, 6–9 and 2–9 (rows) in DJF, MAM, JJA and SON (columns). The absence (presence) of black dots denote the locations where the forecast uncertainty is (not) represented by the confidence intervals. Pink triangles identify the locations where the predictions are reliable but the confidence intervals do not contain the AEMET anomalies.

Subplateau. On the other hand, they are mainly positive in the eastern part of the IP. For lead years 2–9 in DJF (Figure 7.2m), the anomalies are generally positive and below 10 mm/month in regions with reliable predictions. The season which shows the largest areas with reliable predictions is MAM. It shows positive anomalies which almost completely cover the domain in lead year 1 (Figure 7.2b). The positive results continue dominating the northern and southern halves of the domain in lead years 2–5 and 6–9, respectively. In lead years 2–9, the strongest anomalies are negative and below -15 mm/month in the northern regions (Figure 7.2n). On the other hand, the highest positive anomalies are slightly less intense and have been found along the

Mediterranean coast and some locations to the south of the Northern Subplateau, with values between 5 and 10 mm/month. For the other seasons, JJA and SON, the results obtained for the anomalies depend on the lead time. In locations with reliable predictions, the positive anomalies are more frequent in SON for lead year 1 and in JJA for lead years 2–5. At the decadal scale, the areas covered by reliable negative anomalies are larger than those covered by positive results in both seasons, but with more intense anomalies in SON.

The results obtained in MAM exhibit AEMET anomalies outside the confidence intervals over multiple regions at both lead years 1 and 2–5, as well as in DJF and JJA to a lesser degree. Although there is a probability of 10 % associated to the occurrence of these episodes, this high frequency suggests that there are certain aspects to consider about the calculation of these intervals. There are two factors which may be favouring these results. Firstly, note that soil was initialized in the DD simulations which generated the downscaled predictions for the decade 2015–2025. By contrast, the confidence intervals were calculated with downscaled hindcasts which did not consider this soil initialization. This fact may affect to some extent the computation of these intervals, especially in lead year 1, because of the biases related to the spin-up. However, note also that the spin-up period needed after the initialization with very extreme soil moisture conditions is generally lower than 10 months for PR (Figures 6.10 and 6.11). Under more normal initial conditions of soil moisture, the spin-up time needed by the dynamically downscaled PR is expected to be even shorter. Therefore, this factor might not totally explain these results by itself in lead year 1, and even less in lead years 2–5. The second factor which may potentially be causing these outcomes is the time gap between the end of the control period and the decade 2015–2025, along with the number of start dates in the former. This gap spans 15 years, being equivalent to half the number of start dates included in the control period. So, in order to consider the confidence intervals calculated in the control period to quantify the uncertainty of the predictions during the decade 2015–2025 in regions with reliable predictions, two requisites must be assumed. The first requisite is that the relationship between the average ensemble spread and the squared standard error (Eqs. [3.32] and [3.33], respectively) in this decade continues being the same as in the control period. The second requisite is that the recalibration coefficients used to correct the WRF-DPLE experiments, calculated in the control period, are also valid for this decade. The sampling biases could lead to a situation where these requisites may not be satisfied if the gap between the control period and the decade 2015-2025 is too large, or if the length of the control period relative to this gap is not long enough. Therefore, these confidence intervals might not be adequate to quantify the uncertainty for this decade, at least in this season and in the regions showing these results.

The spatial distributions of $\pm\Delta\mathrm{PR}_{90}$ and E_{R} at seasonal scale can be consulted in Figures B.58 and B.59, respectively, both available in Appendix B.3.1. Since the variability at seasonal scale is higher than at annual scale, the results obtained for $\pm\Delta\mathrm{PR}_{90}$ lead to wider confidence intervals. The results obtained for E_{R} depend on the season and lead time. The lowest errors are usually found in MAM and SON for lead years 2–5, as well as in DJF for some northwestern and southeastern regions in lead year 1 or along the Mediterranean coast in lead years 2–5. The regions in the northern half of the IP in lead years 2–5 and JJA also show errors with similar magnitudes. By contrast, the highest E_{R} values have been found in JJA over the whole domain for lead year 1 and mainly in the southern part for lead years 2–5, caused by the low PR rates which are commonly observed in this season (e.g., see Figure 4.4).

7.1.2. *Comparison with the WRF-DPLE$_{10}$ ensemble*

The results for the multiannual mean anomalies of PR from the WRF-DPLE$_{10}$ ensemble mean have been depicted in Figure 7.3. There is not any indication about the locations which have reliable predictions because it was only assessed for the WRF-DPLE$_4$ hindcasts. The spatial distributions are qualitatively similar to those obtained for WRF-DPLE$_4$ in Figure 7.1. In lead year 1, negative anomalies have been found in the eastern regions of the domain, whereas most part of the rest of the IP is covered by positive anomalies. While the maximum values above 12 mm/month have been found to the west of the Central System and in the northwestern regions, the minimum outcomes, below -12 mm/month, have been found to the south of the Pyrenees. At this lead time, WRF-DPLE$_{10}$ shows more moderate positive anomalies than WRF-DPLE$_4$ in the southern regions (Figure 7.1a). The area covered by negative anomalies is larger in lead years 2–5 (Figure 7.3b), when the positive results are found mainly to the north of the Ebro Valley, in the Northern Subplateau and, especially, in the northwest. In this case, the positive anomalies over the Northern Subplateau and Ebro Valley are slightly higher than those found for WRF-DPLE$_4$ (Figure 7.1d), turning from negative into positive for the latter. In lead years 6–9 (Figure 7.3c), the negative anomalies are much more frequent than those positive, which are constrained to regions surrounding the Iberian System and close to the Mediterranean coast. The strongest negative results, found in the northwest of the IP, with values generally below -6 mm/month and even -12 mm/month in some cases,

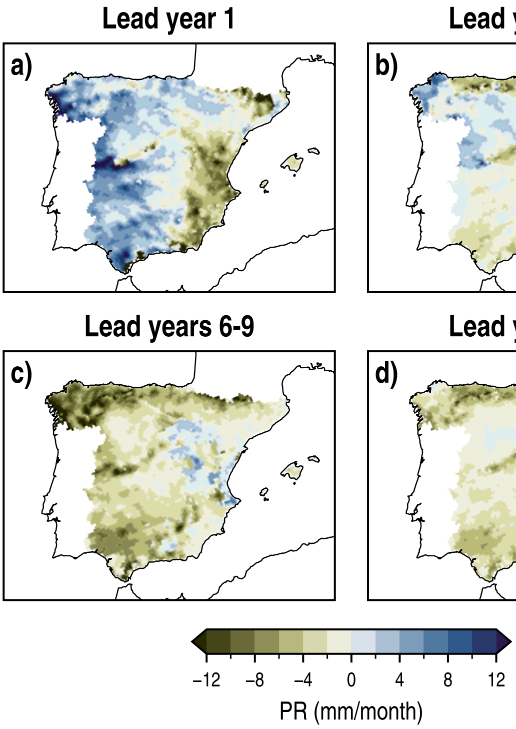

Figure 7.3: Spatial distributions of the WRF-DPLE$_{10}$ multiannual mean anomalies of PR in lead years 1, 2–5, 6–9 and 2–9 at annual scale.

are less accentuated than for WRF-DPLE$_4$ (Figure 7.1g). In lead years 2–9, the domain is almost fully covered with negative anomalies. The minimum anomalies have been found in the Cantabrian Range, with values generally below -6 mm/month. As in previous lead times, the most extreme values tend to be slightly more moderate than in the case of WRF-DPLE$_4$. See, for example, the northwestern regions, the Pyrenees or the Central System in Figures 7.1i and 7.3d. The results obtained at seasonal scale are available at Figure B.60 in Appendix B.3.1. These spatial distributions are also qualitatively similar to those obtained from WRF-DPLE$_4$ in Figure 7.2. However, there are still some differences. For example, the southeastern Mediterranean coast exhibits slightly more intense negative anomalies from WRF-DPLE$_{10}$ in lead year 1 in DJF, and higher positive anomalies to the north of the Iberian System have been found at the same time (Figure B.60a). Differences of similar magnitude can be found across seasons and lead times, although the most important differences are observed for lead year 1 in MAM, when the strong positive anomalies above 30 mm/month for

WRF-DPLE$_4$ (Figure 7.2b) are drastically reduced and even turn into negative for WRF-DPLE$_{10}$ (Figure B.60b).

The spatial distributions of the results obtained from the WRF-DPLE$_{10}$ ensemble mean in terms of E_R and MSSS$_4$ (MSSS for WRF-DPLE$_{10}$ with WRF-DPLE$_4$ as reference dataset; Eq. [3.16]) at annual scale for the decade starting in 2015 are shown in Figure 7.4. The spatial distributions of E_R for WRF-DPLE$_{10}$ and WRF-DPLE$_4$ (Figure 7.1) commonly share the same sign, although differences between their magnitudes lead to the differences observed in the results obtained for MSSS$_4$. WRF-DPLE$_{10}$ generally shows a better ability than WRF-DPLE$_4$ to predict the AEMET anomalies, especially in lead year 1. This improvement in predictive skill is observed mainly in regions over the southern half of the IP, as well as for some in the northwest, northeast and to the southwest of the Pyrenees. On the other hand, a higher skill for WRF-DPLE$_4$ is observed in locations close to the Central System and in some other regions scattered along the northern part of the IP. Similar performances between both ensembles have been obtained mainly along the Mediterranean coast and the Northern Subplateau. The MSSS$_4$ values are exactly 0 in the latter regions in particular

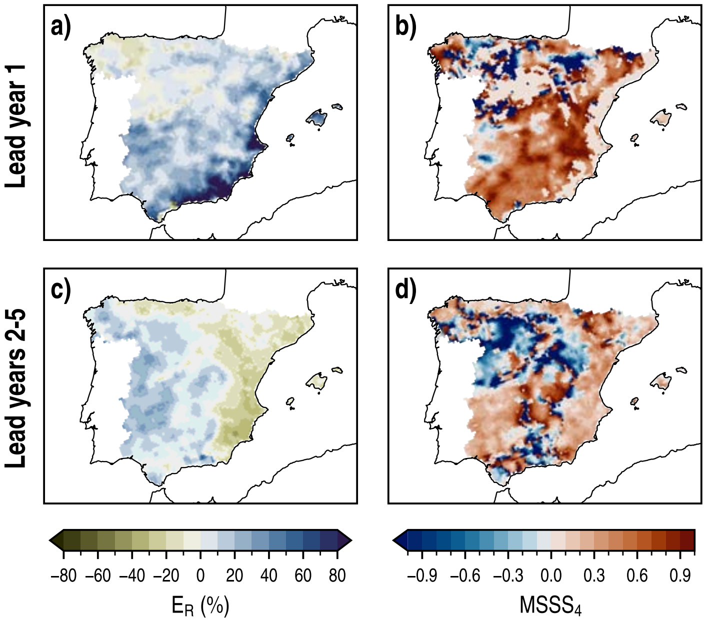

Figure 7.4 : Spatial distributions of relative anomaly errors (E_R, left column) for the WRF-DPLE$_{10}$ multiannual mean anomalies of PR, with AEMET as the observational dataset, and MSSS calculated with WRF-DPLE$_4$ as reference (MSSS$_4$, right column) in lead years 1 and 2–5 (rows) at annual scale.

because of the recalibration which have been applied to the downscaled DCPs and the filter which selects the locations that are or not subjected to such recalibration (see Section 3.5). Since the recalibration coefficients were calculated with the WRF-DPLE$_4$ hindcasts, the results obtained for WRF-DPLE$_{10}$ are equal to those obtained by WRF-DPLE$_4$ in those regions where the predictions have been recalibrated. In lead years 2–5, the area where WRF-DPLE$_{10}$ outperforms WRF-DPLE$_4$ is still large, but its extension decreases compared to lead year 1. In this case, the negative MSSS$_4$ scores showing better skill for WRF-DPLE$_4$ have been found mainly over the Northern Subplateau and regions to the east, the Central System, some locations over the Baetic System and other smaller regions scattered along the domain. On the other hand, WRF-DPLE$_{10}$ provides better estimations of the anomalies along the Cantabrian and Mediterranean coasts, the northeast of the IP and most part of the southern half of the domain. Note that MSSS$_4$ has been calculated from RMSE scores computed only for one start date. Therefore, these results should not be used to compare WRF-DPLE$_4$ and WRF-DPLE$_{10}$ performances in the evaluation of the impact of the ensemble size on the predictive skill for the DPS as a whole, since these results are influenced by a very large sampling bias. A more comprehensive set of experiments would be needed to address that task. Instead, they are only valid to compare WRF-DPLE$_4$ and WRF-DPLE$_{10}$ for this specific decade starting in 2015. Regardless of the results obtained here, the increase of the ensemble size is expected to provide more robust estimations of the actual climate and enhance the predictive skill of the DPS by reducing the unpredictable background noise and contributing to adequately capturing the climate signal, as discussed in previous chapters.

The results obtained for MSSS$_4$ at seasonal scale have been depicted in Figure 7.5. The spatial structure of this score depends on the season and the lead time. The areas covered by positive results are commonly larger than those which have negative scores, with the clear exceptions of MAM and JJA in lead year 1. In these cases, larger E_R values observed in Figures B.61b and B.61c for WRF-DPLE$_{10}$ than for WRF-DPLE$_4$ in Figures B.59b and B.59c (see Appendix B.3.1) lead to the generalized negative MSSS$_4$ scores observed in these seasons. At this lead time, WRF-DPLE$_{10}$ dominates in the southern half of the IP in DJF (with some exceptions mainly in the eastern flank), whereas WRF-DPLE$_4$ often obtains better results in the northern regions (Figure 7.5a). In SON, on the other hand, similar performances between both have been found over the eastern part of the domain (Figure 7.5d). By contrast, WRF-DPLE$_{10}$ is still better in northern and southern regions in this season, although WRF-DPLE$_4$ performs better mainly over some inner central areas. In lead years 2–5, WRF-DPLE$_{10}$ outperforms

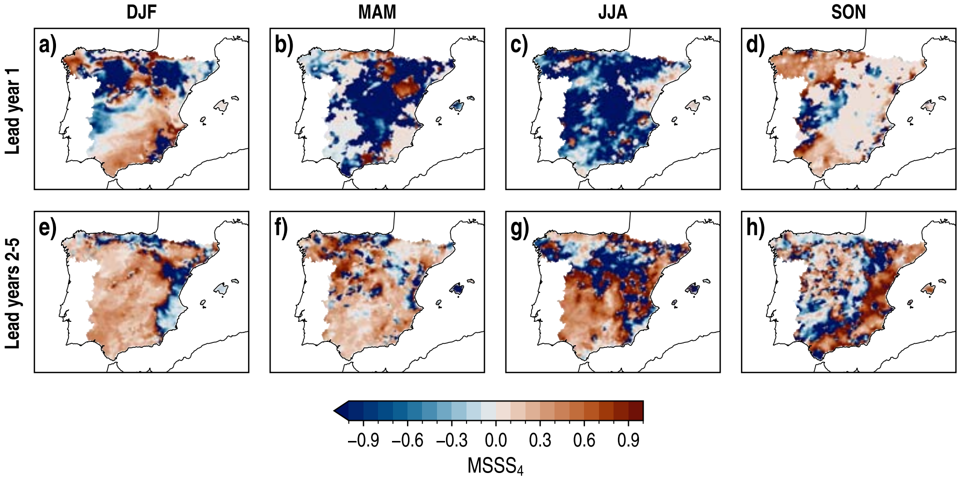

Figure 7.5: Spatial distributions of the MSSS for the WRF-DPLE$_{10}$ multiannual mean anomalies of PR, with WRF-DPLE$_4$ as reference (MSSS$_4$), for lead years 1 and 2–5 (rows) in DJF, MAM, JJA and SON (columns).

WRF-DPLE$_4$ over much large areas of the domain, not only in MAM and JJA but also in DJF and, to a lesser extent, in SON.

7.1.3. *Predictions for regional averages*

As done in previous chapters for the analysis of the hindcasts, the anomalies of the downscaled DCPs for the decade starting in 2015 have been spatially averaged over the regions obtained from the regionalization of AEMET PR described in Section 3.6 (see Figure 3.5a). This regionalization groups together those locations which have similar PR regimes, so these averages can be interpreted as a general representation of the anomaly of PR in each region. From this regional perpective, the results obtained at annual scale for the same metrics computed in sections above have been summarized in Table 7.1.

As happened in the analysis at grid-point scale, the second half of the decade shows generalized negative PR anomalies from WRF-DPLE$_4$ in all regions. In the first 5 years of the decade, however, some regions, such as the NW or WI regions, exhibit positive results, especially in lead year 1. Negative results for the WRF-DPLE$_4$ PR anomalies have also been found in lead years 2–9, even in that regions presenting positive anomalies in lead years 2–5. In lead years 2–9, the confidence intervals defined by ±ΔPR_{90} for WRF-DPLE$_4$ can be used to quantify the forecast uncertainty

Table 7.1: Skill scores for the spatially averaged WRF-DPLE multiannual mean anomalies of PR in lead years 1, 2-5, 6-9 and 2-9 for the decade starting in 2015 at annual scale. PR is the anomaly for the WRF-DPLE$_N$ (the N-member WRF-DPLE ensemble) ensemble mean, $\pm\Delta$PR$_{90}$ represents half the width of the 90 % conficence interval for a single WRF-DPLE member, E_R is the relative anomaly error, with AEMET as the observational dataset, and MSSS$_4$ denotes the added value of WRF-DPLE$_{10}$ over WRF-DPLE$_4$. Only for the WRF-DPLE$_4$ PR, the bold formatting denotes that WRF-DPLE$_4$ is able to represent the forecast uncertainty and that the 90 % confidence interval encloses the AEMET anomaly; the symbol "(∗)", if any, means that the former is satisfied but the latter is not; finally, the plain formatting is used when WRF-DPLE$_4$ cannot represent the forecast uncertainty. Dashes denote data unavailability at that lead time.

Region	Lead years	WRF-DPLE$_4$			WRF-DPLE$_{10}$		
		PR (mm/month)	$\pm\Delta$PR$_{90}$ (mm/month)	E_R (%)	PR (mm/month)	E_R (%)	MSSS$_4$
EI	1	**0.68**	11.63	11.48	-0.72	7.95	0.52
	2-5	**-1.56**	7.34	-5.84	-1.53	-5.77	0.02
	6-9	**-1.97**	7.22	–	-1.48	–	–
	2-9	**-1.76**	5.15	–	-1.50	–	–
WI	1	4.97	16.27	8.61	4.24	7.24	0.29
	2-5	**0.06**	11.60	12.60	0.37	13.24	-0.10
	6-9	-6.66	11.03	–	-3.97	–	–
	2-9	**-3.30**	8.01	–	-1.80	–	–
NE	1	-0.57	19.39	19.47	-3.47	14.04	0.48
	2-5	-5.39	9.03	-15.37	-4.26	-13.72	0.20
	6-9	**-2.71**	10.59	–	-2.68	–	–
	2-9	**-4.05**	6.34	–	-3.47	–	–
CS	1	5.03	12.87	36.03	1.11	25.20	0.51
	2-5	**-0.95**	9.27	8.05	-1.37	7.01	0.24
	6-9	-4.38	8.89	–	-3.46	–	–
	2-9	**-2.67**	6.12	–	-2.41	–	–
NW	1	**3.26**	27.94	-7.03	5.82	-5.04	0.49
	2-5	**2.35**	18.81	9.08	2.03	8.79	0.06
	6-9	-16.34	16.80	–	-9.66	–	–
	2-9	**-7.00**	13.10	–	-3.82	–	–
EA	1	-4.02	15.63	53.09	-5.85	44.90	0.28
	2-5	-6.77	7.55	-28.43	-5.04	-24.36	0.27
	6-9	-0.32	7.13	–	-0.88	–	–
	2-9	-3.55	4.90	–	-2.96	–	–
SW	1	8.01	17.61	32.15	2.82	21.31	0.56
	2-5	**-2.23**	11.77	7.32	-2.27	7.24	0.02
	6-9	-7.37	12.81	–	-5.78	–	–
	2-9	**-4.80**	8.32	–	-4.03	–	–
CN	1	3.72	22.11	2.39	1.61	0.51	0.95
	2-5	**-5.04** $^{(*)}$	13.24	-11.85	-4.13	-11.09	0.13
	6-9	**-5.66**	12.08	–	-6.42	–	–
	2-9	**-5.35**	8.73	–	-5.27	–	–

and generally contain the AEMET anomaly in lead years 2–9, with the exception of the EA region. There, the predictions do not show reliability at all. Likewise in Figure 7.1, the width of the confidence intervals is much larger than the anomalies. It means that, although the observed value can be found in them, the sign of the predicted anomaly may not be the same as for the AEMET value. The lowest E_R value has been found for lead year 1 in the CN region, although there are not reliable predictions in this case. Promising reliable results in terms of E_R have also been found, for example, for lead years 1 and 2–5 in the NW region, with absolute values below 10 % for both ensemble sizes. By contrast, the highest errors can be observed in the EA region. The results obtained in terms of the WRF-DPLE$_{10}$ PR anomalies in lead years 6–9 and 2–9 are generally more moderate than those obtained from WRF-DPLE$_4$, in the sense that the absolute values of the anomalies are usually lower for WRF-DPLE$_{10}$, with the exception of the CN and EA regions in lead years 6–9. In lead years 1 and 2–5, however, there is not such a clear pattern. In the case of WRF-DPLE$_{10}$, the results also indicate a progressive decrease of the anomalies into negative values along the decade. The lower magnitude of the WRF-DPLE$_{10}$ E_R outcomes compared to WRF-DPLE$_4$ are responsible of the general positive scores obtained for MSSS$_4$, indicating that WRF-DPLE$_{10}$ often outperforms WRF-DPLE$_4$ in estimating the AEMET anomaly at annual scale. The highest scores have been found in lead year 1 for all regions, with remarkable results especially for the CN region (MSSS$_4$ = 0.95). In lead years 2–5, the results are still positive for almost the whole domain (excepting the WI region), but the gain of increasing the ensemble size is normally much lower.

7.2. Daily maximum near-surface air temperature

7.2.1. *Analysis of the WRF-DPLE$_4$ predictions*

The spatial distributions of the WRF-DPLE$_4$ multiannual mean anomalies of T_max, half of the width of the confidence intervals at the 90 % level for a single member ($\pm\Delta T_{\mathrm{max},90}$) and the anomaly error in the decade starting in 2015 (E; see Eq. [3.9]) at annual scale have been depicted in Figure 7.6.

As opposed to the results obtained for PR in Section 7.1.1, the T_max anomalies are unanimously positive at all lead times for the whole domain. Such a result is consistent with the analysis of the NSAT trends done in the previous Chapter 5, which showed the existence of generalized significant positive trends for the T_max hindcasts in the IP at all lead times. In lead year 1, the regions where the predictions are reliable (see Figure 5.5 in Section 5.1.1) show anomalies ranging from 0.25 K to 0.75 K (Figure 7.6a).

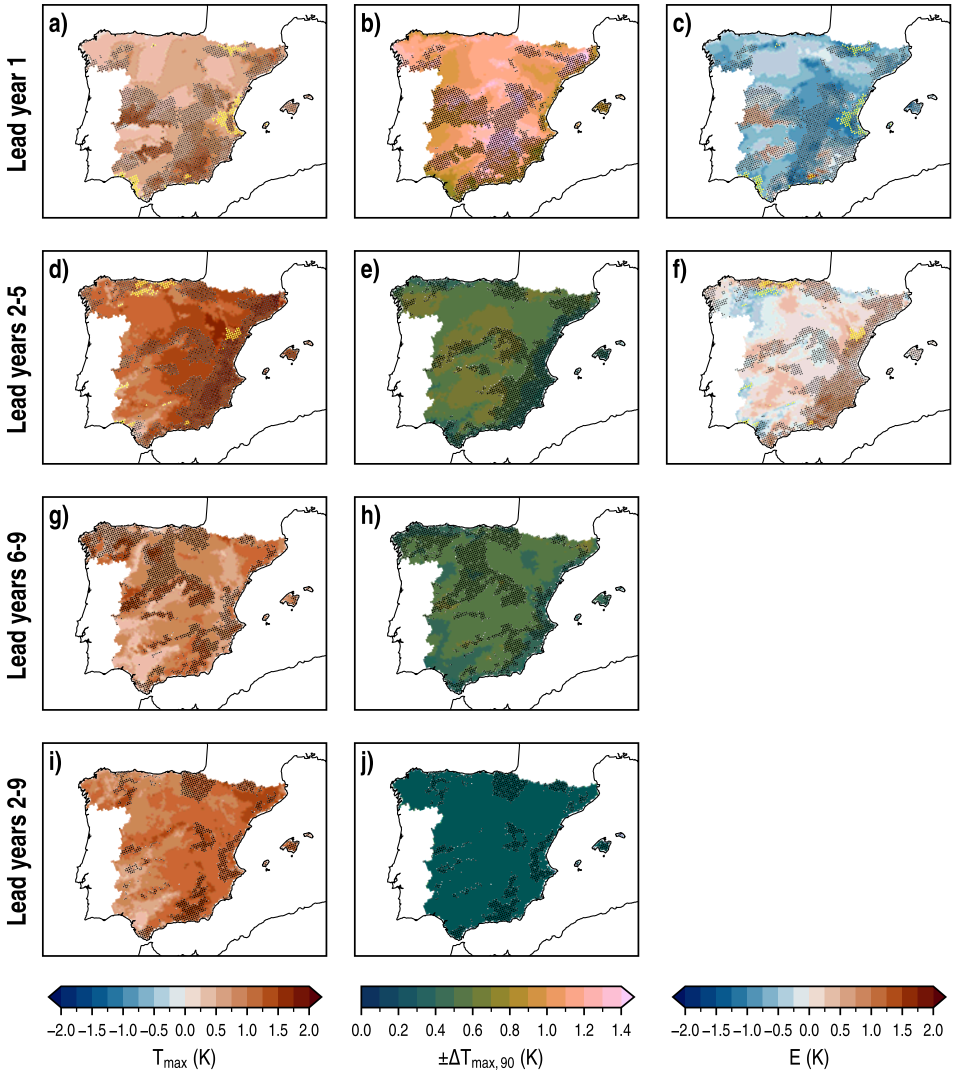

Figure 7.6: Spatial distributions of the WRF-DPLE$_4$ multiannual mean anomalies of T_{max} (left column), half the width of the 90 % confidence intervals for a single WRF-DPLE$_4$ member ($\pm\Delta T_{\text{max},90}$, center column) and the anomaly errors (E, right column), with AEMET as the observational dataset, at annual scale for several lead times (rows). The absence (presence) of black dots denote the locations where the forecast uncertainty is (not) represented by the confidence intervals. In T_{max} and E maps, yellow triangles identify the locations where the predictions are reliable but the confidence intervals do not contain the AEMET anomalies.

These regions are mainly in the northern part of the IP, to the south of the Central System, over the Guadalquivir Valley and along part of the Mediterranean coast.

This lead time exhibits the widest confidence intervals found for a single member in terms of T_{max} (Figure 7.6b). They are sufficiently large to enclose AEMET anomalies with opposite signs over vast areas. The AEMET anomalies are underestimated over almost all regions where the predictions are reliable (Figure 7.6c). The negative E values are even below -0.75 K in regions over the eastern half of the IP and some other in the northwest. Some locations with reliable predictions have shown AEMET anomalies which fall outside the confidence intervals. They are found to the south of the Pyrenees, along part of the Mediterranean coast and in some small southern regions.

The highest positive anomalies have been found in lead years 2–5 (Figure 7.6d). In this case, the reliable predictions can be found over most part of the domain (excepting the eastern flank, some central inner areas and northern regions), with values commonly between 0.75 K and 1.5 K. The maximum values at this lead time have been found to the south of the Ebro Valley, with anomalies up to 2 K. The confidence intervals defined by $\pm\Delta T_{\mathrm{max},90}$ are narrower than in lead year 1 (Figure 7.6e). Since $\pm\Delta T_{\mathrm{max},90}$ is below 0.7 K over the whole domain at this lead time, with minimum values between 0.4 K and 0.5 K in southern regions, a better precision than in lead year 1 is provided for the predictions of T_{max}. The spatial distribution of E (Figure 7.6f) depicts errors with lower magnitudes than in lead year 1, showing values commonly between -0.5 K and 0.5 K in regions with reliable outcomes. Most part of these results ranges from -0.25 K to 0.25 K. At this lead time, some AEMET anomalies fall outside the confidence intervals in places where there are reliable predictions. They can be mainly observed in northern and eastern regions, as well as in some locations scattered over the southern part of the IP.

More moderate anomalies are shown in lead years 6–9, when the minimum values are between 0.25 K and 0.5 K in regions with reliable predictions over the southwestern sector of the domain, as well as in some locations in the northwest (Figure 7.6g). The highest reliable anomalies, with values between 1.25 K and 1.5 K, are observed in central western and northwestern regions, as well as in a small area to the south of the Pyrenees. The results obtained for $\pm\Delta T_{\mathrm{max},90}$ show values below 0.6 K over almost the whole domain, providing slightly more precision than those in lead years 2–5 (Figure 7.6h). Finally, the anomalies in lead years 2–9 (Figure 7.6i) are in general higher than at the previous lead time, but slightly lower compared to lead years 2–5. The maximum reliable anomalies have been found in the northeastern regions, in the Central System, in the northwest and in some locations over the Southern Subplateau, with values between 1.25 K and 1.5 K. Large areas in the eastern half of the IP have values between 1 K and 1.25 K. The lowest $\pm\Delta T_{\mathrm{max},90}$ outcomes have been obtained at this lead time, with vast areas showing results even below 0.4 K (Figure 7.6j). The width of the confidence intervals relative to the magnitude of the anomalies is lower than for PR, because the signal-to-noise paradox is stronger in PR than in T_{max} (see Figures 4.5 and 5.3 for an evaluation of both RPCs in lead years 2–9).

The spatial distributions of the WRF-DPLE$_4$ multiannual mean anomalies of T_{max} at seasonal scale have been depicted in Figure 7.7. As at annual scale, the results found for the regions with reliable predictions are generally positive, as could be expected from the generalized positive T_{max} trends observed in Tables A.3 to A.6 (Appendix A.2). Some exceptions have been found mainly in lead year 1 for DJF and SON (Figures 7.7a and 7.7d, respectively), when negative anomalies can be observed across large areas of the domain, as well as in JJA in the Ebro Valley (Figure 7.7c). More negative anomalies have also been found in SON over some central and southwestern areas in lead years 6–9 (Figure 7.7l) and western regions in lead years 2–9 (Figure 7.7p), but these results are shown in locations without reliable predictions. The highest anomalies have been found in JJA, with values generally above 1.5 K at all lead times, excepting some regions close to the Mediterranean coast in the southeast for lead year 1. These regions show positive anomalies below 1 K in locations with reliable results. In MAM, high anomalies which increase eastwards from 1 K to 2 K are also observed in lead years 2–5 (Figure 7.7f), whereas they are commonly lower than 1 K in locations with reliable predictions at other lead times. The season which shows the largest areas of the domain with reliable results is DJF, when there is prediction reliability over most part of the domain at all lead times, excluding some regions mainly placed in the northern part and others also found over the Baetic System or in some inner and coastal locations. In this season, a similar situation to that depicted in Figure 7.2 for MAM and JJA is observed, in the sense that there are many locations with reliable predictions where the AEMET anomaly is outside the confidence intervals delimited by $\pm\Delta T_{\mathrm{max},90}$. Although this also happens in other seasons for T_{max} and there is a probability of 10 % associated to these occurrences, the results obtained in DJF, especially in lead year 1, suggest that the factors mentioned above in Section 7.1.1 may also be affecting these outcomes. In this case, the role of the spin-up biases may be more important than in the case of PR, since the spin-up period needed by T_{max} after the initialization with extreme soil moisture conditions can be longer than 3 years under the most unfavourable circumstances (see Figures 6.10 and 6.11), although it is expected to be shorter whether simulations are initialized with more normal soil moisture conditions.

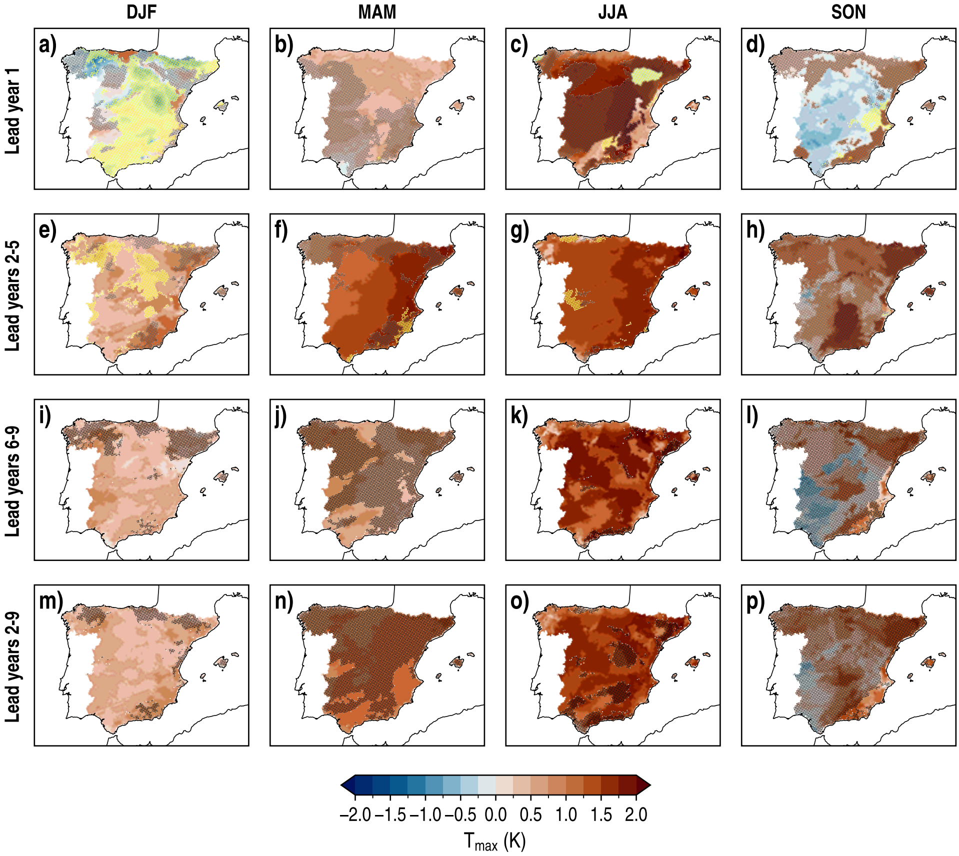

Figure 7.7: Spatial distributions of the WRF-DPLE$_4$ multiannual mean anomalies of T_{max} for lead years 1, 2–5, 6–9 and 2–9 (rows) in DJF, MAM, JJA and SON (columns). The absence (presence) of black dots denote the locations where the forecast uncertainty is (not) represented by the confidence intervals. Yellow triangles identify the locations where the predictions are reliable but the confidence intervals do not contain the AEMET anomalies.

The results obtained from the analysis at seasonal scale for $\pm\Delta T_{\text{max},90}$ and E can be consulted in Figures B.62 and B.63 (Appendix B.3.2), respectively. As occurred for PR, the seasonal variability is higher than at annual scale, leading to wider confidence intervals. Although JJA is the season showing the highest positive anomalies, it is not the season with the highest magnitudes for the E results. In JJA, the lowest errors have been found mainly in the eastern part of the IP for lead years 2–5, with values between -0.25 K and 0.25 K over many regions. Errors of similar magnitudes can also be found in lead year 1 for the same season, although they are constrained to smaller areas. Relatively moderate errors are also observed in lead years 2–5 for

MAM. The seasons which show the highest magnitudes of E are DJF and SON in lead year 1. The latter depicts values below -1.25 K over most part of the domain, where the predictions are generally reliable.

7.2.2. *Comparison with the WRF-DPLE$_{10}$ ensemble*

The multiannual mean anomalies of $T_{\max}$ obtained by WRF-DPLE$_{10}$ in the decade starting in 2015, available in Figure 7.8, are very similar to those obtained by WRF-DPLE$_4$ (Figure 7.6). In this case, the information about the prediction reliability is not provided because the CRPSS analyzed in Section 5.1.1 was calculated only for the 4-member ensemble. The results are also unanimously positive over the whole domain. The highest anomalies are observed for lead year 1 in the Baetic System and for lead years 2–5 in some locations in the northeast and southeast of the domain, showing values between 1.5 K and 1.75 K. Similar anomalies have been found between lead years 6–9 and 2–9, but with slightly higher results for the latter by nearly 0.25 K in general. Although the spatial patterns of the anomalies depicted for WRF-DPLE$_{10}$ and WRF-DPLE$_4$ show similar features, there are still some differences. For example, WRF-DPLE$_{10}$ shows lower anomalies in regions to the south of the Ebro Valley and along the Mediterranean coast for lead years 2–5. The anomalies for WRF-DPLE$_{10}$ are also slightly lower in lead years 6–9 in regions such as the southwestern part of the Northern Subplateau or in the northwest of the IP. The same occurs in the northwest for lead years 2–9 and also in some locations over the Central System.

The anomalies obtained for WRF-DPLE$_{10}$ at seasonal scale are available in Figure B.64 (Appendix B.3.2). In this case, the differences between WRF-DPLE$_{10}$ and WRF-DPLE$_4$ are more noticeable than at annual scale. Compare, for example, Figure B.64a and Figure 7.7a for DJF in lead year 1. While strong negative anomalies are observed for WRF-DPLE$_4$ to the north of the Southern Subplateau at this time, the same regions show positive anomalies for WRF-DPLE$_{10}$. Also in lead year 1 but for MAM, WRF-DPLE$_4$ (Figure 7.7b) shows generalized lower anomalies than WRF-DPLE$_{10}$ (Figure B.64b) over the whole domain. This situation is inverted in JJA at the same lead time (Figures 7.7c and B.64c). Differences continue at other lead times depending on the season. Some of the most accentuated differences can be observed in MAM and JJA in lead years 2–5 and 6–9, respectively, when the anomalies are generally higher for WRF-DPLE$_4$ (Figures 7.7f and 7.7k) than for WRF-DPLE$_{10}$ (Figures B.64f and B.64k).

The spatial distributions of E and MSSS$_4$ for WRF-DPLE$_{10}$ in lead years 1 and

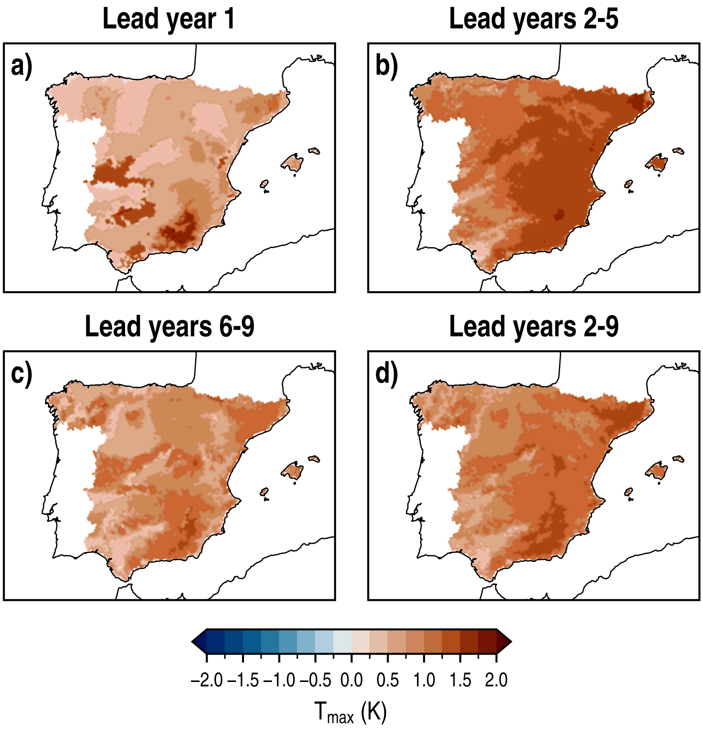

Figure 7.8: Spatial distributions of the WRF-DPLE$_{10}$ multiannual mean anomalies of T_{max} in lead years 1, 2–5, 6–9 and 2–9 at annual scale.

2–5 at annual scale have been represented in Figure 7.9. The results obtained for E generally show similar spatial patterns to those observed in Figures 7.6c and 7.6f for WRF-DPLE$_4$, but there are some differences which lead to the MSSS$_4$ results. In lead year 1, the performances of both ensembles are identical in regions where the recalibration has been applied, as explained in Section 7.1.2. On the other hand, there are regions to the south of the Northern Subplateau and Sierra Morena where WRF-DPLE$_4$ commonly gets better results, and other southern locations such as the Baetic System where WRF-DPLE$_{10}$ shows both higher and lower errors than WRF-DPLE$_4$. In lead years 2–5, there are more discrepancies between the performances of the ensembles. The eastern flank and the southern half of the domain mostly show a better performance of WRF-DPLE$_{10}$, whereas better results for WRF-DPLE$_4$ are observed in part of the northwest of the domain and the Northern Subplateau, the south of the Pyrenees, some central inner regions and other locations in the south. As mentioned above in the case of PR, it is worth remarking that these results obtained

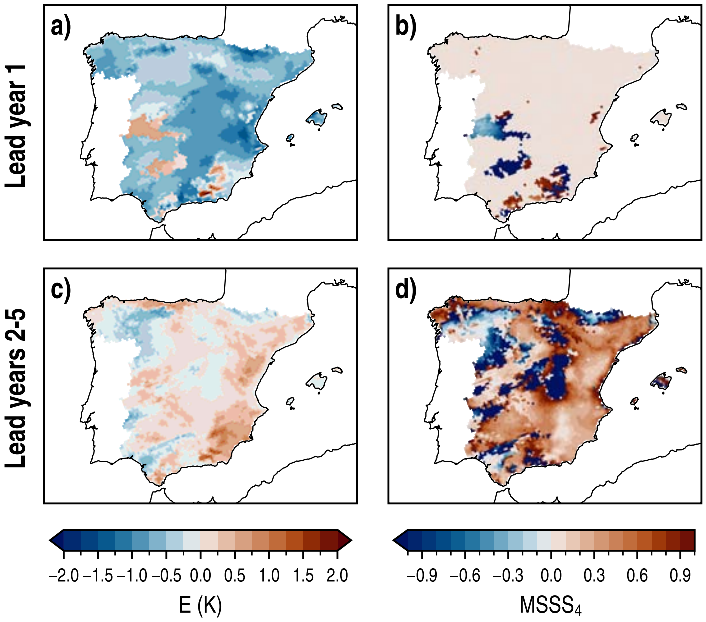

Figure 7.9: Spatial distributions of anomaly errors (E, left column) for the WRF-DPLE$_{10}$ multiannual mean anomalies of $T_{\max}$, with AEMET as the observational dataset, and MSSS calculated with WRF-DPLE$_4$ as reference (MSSS$_4$, right column) in lead years 1 and 2–5 (rows) at annual scale.

for MSSS$_4$ should not be used as a measure of the added value to the predictive skill of increasing the ensemble size of WRF-DPLE in general. These results only show the differences between the predictions obtained by both downscaled ensembles for the decade starting in 2015. As discussed in Chapter 5, the increase of the ensemble size of a DPS provides more robust estimations of the predictive skill because of the need of computing large ensemble averages to remove the unpredictable background noise in DCPs.

At seasonal scale, the highest MSSS$_4$ values have been found in DJF for both lead years 1 and 2–5 (Figure 7.10). In this season, positive MSSS$_4$ outcomes have been obtained for most part of the domain, excluding the Ebro Valley, the southeast, some of the southernmost locations and the Cantabrian coast in lead years 2–5. On the other hand, WRF-DPLE$_4$ has obtained better results for lead year 1 in MAM. In lead years 2–5, by contrast, the regions with better results for WRF-DPLE$_4$ are fundamentally constrained to the northern half of the domain, excluding the northeastern sector. Some of the best results for WRF-DPLE$_4$ in comparison with WRF-DPLE$_{10}$ have been obtained in JJA, with large areas of the domain indicating that WRF-DPLE$_4$

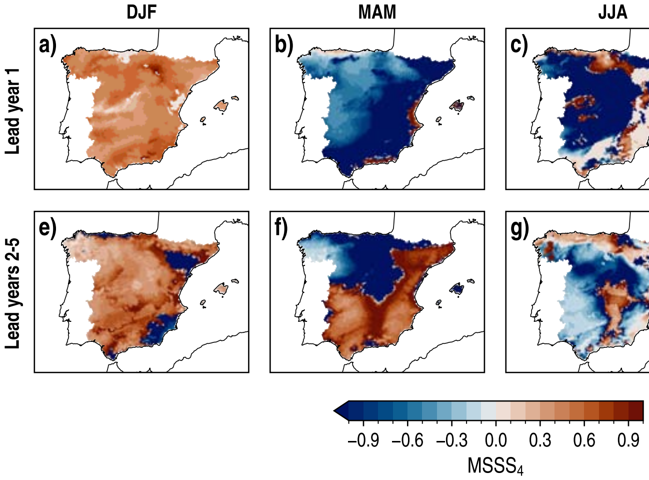

Figure 7.10: Spatial distributions of MSSS for the WRF-DPLE$_{10}$ multiannual mean anomalies of T_{max}, with WRF-DPLE$_4$ as reference (MSSS$_4$), for lead years 1 and 2–5 (rows) in DJF, MAM, JJA and SON (columns).

better captures the magnitude of the AEMET anomaly. Finally, similar performances between ensembles have been found for lead year 1 in SON over almost the whole domain, whereas WRF-DPLE$_{10}$ generally improves the results obtained by WRF-DPLE$_4$ in most part of the eastern half in lead years 2–5.

7.2.3. *Predictions for regional averages*

The multiannual mean anomalies of T_{max} have been spatially averaged over the regions resulting from the regionalization applied to the AEMET NSAT, which was described in Section 3.6 (Figure 3.5b). The results obtained at annual scale for the same metrics computed in the previous sections have been summarized in Table 7.2.

The results obtained for the WRF-DPLE$_4$ T_{max} are in the line of those observed at grid-point scale (Figure 7.6), in the sense that the anomalies are completely positive at all lead times for all regions. The highest averaged anomalies for each region have been found in lead years 2–5 and 2–9, surpassing 1 K at both lead times in almost all regions, with the exception of the SW and NO regions. The AEMET anomalies are contained in the confidence intervals defined by $\pm\Delta T_{\mathrm{max},90}$ for WRF-DPLE$_4$ in all regions and lead times where the predictions are reliable. In contrast to PR (Table 7.2), the confidence intervals for T_{max} are narrow enough not to contain values of different signs at almost all lead times, with the exception of lead year 1. The dependence of $\pm\Delta T_{\mathrm{max},90}$ on the length of the averaging lead time window is clear, as it decreases

Table 7.2: Skill scores for the spatially averaged WRF-DPLE multiannual mean anomalies of T_{max} in lead years 1, 2-5, 6-9 and 2-9 for the decade starting in 2015 at annual scale. T_{max} is the anomaly for the WRF-DPLE$_{\mathrm{N}}$ (the N-member WRF-DPLE ensemble) ensemble mean, $\pm\Delta T_{\mathrm{max,90}}$ represents half the width of the 90 % conficence interval for a single WRF-DPLE member, E is the anomaly error, with AEMET as the observational dataset, and MSSS$_4$ denotes the added value of WRF-DPLE$_{10}$ over WRF-DPLE$_4$. Only for the WRF-DPLE$_4$ T_{max}, the bold formatting denotes that WRF-DPLE$_4$ is able to represent the forecast uncertainty and that the 90 % confidence interval encloses the AEMET anomaly; the symbol "(∗)", if any, means that the former is satisfied but the latter is not; finally, the plain formatting is used when WRF-DPLE$_4$ cannot represent the forecast uncertainty. Dashes denote data unavailability at that lead time.

Region	Lead years	WRF-DPLE$_4$			WRF-DPLE$_{10}$		
		T_{max} (K)	$\pm\Delta T_{\mathrm{max,90}}$ (K)	E (K)	T_{max} (K)	E (K)	MSSS$_4$
SW	1	0.63	0.95	-0.46	0.66	-0.42	0.16
	2-5	**0.90**	0.54	0.03	0.85	-0.02	0.70
	6-9	**0.62**	0.48	–	0.63	–	–
	2-9	**0.76**	0.36	–	0.74	–	–
NO	1	**0.45**	1.07	-0.56	0.45	-0.56	0.00
	2-5	1.07	0.52	0.13	0.98	0.04	0.89
	6-9	0.76	0.46	–	0.72	–	–
	2-9	**0.91**	0.34	–	0.85	–	–
CI	1	0.67	1.17	-0.73	0.69	-0.71	0.06
	2-5	**1.30**	0.59	0.15	1.26	0.10	0.52
	6-9	**0.89**	0.54	–	0.91	–	–
	2-9	**1.10**	0.39	–	1.08	–	–
NE	1	**0.65**	1.07	-0.60	0.66	-0.60	0.00
	2-5	**1.38**	0.51	0.21	1.31	0.14	0.57
	6-9	**0.91**	0.51	–	0.94	–	–
	2-9	**1.15**	0.35	–	1.12	–	–
CS	1	1.04	1.03	-0.51	1.23	-0.31	0.62
	2-5	1.43	0.51	0.44	1.36	0.37	0.29
	6-9	**1.03**	0.51	–	1.15	–	–
	2-9	1.23	0.34	–	1.26	–	–
EA	1	**0.70**	1.00	-0.77	0.71	-0.76	0.02
	2-5	1.48	0.43	0.46	1.39	0.37	0.34
	6-9	**0.80**	0.45	–	0.85	–	–
	2-9	**1.14**	0.30	–	1.12	–	–
MT	1	**0.66**	1.10	-0.68	0.68	-0.66	0.07
	2-5	**1.24**	0.57	-0.10	1.19	-0.15	-1.36
	6-9	**1.05**	0.52	–	0.99	–	–
	2-9	**1.15**	0.38	–	1.09	–	–
WI	1	**0.56**	1.05	-0.39	0.57	-0.39	0.02
	2-5	**1.14**	0.59	0.01	1.11	-0.03	-16.20
	6-9	**0.87**	0.51	–	0.78	–	–
	2-9	**1.00**	0.38	–	0.94	–	–

from lead year 1 down to reach the minimum values in lead years 2–9. In lead year 1, all WRF-DPLE$_4$ E values show an underestimation of the T_{max} anomalies by the downscaled predictions, whereas the errors in lead years 2–5 are usually related to an overestimation of the AEMET anomaly, excepting in the MT region. The lowest errors for WRF-DPLE$_4$ are shown always for lead years 2–5 in all regions. The best results obtained in these terms have been found in the WI region, which shows the lowest magnitudes of E in both lead years 1 and 2–5. As mentioned in previous sections, the results obtained for WRF-DPLE$_{10}$ are close to those for WRF-DPLE$_4$. However, there are still some differences, as the results obtained for E and MSSS$_4$ reveal. With respect to the former, it is worth remarking that WRF-DPLE$_{10}$ generally gets equal or smaller absolute errors than WRF-DPLE$_4$ at almost all lead times in all regions. The only exceptions, with small differences, are the WI and MT regions in lead years 2–5. These improvements over WRF-DPLE$_4$ in predicting the AEMET anomaly are reflected in the results obtained for MSSS$_4$, with the only negative results shown in the aforementioned cases. The highest scores are observed in lead years 2–5 in the SW and NO regions, with values of 0.70 and 0.89, respectively. However, in the case of the SW region, the differences between the WRF-DPLE$_4$ and WRF-DPLE$_{10}$ leading to this result are very small, since E values about 0.03 K and -0.02 K have been respectively obtained for each ensemble. In the same line, the most negative MSSS$_4$ result, MSSS$_4$ $= -16.20$ in the MT region for lead years 2–5, is also caused by differences of the order of 10^{-2} K in the E results obtained by both ensembles.

7.3. Daily minimum near-surface air temperature

7.3.1. *Analysis of the WRF-DPLE$_4$ predictions*

The WRF-DPLE$_4$ multiannual mean anomalies of T_{min} at annual scale for the decade starting in 2015, depicted in Figure 7.11, only show positive values over the whole domain at all lead times, as in the case of T_{max} (Figure 7.6). These predictions are not reliable over large parts of the domain because of the results obtained in Figure 5.16, which were discussed in Section 5.2.1.

The reliability is present in lead year 1 mainly in the northern part of the IP and in the Balearic Islands, as well as along part of the Mediterranean coast and some southern and southwestern regions. In these places, the T_{min} anomalies are usually between 0.4 K and 0.8 K (Figure 7.11a). The highest values can be observed in the northeast and west of the IP, where the outcomes are above 0.6 K. The confidence intervals for a single member defined by $\pm\Delta T_{\text{min},90}$ (Figure 7.11b) show narrower

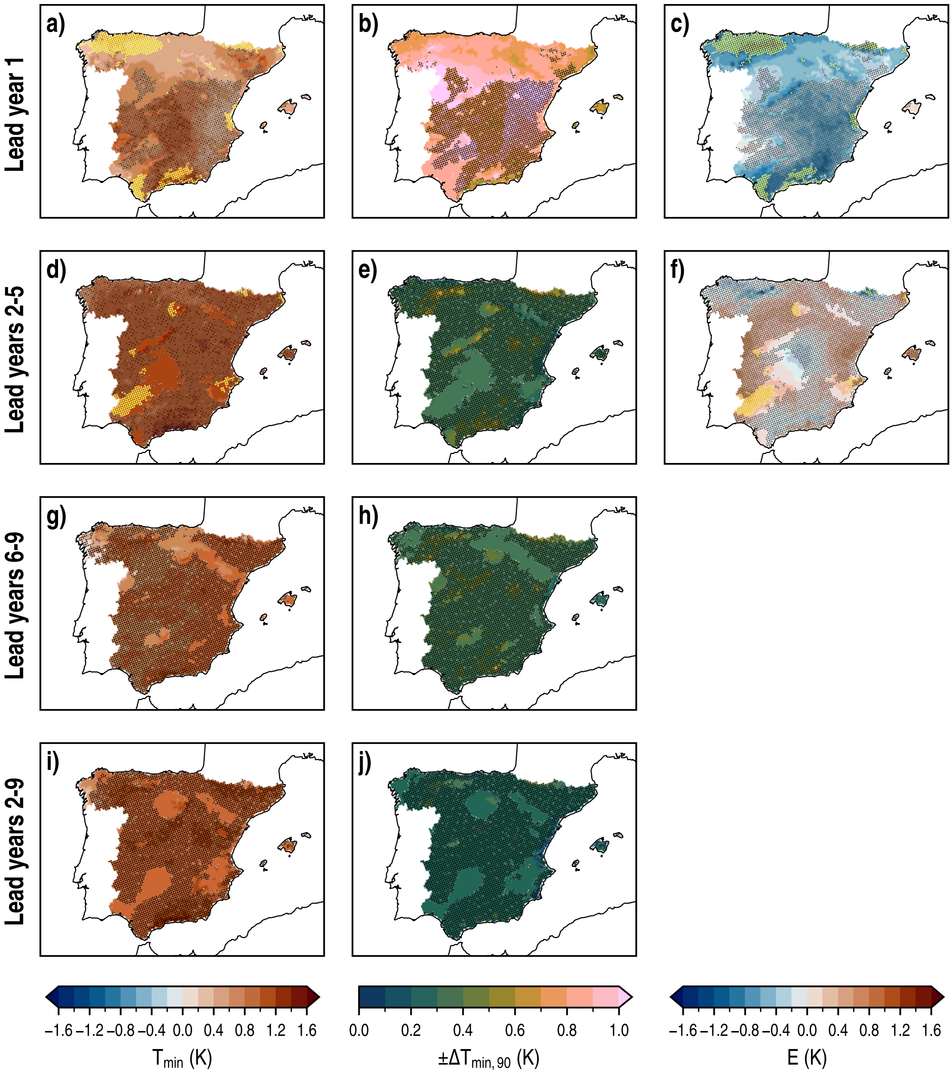

Figure 7.11: Spatial distributions of the WRF-DPLE$_4$ multiannual mean anomalies of T_{min} (left column), half the width of the 90 % confidence interval for a single WRF-DPLE$_4$ member ($\pm\Delta T_{\text{min,90}}$, center column) and the anomaly errors (E, right column), with AEMET as the observational dataset, at annual scale for several lead times (rows). The absence (presence) of black dots denote the locations where the forecast uncertainty is (not) represented by the confidence intervals. In T_{min} and E maps, yellow triangles identify the locations where the predictions are reliable but the confidence intervals do not contain the AEMET anomalies.

widths than those defined by $\pm\Delta T_{\text{max,90}}$ in Figure 7.6b, but they are still enclosing anomalies of different signs. The highest $\pm\Delta T_{\text{min,90}}$ results are observed over the

Northern Subplateau, with values generally above 1 K. On the other hand, the lowest $\pm\Delta T_{\text{min},90}$ values are observed along the northeastern and southeastern regions close to the Mediterranean coast, with values between 0.5 K and 0.7 K. As happened for T_{max} in Figure 7.6c, the AEMET anomalies are underestimated over almost the whole domain at this lead time (Figure 7.11c), with the exception of the Balearic Islands and some southwestern locations (but mostly without reliable results in the latter). The smallest absolute errors are observed in the northeast and southwest of the IP (the regions showing the highest anomalies) and over the Balearic Islands and part or the Northern Subplateau, with absolute E values commonly below 0.4 K. There are some regions where the AEMET anomalies are not inside the confidence intervals defined by $\pm\Delta T_{\text{min},90}$ in places where there are reliable predictions. These results have been found in the northwest, the northeast, a small area along the Mediterranean coast and other southern regions. In all cases, the AEMET anomaly is higher than the upper boundary of the confidence intervals. Although there is a probability of 10 % associated to these occurrences, the same factors mentioned in previous sections (i.e., the soil spin-up and the length of the control period alongside the gap between the end of this period and the decade 2015–2025) may be contributing to this situation. In this case, since the length of the spin-up period after the initialization with extreme soil moisture conditions is commonly below 1 year (Figures 6.10 and 6.11), the impact of this factor on these results would be lower than in the case of T_{max}.

Likewise for T_{max}, the highest anomalies of T_{min} have been found in lead years 2–5 (Figure 7.11d). The regions with reliable predictions are mainly located in the southern half of the IP, as well as over some other regions such as the Ebro Valley, the Central System or the northernmost part of the Iberian System. The anomalies in those regions are between 1 and 1.2 K. The results obtained for $\pm\Delta T_{\text{min},90}$ at this lead time (Figure 7.11e) are much lower than in lead year 1. With the exception of the Central and Iberian Systems, where values around 0.7 K can be found, results below 0.4 K are generalized in regions with reliable predictions. In the same regions, the results obtained for the spatial distribution of E (Figure 7.11f) are commonly positive, showing a general overestimation of the AEMET anomalies. Some regions with negative E values can also be observed in the central part of the IP and in some northern locations. At this lead time, the AEMET anomalies which fall outside the confidence intervals in regions with reliable predictions are commonly smaller than the lower boundaries of the confidence intervals, with the exception of some locations in the Pyrenees.

In lead years 6–9, the T_{min} anomalies, depicted in Figure 7.11g, are lower than those observed in lead years 2–5. They are mainly between 0.4 K and 0.8 K in regions with reliable predictions, with the highest values above 0.6 K found in the eastern part of the domain. The confidence intervals defined by $\pm\Delta T_{\mathrm{min},90}$ in Figure 7.11h show very similar results to those obtained in lead years 2–5.

The T_{min} anomalies generally increase at decadal scale compared to lead years 6–9, since values above 0.8 K have been found over most part of regions showing reliable results (Figure 7.11i). The lowest outcomes can be observed in the northwestern regions, with values between 0.2 K and 0.6 K. The narrowest confidence intervals are obtained in lead years 2–9 (Figure 7.11j), when values between 0.2 K and 0.3 K are found in most part of the domain, with some regions in the north having values between 0.3 K and 0.4 K.

With a few exceptions, the WRF-DPLE$_4$ T_{min} anomalies resulting from the analysis at seasonal scale are predominantly positive in regions with reliable predictions (Figure 7.12). The highest anomalies are commonly observed in JJA (Figure 7.12, third column), but generally with lower values than those obtained for T_{max} in Figure 7.7. This season has some of the largest areas where predictions are reliable at all lead times. The most accentuated positive anomalies are shown in the eastern part of the IP, where they reach maximum values above 1.5 K in lead years 6–9. The largest area with reliable negative anomalies has been found for lead year 1 in JJA, with values down to -0.5 K. More moderate but still high positive anomalies are also observed in SON, especially in lead years 6–9 and 2–9. At these lead times, predictions are reliable mainly in the western half of the IP, being these areas generally constrained to the northwestern sector in lead years 6–9. These anomalies are higher in lead years 2–9, with values predominantly between and 1 K and 1.5 K. The most accentuated anomalies in MAM have been found for lead years 2–5 over the eastern flank of the IP, where the results show values between 1 K and 1.25 K. In the same season, anomalies between 0.75 K and 1 K are observed in the southwestern and eastern sectors in lead years 1 and 2–9 for regions with reliable predictions. The highest anomalies in DJF have been found in lead years 1 and 6–9, with maximum values normally ranging from 1 K to 1.5 K.

The amount of cases with reliable predictions but also with AEMET anomalies outside the confidence intervals (Figure B.66 in Appendix B.3.3) is lower than for T_{max}. For T_{min}, these cases are mainly observed in JJA, and they are predominantly associated to underestimations of the AEMET anomalies, not only in JJA but in all seasons, as the spatial distributions of the seasonal E indicate (Figure B.67 in

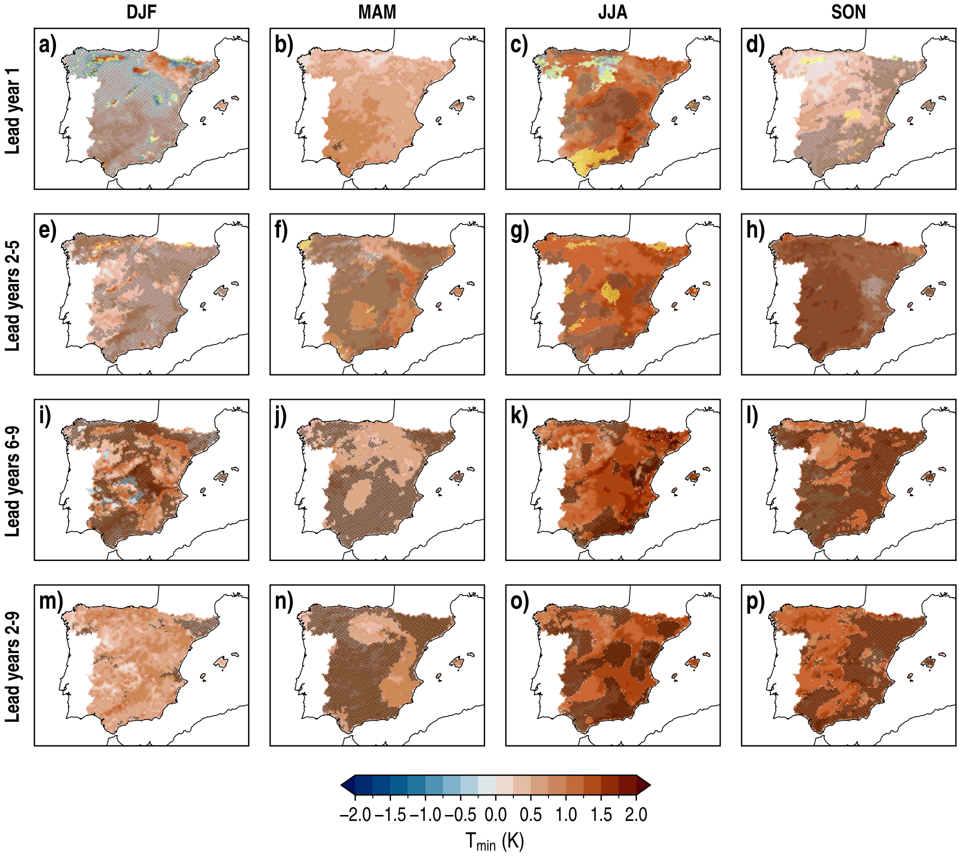

Figure 7.12: Spatial distributions of the WRF-DPLE$_4$ multiannual mean anomalies of T_{min} for lead years 1, 2–5, 6–9 and 2–9 (rows) in DJF, MAM, JJA and SON (columns). The absence (presence) of black dots denote the locations where the forecast uncertainty is (not) represented by the confidence intervals. Yellow triangles identify the locations where the predictions are reliable but the confidence intervals do not contain the AEMET anomalies.

Appendix B.3.3). There is a few exceptions in JJA in lead years 2–5 for some regions in the west of the domain, where these results are due to an overestimation of the AEMET anomalies. The results obtained for E show the highest errors for lead year 1 in DJF, when values below -1.6 K are found over most part of the domain. On the other hand, the largest areas with the lowest absolute errors are observed in JJA for lead years 2–5, showing values between -0.2 K and 0.2 K over the eastern and northern sectors of the IP. In general, the results obtained for $\pm\Delta T_{\mathrm{min},90}$ relative to the magnitude of the anomalies, at both annual and seasonal scales, are sligthly lower than those obtained for $\pm\Delta T_{\mathrm{max},90}$. The weaker signal-to-noise paradox observed in

T_{min} compared to T_{max} in lead years 2–9 contributes to producing these outcomes (see Figures 5.3 and 5.14).

7.3.2. *Comparison with the WRF-DPLE$_{10}$ ensemble*

The T_{min} anomalies obtained from WRF-DPLE$_{10}$ at annual scale are also positive for the whole domain (Figure 7.13), as for the 4-member ensemble. In the same vein as in the analysis of the WRF-DPLE$_{10}$ results for the other variables, there is not any indication related to the confidence intervals since they were calculated only with the WRF-DPLE$_4$ hindcasts. In lead year 1, the highest anomalies have been found in central inner regions and over the Baetic System, with values ranging from 1 K to 1.4 K. In lead years 2–5, most part of the domain depicts outcomes above 0.8 K, with results between 1 K and 1.2 K covering large areas of the domain, and maximum values which surpass 1.6 K to the south of the Baetic System. The anomalies are

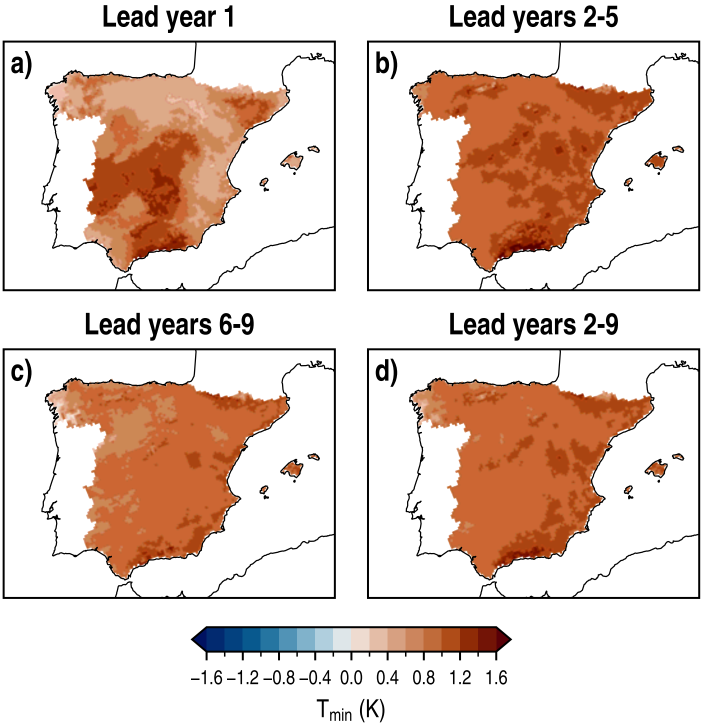

Figure 7.13: Spatial distributions of the WRF-DPLE$_{10}$ multiannual mean anomalies of T_{min} in lead years 1, 2–5, 6–9 and 2–9 at annual scale.

slightly lower in lead years 6–9 compared to the previous lead time, and most part of the regions with values above 1 K have disappeared, although some of them are still present along the Mediterranean coast and the Pyrenees. The results obtained in lead years 2–9 represent a midpoint between those for lead years 2–5 and 6–9, when the maximum values in the Baetic System emerge again along with some regions showing anomalies above 1 K, mainly found in the eastern part of the domain. The main differences with the results obtained for WRF-DPLE$_4$, showed in Figure 7.11, are observed in lead years 2–5, with anomalies generally lower for WRF-DPLE$_{10}$, and in lead year 1, when they are higher over the central inner part and southern regions of the IP. At seasonal scale, the anomalies have been represented in Figure B.68 (Appendix B.3.3). Most part of these results show very similar spatial distributions to those obtained for WRF-DPLE$_4$ (Figure 7.12), but there are differences whose magnitude depends on the lead time and season. Some of the more remarkable differences are those observed in DJF for lead years 1 and 6–9, when the negative anomalies observed for WRF-DPLE$_4$ turn into positive for WRF-DPLE$_{10}$. Moreover, the anomalies for WRF-DPLE$_{10}$ are generally slightly lower in lead years 1, 6–9 and 2–9 in JJA, as well as in lead years 2–9 in SON.

The spatial distributions of E and MSSS$_4$ for WRF-DPLE$_{10}$ in lead years 1 and 2–5 at annual scale have been depicted in Figure 7.14. In lead year 1, WRF-DPLE$_{10}$ gets more moderate negative E results in some locations in the northeast, part of the Northern Subplateau and other central inner and southern regions, indicating that the underestimation of the AEMET anomaly in these areas is less pronounced than for WRF-DPLE$_4$ (Figure 7.11c). These results lead to the positive MSSS$_4$ values observed in those regions (Figure 7.14b). However, the performance of WRF-DPLE$_4$ is better in part of the western regions and some locations in the northeastern Mediterranean coast. In the rest of the domain, both ensembles perform similarly. A different situation is observed in lead years 2–5, when the WRF-DPLE$_{10}$ underestimation of the anomaly is more accentuated in the northwest, the northeast, the central part of the IP and regions over the Baetic System and surrounding the Guadalquivir Valley, leading to negative MSSS$_4$ values. WRF-DPLE$_{10}$ reduces the overestimation observed for WRF-DPLE$_4$ in the rest of the domain (Figure 7.11f), getting positive MSSS$_4$ values in those places.

At seasonal scale, the results obtained for the comparison between WRF-DPLE$_{10}$ and WRF-DPLE$_4$ highly depend on the season and lead time (Figure 7.15). For example, the best results in favour of WRF-DPLE$_{10}$ have been obtained for lead years 1 and 2–5 in DJF and SON, respectively, as the domain is almost entirely dominated by the

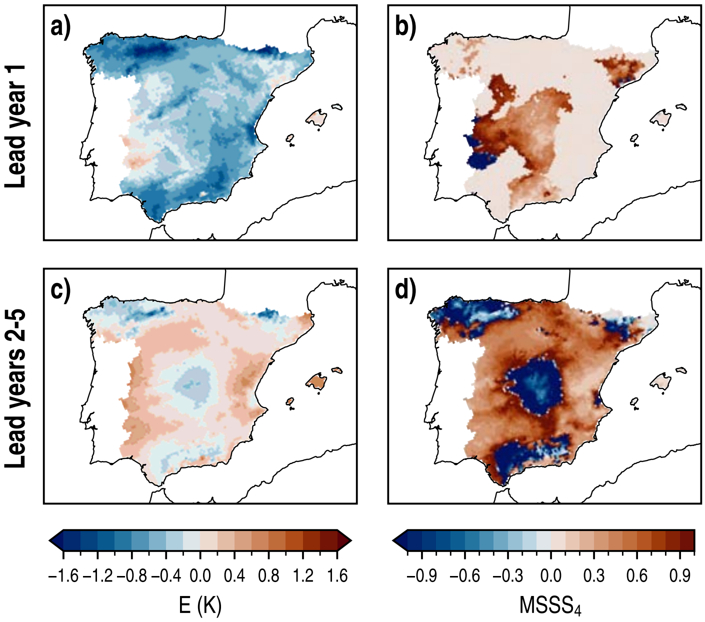

Figure 7.14: Spatial distributions of anomaly errors (E, left column) for the WRF-DPLE$_{10}$ multiannual mean anomalies of T_{min}, with AEMET as the observational dataset, and MSSS calculated with WRF-DPLE$_4$ as reference (MSSS$_4$, right column) in lead years 1 and 2–5 (rows) at annual scale.

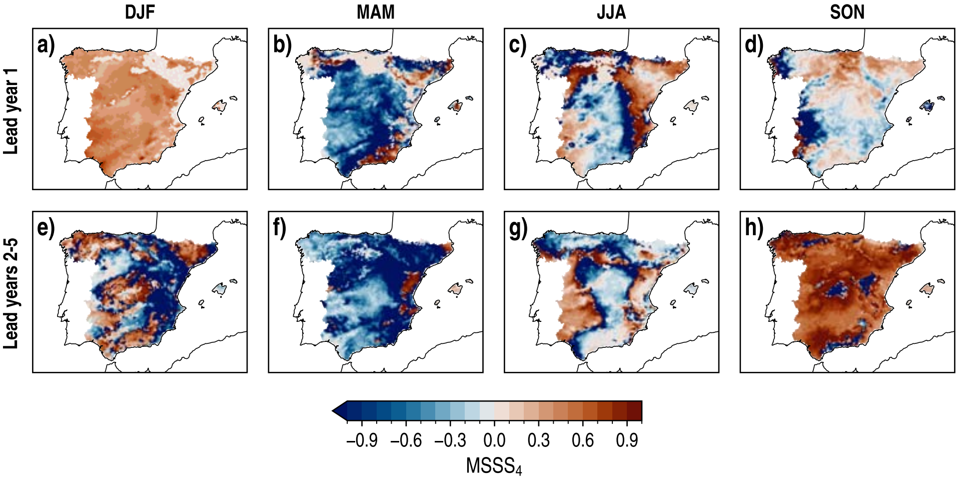

Figure 7.15: Spatial distributions of the MSSS for the WRF-DPLE$_{10}$ multiannual mean anomalies of T_{min}, with WRF-DPLE$_4$ as reference (MSSS$_4$), for lead years 1 and 2–5 (rows) in DJF, MAM, JJA and SON (columns).

positive results. On the other hand, the best results for WRF-DPLE$_4$ in comparison with WRF-DPLE$_{10}$ are observed in MAM at both lead times, when the MSSS$_4$ outcomes are fundamentally negative for most part of the domain. In JJA, WRF-DPLE$_{10}$ has obtained much better estimations of the anomalies in some southeastern regions, part of the Northern Subplateau and other northern locations in lead year 1, as the high MSSS$_4$ outcomes indicate. The placement of the positive MSSS$_4$ values in this season slightly changes in lead years 2–5, with the Northern Subplateau maintaining some of the highest scores. The spatial distributions of Figure 7.15 are explained by the comparison between the higher or lower absolute errors obtained by WRF-DPLE$_{10}$ (Figure B.69 in Appendix B.3.3) relative to WRF-DPLE$_4$ (Figure B.67).

7.3.3. *Predictions for regional averages*

The analysis of the downscaled future DCPs for T_{min} has been completed with the evaluation of the results for the multiannual mean anomalies of the field after being spatially averaged over the regions resulting from the regionalization described in Section 3.6 (Figure 3.5b). The results have been summarized in Table 7.3.

The outcomes found for the WRF-DPLE$_4$ T_{min} reflect what has already been mentioned in the analysis at grid-point scale. The anomalies are unanimously positive over the whole domain, a result consistent with the marked positive trend of T_{min} mentioned in Section 5.2.1 (see also Table A.2 in Appendix A.2). For each region, the highest WRF-DPLE$_4$ anomalies are observed always in lead years 2–5, with the most accentuated outcome found in the CS region ($T_{\mathrm{min}} = 1.27$ K). However, there is a general lack of reliable predictions in almost all regions at all lead times also in this case. The only regions where the predictions are reliable are the NE, WI and NO regions in lead year 1, as well as the latter in lead years 6–9. As happened for PR (Table 7.1) or T_{max} (Table 7.2), the confidence intervals defined by $\pm\Delta T_{\mathrm{min},90}$ exhibit widths which decrease as the length of the averaging window increases. Moreover, note that the confidence intervals are even narrower than those for T_{max}, partially due to the weaker presence of the signal-to-noise paradox in this variable compared to the other (see Figures 5.3 and 5.14). As also happened for T_{max}, the WRF-DPLE$_4$ E outcomes are lower (in absolute value) in lead years 2–5 compared to lead year 1, despite the anomalies are higher precisely in lead years 2–5. The lowest magnitudes of E have been found in the CI, NO and CS regions in lead years 2–5 with values of 0.08 K, 0.06 K and 0.03 K, respectively. The errors made by WRF-DPLE$_{10}$ are generally less pronounced than those for WRF-DPLE$_4$ at both lead times. For this reason, the results obtained for MSSS$_4$ are fundamentally positive. The only exception is the MT

Table 7.3 : Skill scores for the spatially averaged WRF-DPLE multiannual mean anomalies of T_{min} in lead years 1, 2-5, 6-9 and 2-9 for the decade starting in 2015 at annual scale. T_{min} is the anomaly for the WRF-DPLE$_{\mathrm{N}}$ (the N-member WRF-DPLE ensemble) ensemble mean, $\pm\Delta T_{\mathrm{min,90}}$ represents half the width of the 90 % conficence interval for a single WRF-DPLE member, E is the anomaly error, with AEMET as the observational dataset, and MSSS$_4$ denotes the added value of WRF-DPLE$_{10}$ over WRF-DPLE$_4$. Only for the WRF-DPLE$_4$ T_{min}, the bold formatting denotes that WRF-DPLE$_4$ is able to represent the forecast uncertainty and that the 90 % confidence interval encloses the AEMET anomaly; the symbol "(∗)", if any, means that the former is satisfied but the latter is not; finally, the plain formatting is used when WRF-DPLE$_4$ cannot represent the forecast uncertainty. Dashes denote data unavailability at that lead time.

Region	Lead years	WRF-DPLE$_4$			WRF-DPLE$_{10}$		
		T_{min} (K)	$\pm\Delta T_{\mathrm{min,90}}$ (K)	E (K)	T_{min} (K)	E (K)	MSSS$_4$
SW	1	0.84	0.80	-0.43	0.91	-0.36	0.32
	2-5	1.05	0.34	0.28	0.96	0.18	0.57
	6-9	0.78	0.34	–	0.84	–	–
	2-9	0.92	0.23	–	0.90	–	–
NO	1	**0.52**	0.77	-0.73	0.54	-0.72	0.03
	2-5	0.97	0.32	0.06	0.88	-0.03	0.82
	6-9	**0.67**	0.33	–	0.73	–	–
	2-9	0.82	0.23	–	0.81	–	–
CI	1	0.89	0.81	-0.51	0.98	-0.42	0.31
	2-5	1.10	0.32	0.08	1.01	0.00	1.00
	6-9	0.83	0.35	–	0.89	–	–
	2-9	0.96	0.22	–	0.95	–	–
NE	1	**0.64**	0.75	-0.45	0.68	-0.41	0.15
	2-5	1.05	0.32	0.15	1.00	0.10	0.53
	6-9	0.89	0.33	–	0.95	–	–
	2-9	0.97	0.22	–	0.97	–	–
CS	1	**0.94** $^{(*)}$	0.74	-0.89	1.02	-0.82	0.16
	2-5	1.27	0.37	0.03	1.22	-0.02	0.45
	6-9	0.91	0.36	–	1.00	–	–
	2-9	1.09	0.24	–	1.11	–	–
EA	1	0.60	0.80	-0.56	0.61	-0.56	0.01
	2-5	1.07	0.30	0.34	1.01	0.29	0.30
	6-9	0.89	0.31	–	0.98	–	–
	2-9	0.98	0.20	–	0.99	–	–
MT	1	**0.63** $^{(*)}$	0.84	-0.93	0.69	-0.88	0.12
	2-5	1.11	0.44	-0.16	1.05	-0.22	-0.93
	6-9	0.88	0.39	–	0.90	–	–
	2-9	0.99	0.28	–	0.98	–	–
WI	1	**0.64**	0.88	-0.47	0.68	-0.42	0.18
	2-5	1.07	0.33	0.35	0.97	0.24	0.51
	6-9	0.78	0.36	–	0.80	–	–
	2-9	0.92	0.24	–	0.88	–	–

region in lead years 2–5, where the E values are -0.16 K and -0.22 K for WRF-DPLE$_4$ and WRF-DPLE$_{10}$, respectively. In the same line as T_{max}, the highest MSSS$_4$ results are observed in lead years 2–5, found in the NO and CI regions with values of 0.82 and 1, respectively.

7.4. Daily mean near-surface air temperature

7.4.1. *Analysis of the WRF-DPLE$_4$ predictions*

This Section is devoted to present the downscaled future DCPs for T_{mean} in the decade starting in 2015. As in previous sections, the anomalies at all lead times and annual scale for WRF-DPLE$_4$ are positive over the whole domain (Figure 7.16). These results are reliable over most part of the domain at all lead times, with the exception of lead year 1, when the reliability is mainly shown in the northern half of the IP, as indicated by the results obtained for CRPSS in Figure 5.28. In lead year 1, the anomalies have values ranging from 0.4 K to 0.8 K over the northern regions with reliable predictions, with the maximum values between 0.8 K and 1 K found in the southwestern sector of the domain. The confidence intervals defined by $\pm\Delta T_{\mathrm{mean,90}}$ are again too wide compared to the anomalies at this lead time, up to the point of enclosing results with different signs (Figure 7.16b). In the same regions which show the aforementioned highest anomalies, $\pm\Delta T_{\mathrm{mean,90}}$ values are between 0.9 K and 1 K. They are even larger in some locations with reliable predictions over the central IP, where $\pm\Delta T_{\mathrm{mean,90}}$ is even above 1 K. The results obtained for E indicate that the AEMET anomaly is underestimated over most part of the domain at this lead time, excepting for a few locations in the Baetic System (Figure 7.16c). This underestimation presents minimum values around -1 K over some locations with prediction reliability to the north of the Iberian System, along the Cantabrian Range and in the Pyrenees. On the other hand, part of the lowest error magnitudes have been found in the same southwestern regions which also show the highest anomalies, with E values down to -0.2 K. There are some regions where, despite they show reliable predictions, the AEMET anomalies are not inside the confidence intervals defined by $\pm\Delta T_{\mathrm{mean,90}}$. In most part of the cases, with the exception of that found in the Baetic System, the AEMET anomalies are higher than the upper boundary of the confidence intervals. Although there is a probability of 10 % associated to these occurrences, the same factors previously mentioned for the other variables (i.e., the spin-up biases, the sample size of the control period and the gap between the end of this period and the decade 2015–2025) may also affect these results (see Section 7.1.1).

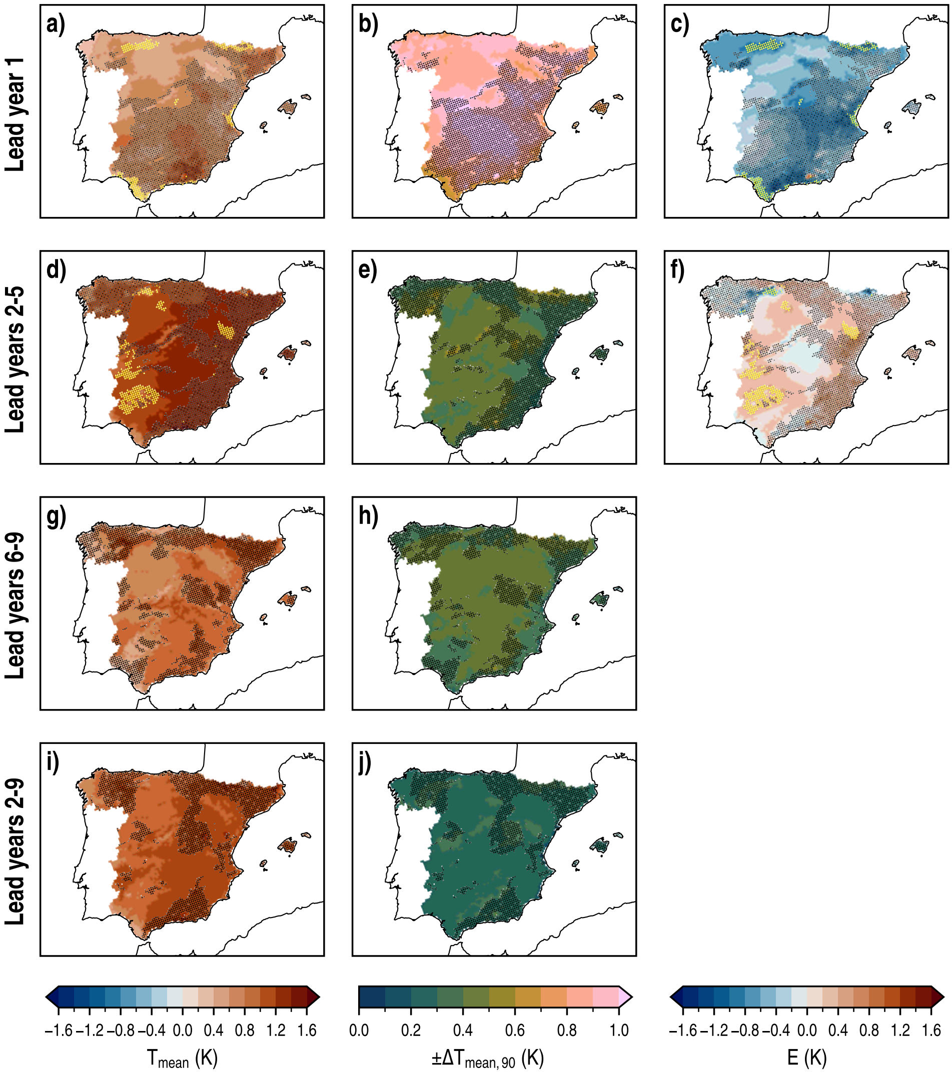

Figure 7.16: Spatial distributions of the WRF-DPLE$_4$ multiannual mean anomalies of T_{mean} (left column), half the width of the 90 % confidence interval for a single WRF-DPLE$_4$ member ($\pm\Delta T_{\mathrm{mean,90}}$, center column) and the anomaly errors (E, right column), with AEMET as the observational dataset, at annual scale for several lead times (rows). The absence (presence) of black dots denote the locations where the forecast uncertainty is (not) represented by the confidence intervals. In T_{min} and E maps, yellow triangles identify the locations where the predictions are reliable but the confidence intervals do not contain the AEMET anomalies.

The anomalies are higher in lead years 2–5, when they are fundamentally between 1 K and 1.4 K in those regions showing reliable results (Figure 7.16d). There are smaller regions with lower anomalies, mainly found in the western half of the domain with values down to 0.4 K. The maximum outcomes are observed over the eastern half of the IP and to the south and southeast of the Northern Subplateau, with results around 1.2–1.4 K. The results obtained for $\pm\Delta T_{\mathrm{mean},90}$ at this lead time (Figure 7.16e) are much lower that in lead year 1, with values commonly below 0.5 K. These confidence intervals do not enclose anomalies with different signs for most part of the domain, with the exception of some locations to the southeast of the Northern Subplateau and close to the Strait of Gibraltar with low anomalies. The spatial distribution of E shows errors with generally lower magnitudes than in lead year 1 across larger areas (Figure 7.16f). In locations with reliable predictions, the absolute value of these errors is almost always below 0.4 K, reaching outcomes between -0.2 K and 0.2 K in the central IP, the eastern part of the Northern Subplateau and southern areas. The number of locations with reliable estimations of the anomaly and confidence intervals which do not contain the AEMET result is slightly larger than in lead year 1. In this case, the effect of the spin-up issue do not affect as much as as it could do at the other lead time, since the spin-up period needed by T_{mean} after the initialization with extreme soil moisture conditions is commonly below 1 year and, mostly, do not surpass 2 years. Indeed, if more normal soil moisture ICs are used to initialize the simulations, even shorter spin-up period could be expected, as mentioned for the previous variables.

The WRF-DPLE$_4$ anomalies of T_{mean} are generally lower in lead years 6–9 than in lead years 2–5 (Figure 7.16g), as happened for the other NSAT variables. The highest outcomes are mainly observed in part of the eastern and southern halves of the domain in regions with reliable predictions. The regions with reliable predictions are fundamentally covered by anomalies above 0.6 K. The maximum outcomes reach values around 1–1.2 K in some central locations and in the northeast of the IP. On the other hand, the lower values have been fundamentally found in the southwestern regions and over some locations in the Northern Subplateau. As in lead years 2–5, the results obtained for $\pm\Delta T_{\mathrm{mean},90}$ show values below 0.5 K across the domain. Even lower values than 0.4 K can be found in some regions along the Mediterranean coast, the northern part of the domain and in the southwest. Higher anomalies have been found in lead years 2–9, especially in the eastern half of the domain in regions with reliable results, with values between 1 K and 1.2 K. The outcomes in the western part are more moderate, but they are fundamentally above 0.8 K in most part of the domain, with the exception of some locations mainly grouped in the northwest and southwest. The narrowest confidence intervals have been found at this lead time, with values mostly lower than 0.3 K for ±$\Delta T_{\text{mean},90}$, although the outcomes can reach values up to 0.4 K in some cases.

The multiannual mean anomalies of T_{mean} at seasonal scale have been represented in Figure 7.17. As at annual scale, most part of the results obtained for those regions with reliable predictions are positive, with the exception of some regions spread across the domain in DJF for lead year 1, as well as a few locations found in JJA and SON at the same lead time. The highest anomalies are observed in JJA, especially in lead years 6–9 and 2–9, which is also the season when the predictions are generally more reliable. At both lead times, large areas with values around 1.5–1.75 K can be

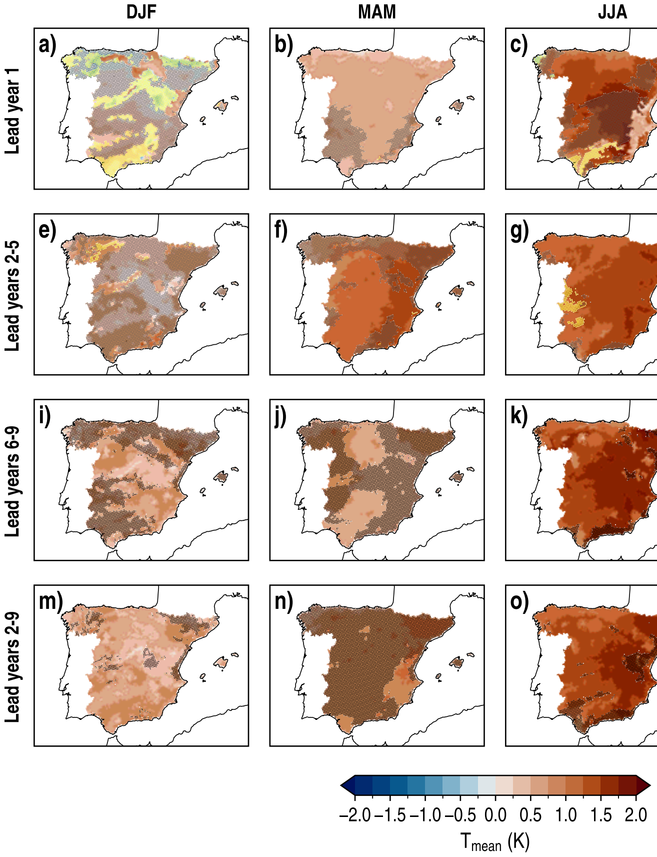

Figure 7.17 : Spatial distributions of the WRF-DPLE$_4$ multiannual mean anomalies of T_{mean} for lead years 1, 2–5, 6–9 and 2–9 (rows) in DJF, MAM, JJA and SON (columns). The absence (presence) of black dots denote the locations where the forecast uncertainty is (not) represented by the confidence intervals. Yellow triangles identify the locations where the predictions are reliable but the confidence intervals do not contain the AEMET anomalies.

found, with maximum values even above 1.75 K in some smaller regions in lead years 2–9. The lowest anomalies in this season have been found in lead year 1, with positive values ranging from 0.25 K to 0.75 K in southern and eastern regions close to the Mediterranean coast. Another season with high reliable predictions, fundamentally in lead years 1 and 2–5, is MAM. In this season, the highest anomalies are observed in the eastern flank for lead years 2–5, with outcomes around 1.25–1.5 K. The other regions commonly show results between 1 K and 1.25 K at this time. The anomalies are slightly lower in MAM at the other lead times, but commonly with values above 0.5 K over most part of the domain with reliable predictions. While SON hardly shows regions with reliable results, they span large parts of the domain in DJF for lead years 6–9 and 2–9, with values predominantly between 0.25 K and 1 K, which are slightly higher in lead years 6–9. The largest number of locations with reliable results and AEMET anomalies outside the confidence intervals is found in DJF, as occurred for T_{max} (Figures 7.17a and 7.17e). In this case, they are much more frequent in lead year 1, covering vast areas across the whole domain.

The spatial distributions of $\pm\Delta T_{\mathrm{mean},90}$ and E for each season are available in Figures B.70 and B.71 (Appendix B.3.4), respectively. As at annual scale, the width of the confidence intervals decreases with the length of the averaging window. This width is larger than at annual scale because the T_{mean} variability at seasonal scale is also larger. The $\pm\Delta T_{\mathrm{mean},90}$ values are lower than the anomaly in many occasions, as happened for the other NSAT variables and in contrast to PR. This is a consequence of the presence of a weaker signal-to-noise paradox in this case compared to PR (see Figures 4.5 and 5.26). The results obtained for E show the most accentuated errors in DJF, when the anomaly is highly underestimated, and in MAM, when it is generally overestimated, both in lead year 1. In both cases, the maximum absolute E values reach results above 1.6 K. In regions with reliable predictions, lower absolute values can be found for lead year 1 in JJA or for lead years 2–5 in MAM and JJA. The lowest errors are between -0.2 K and 0.2 K and their location vary depending on the season and lead time.

7.4.2. *Comparison with the WRF-DPLE$_{10}$ ensemble*

The anomalies obtained for the WRF-DPLE$_{10}$ T_{mean} at annual scale have been depicted in Figure 7.18. There are evident similarities between these spatial patterns and those represented in Figure 7.16 in both qualitative and quantitative terms. There are still some differences which can be observed if the spatial distributions are carefully examined. For example, the magnitude of the anomalies predicted by WRF-DPLE$_{10}$

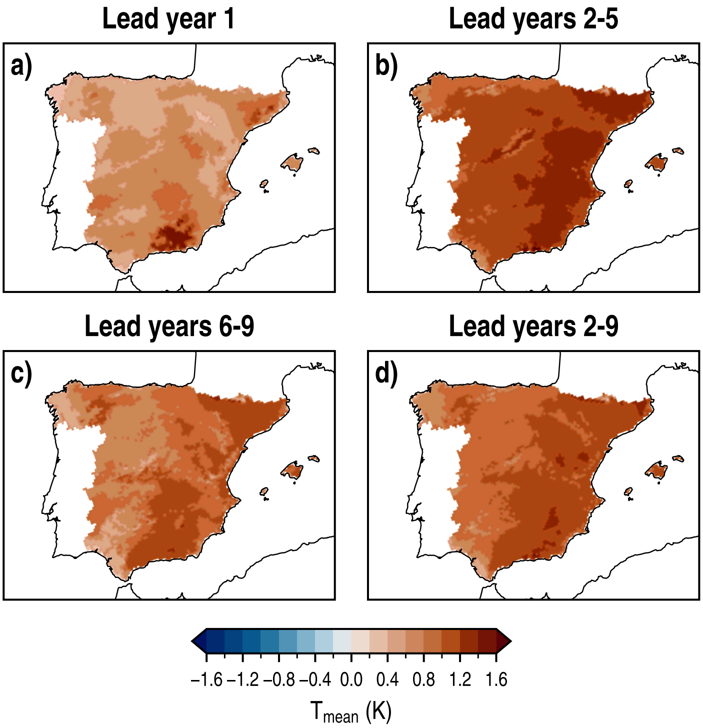

Figure 7.18: Spatial distributions of the WRF-DPLE$_{10}$ multiannual mean anomalies of T_{mean} in lead years 1, 2–5, 6–9 and 2–9 at annual scale.

(Figure 7.18a) in the Baetic System is higher than that predicted by WRF-DPLE$_4$ (Figure 7.16a). In lead years 2–5, the WRF-DPLE$_{10}$ anomalies (Figure 7.18b) along the Mediterranean coast, the Cantabrian coast and the central IP are slightly lower than for the 4-member ensemble (Figure 7.16d). By contrast, WRF-DPLE$_{10}$ has obtained more accentuated anomalies in lead years 6–9 (Figure 7.18c), especially over the central southern regions of the IP and part of the Balearic Islands. In lead years 2–9, the largest differences can be observed in the southwestern regions, where the anomalies are lower in the case of WRF-DPLE$_{10}$ (Figure 7.18d). There are also differences between the anomalies of both ensembles at seasonal scale. The results obtained for WRF-DPLE$_{10}$ have been represented in Figure B.72 (available in Appendix B.3.4). Some of the most remarkable discrepancies are shown in the northern regions for lead year 1 in DJF, where negative anomalies obtained for WRF-DPLE$_4$ (Figure 7.17a) turn into positive in the case of WRF-DPLE$_{10}$ (Figure B.72a). Moreover, the anomalies observed in JJA also for lead year 1, over some central inner regions, are lower for

WRF-DPLE$_{10}$ (Figure B.72c). In the same season, also lower anomalies for WRF-DPLE$_{10}$ have been generally found over most part of the domain in lead years 6–9 and 2–9 (Figures B.72k and B.72o, respectively). More differences can be found depending on the lead time and season.

The spatial distributions obtained for the WRF-DPLE$_{10}$ E and MSSS$_4$ have been represented in Figure 7.19. The differences between WRF-DPLE$_{10}$ and WRF-DPLE$_4$ errors lead to the patterns observed for MSSS$_4$. In lead year 1, there are almost not discrepancies between the ensemble performances, with the exception of small areas in the northwest, the northeast and the Baetic System. The MSSS$_4$ values equal to zero are observed because the predictions have been recalibrated in such regions for both ensembles, as mentioned in Section 7.1.2. In the rest of the domain, only a few locations in the Baetic System show a worse skill for WRF-DPLE$_{10}$ in estimating the AEMET anomaly, whereas it outperforms WRF-DPLE$_4$ in the others. Higher differences, although not easily appreciable in terms of E, are observed in lead years 2–5, when negative the MSSS$_4$ results mainly found in the northwest, the Pyrenees, the central IP and in some southern regions indicate a better estimation of

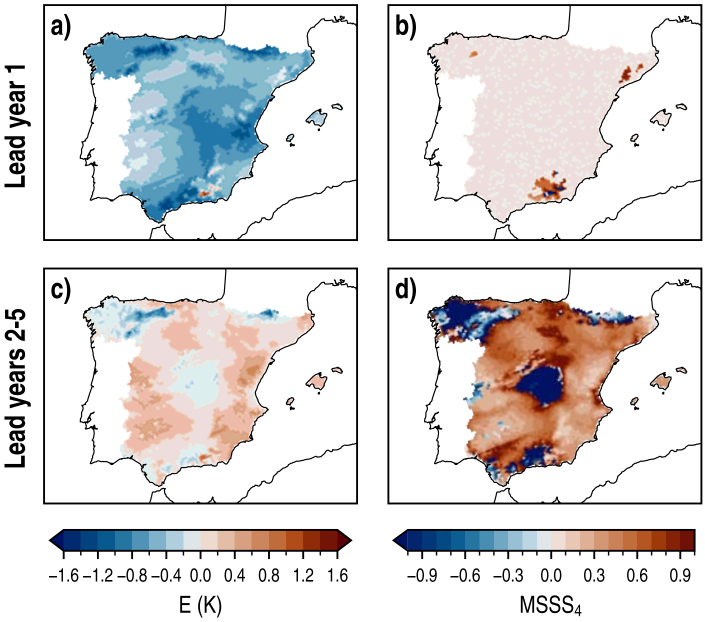

Figure 7.19: Spatial distributions of anomaly errors (E, left column) for the WRF-DPLE$_{10}$ multiannual mean anomalies of T_{mean}, with AEMET as the observational dataset, and MSSS calculated with WRF-DPLE$_4$ as reference (MSSS$_4$, right column) in lead years 1 and 2–5 (rows) at annual scale.

the actual anomaly by WRF-DPLE$_4$ compared to WRF-DPLE$_{10}$. Better results have been obtained for WRF-DPLE$_{10}$ elsewhere.

At seasonal scale, the results obtained for MSSS$_4$ and the WRF-DPLE$_{10}$ E are available in Figures 7.20 and B.73 (Appendix B.3.4), respectively. While the large areas with positive MSSS$_4$ results (and slightly lower absolute E values for WRF-DPLE$_{10}$) in DJF and SON for lead year 1 indicate a general better performance for WRF-DPLE$_{10}$ (with the exception, mainly, of some western regions and northeastern regions in the case of SON), better results for WRF-DPLE$_4$ are predominant in MAM and, to a lesser degree, in JJA at this lead time. In lead years 2–5, the 10-member ensemble continues performing better in general in DJF and SON, whereas WRF-DPLE$_4$ has obtained better results in most part of the domain in MAM, although the eastern flank and some regions in the southwestern sector are dominated by WRF-DPLE$_4$. Differences are also observed in JJA, although the areas covered by MSSS$_4$ results between -0.1 and 0.1 have larger extensions than at the other lead time and seasons.

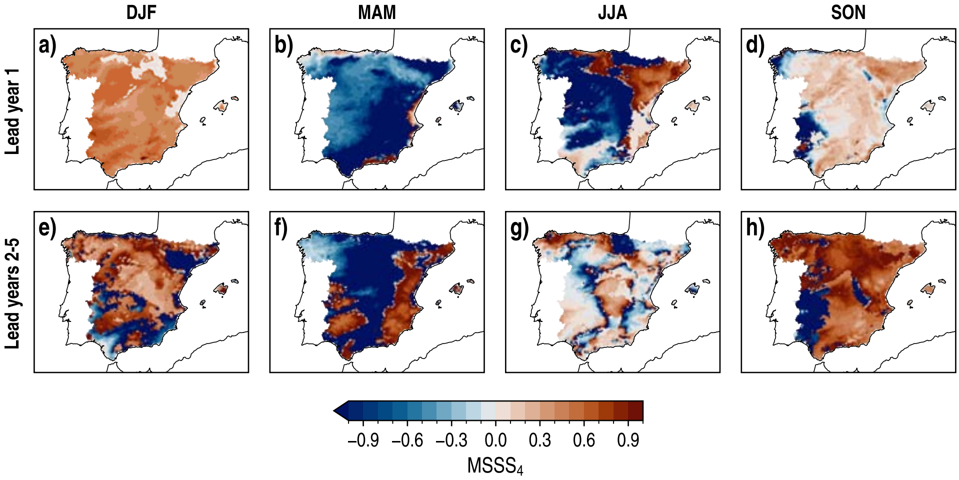

Figure 7.20: Spatial distributions of MSSS for the WRF-DPLE$_{10}$ multiannual mean anomalies of T_{mean}, with WRF-DPLE$_4$ as reference (MSSS$_4$), for lead years 1 and 2–5 (rows) in DJF, MAM, JJA and SON (columns).

7.4.3. *Predictions for regional averages*

The predictions for the spatially averaged multiannual mean anomalies of T_{mean} in each region resulting from the regionalization described in Section 3.6 (Figure 3.5b) are available in Table 7.4. The WRF-DPLE$_4$ anomalies depict the same behaviour observed at grid-point scale, since the highest values have been found in all regions

Table 7.4: Skill scores for the spatially averaged WRF-DPLE multiannual mean anomalies of T_{mean} in lead years 1, 2-5, 6-9 and 2-9 for the decade starting in 2015 at annual scale. T_{mean} is the anomaly for the WRF-DPLE$_{\text{N}}$ (the N-member WRF-DPLE ensemble) ensemble mean, $\pm\Delta T_{\text{mean,90}}$ represents half the width of the 90 % conficence interval for a single WRF-DPLE member, E is the anomaly error, with AEMET as the observational dataset, and MSSS$_4$ denotes the added value of WRF-DPLE$_{10}$ over WRF-DPLE$_4$. Only for the WRF-DPLE$_4$ T_{mean}, the bold formatting denotes that WRF-DPLE$_4$ is able to represent the forecast uncertainty and that the 90 % confidence interval encloses the AEMET anomaly; the symbol "(∗)", if any, means that the former is satisfied but the latter is not; finally, the plain formatting is used when WRF-DPLE$_4$ cannot represent the forecast uncertainty. Dashes denote data unavailability at that lead time.

Region	Lead years	WRF-DPLE$_4$			WRF-DPLE$_{10}$		
		T_{mean} (K)	$\pm\Delta T_{\text{mean,90}}$ (K)	E (K)	T_{mean} (K)	E (K)	MSSS$_4$
SW	1	0.68	0.81	-0.50	0.68	-0.50	0.00
	2-5	**1.09**	0.40	0.28	1.03	0.21	0.42
	6-9	**0.73**	0.37	–	0.74	–	–
	2-9	**0.91**	0.26	–	0.89	–	–
NO	1	**0.52**	0.87	-0.62	0.52	-0.62	0.00
	2-5	1.03	0.38	0.11	0.94	0.02	0.97
	6-9	0.71	0.37	–	0.74	–	–
	2-9	0.87	0.26	–	0.84	–	–
CI	1	0.72	0.98	-0.69	0.72	-0.69	0.00
	2-5	**1.25**	0.42	0.16	1.19	0.10	0.59
	6-9	**0.88**	0.42	–	0.93	–	–
	2-9	**1.07**	0.28	–	1.06	–	–
NE	1	0.66	0.87	-0.51	0.67	-0.50	0.03
	2-5	1.26	0.37	0.22	1.19	0.16	0.51
	6-9	0.94	0.40	–	0.99	–	–
	2-9	1.10	0.26	–	1.09	–	–
CS	1	0.99	0.87	-0.70	1.08	-0.61	0.24
	2-5	1.30	0.40	0.19	1.24	0.12	0.55
	6-9	**0.97**	0.41	–	1.08	–	–
	2-9	1.13	0.27	–	1.16	–	–
EA	1	0.64	0.85	-0.68	0.64	-0.68	0.00
	2-5	1.28	0.33	0.40	1.20	0.33	0.33
	6-9	**0.88**	0.37	–	0.95	–	–
	2-9	**1.08**	0.23	–	1.08	–	–
MT	1	**0.65**	0.91	-0.80	0.65	-0.80	0.01
	2-5	**1.15**	0.46	-0.16	1.10	-0.21	-0.79
	6-9	0.90	0.43	–	0.90	–	–
	2-9	**1.02**	0.30	–	1.00	–	–
WI	1	**0.58**	0.87	-0.45	0.58	-0.45	0.00
	2-5	**1.15**	0.42	0.23	1.09	0.16	0.49
	6-9	**0.76**	0.41	–	0.76	–	–
	2-9	**0.96**	0.28	–	0.93	–	–

for lead years 2–5 and 2–9. Almost all regions show reliable results at least at one lead time window, with the exception of the NE region, where there is a complete lack of reliable predictions. The regions which most stand out in terms of reliable predictions are the WI, MT, CI and SW regions, with 3 or 4 lead time windows when results are reliable. Among these regions, the highest anomalies are observed in the CI region, with values of 1.05 K, 0.88 K and 1.07 K in lead years 2–5, 6–9 and 2–9, respectively. The lowest value, on the other hand, is observed in lead year 1 in the NO region, with T_{mean} = 0.52 K. This is the region with the lowest values in general, but they are not reliable at other lead times. As happened in previous sections, the confidence intervals defined by the WRF-DPLE$_4$ $\pm\Delta T_{\mathrm{mean,90}}$ are narrower as the length of the averaging window increases, so the lowest width is always observed in lead years 2–9. The E values obtained for WRF-DPLE$_4$ in lead years 2–5 are lower than those in lead year 1. In addition, while there is a underestimation of the AEMET anomaly in lead year 1 in all regions, it is generally overestimated in lead years 2–5, with the exception of the MT region. The lowest magnitudes of the error observed in regions with reliable results have been found for lead years 2–5 in the CI and MT regions (E = 0.16 K and E = −0.16 K, respectively), both with T_{mean} anomalies above 1 K. The errors obtained by WRF-DPLE$_{10}$ are qualitatively similar to those for WRF-DPLE$_4$, although they are lower in absolute value in almost all cases, with the exception of the MT region in lead years 2–5, when the estimation of the AEMET anomaly is more accurate for WRF-DPLE$_4$. This is also reflected in the results shown for MSSS$_4$, for which this region is the only one with a negative result. In the same vein as for the other NSAT variables, the largest differences between WRF-DPLE$_4$ and WRF-DPLE$_{10}$ performances in terms of MSSS$_4$ are observed in lead years 2–5. Indeed, the MSSS$_4$ values are very close to zero in most part of the regions for lead year 1, with the exception of the CS region, with a result of MSSS$_4$ = 0.24.

7.5. Concluding remarks

In this Chapter, the WRF-DPLE predictions for the decade 2015–2025 have been analyzed. Given the results obtained in Chapter 6, these simulations were initialized from a dynamically equilibrated soil state to reduce the biases related to the spin-up at the beginning of the decade. The analysis has been focused on the same variables examined in Chapters 4 and 5, i.e., PR and the NSAT variables (T_{max}, T_{min} and T_{mean}). Additionally, the full 10-member CESM-DPLE subensemble available for DD has been downscaled in this case. The impact of the ensemble size on the predictions has been also evaluated. As the AEMET dataset provides observational information up to 2022, the predictions have been compared with the observational value in lead years 1 and 2–5. The most relevant findings are the following:

- **At annual scale, the predicted WRF-DPLE$_4$ anomalies for PR are generally positive at the beginning of the decade in regions with reliable predictions, whereas they mostly turn into negative during the second half of the decade.** The anomalies are also negative for lead years 2–9. The most intense negative anomalies in lead years 6–9 and 2–9 have been found in the Pyrenees, some northwestern regions and the Central System. In some cases, these anomalies can reach values below -12 mm/month. However, the width of the confidence intervals for these results is wide enough to enclose values with opposite signs at all lead times. At seasonal scale, some of the largest areas with reliable negative anomalies are shown for SON in lead years 2–9. There are many regions with reliable predictions showing observational values out of the confidence intervals in MAM. Although there is a probability of 10 % associated to these occurrences, they may be also partially due to the fact that the confidence intervals have been calculated in the control period. Since there is a 15-year gap between the end of the control period (which has a sample size of 30 start dates) and the beginning of the decade 2015–2025, the confidence intervals may not be totally appropriate to account for the uncertainty in this decade in those specific locations. Additionally, the experiments for the decade 2015–2025 were initialized from a dynamically equilibrated soil state, in contrast to the hindcasts, so the spin-up biases may have also affected the calculation of the confidence intervals. However, as the spin-up period for precipitation is commonly shorter than 10 months, these biases would only influence on the results for the lead year 1.

- **Qualitatively similar predictions have been obtained from WRF-DPLE$_{10}$ for PR, but in general with lower errors than WRF-DPLE$_4$ at annual scale for lead years 1 and 2–5.** Moreover, the WRF-DPLE$_{10}$ predictions tend to show more moderate anomalies than WRF-DPLE$_4$. At seasonal scale, the areas where WRF-DPLE$_{10}$ outperforms WRF-DPLE$_4$ are commonly larger than those where WRF-DPLE$_4$ gets a better accuracy, with the exceptions of MAM and JJA in lead year 1. In these cases, WRF-DPLE$_4$ performs better over most part of the domain.

- **From a regional perspective, the spatially averaged anomalies reproduce what has been observed in the grid-point analysis**. The most intense reliable predicted anomalies are negative and have been found in the CN and NW regions, in lead years 6–9 and 2–9, for WRF-DPLE$_4$. Qualitatively similar results are shown by the WRF-DPLE$_{10}$ predictions, but typically getting lower errors in lead years 1 and 2–5.

- **The predicted anomalies for the NSAT variables are positive at annual scale over the whole domain for all lead times.** In regions with reliable predictions from WRF-DPLE$_4$, the most intense anomalies have been found at lead years 2–5 for the three NSAT variables. The anomalies are commonly higher than 1 K, with the maximum values between 1.5 K and 1.75 K, shown for T_{max}, in some regions over the Iberian System. The anomalies can be even higher at seasonal scale. The highest values are shown in JJA for the three variables, reaching outcomes up to 2 K in large areas of the domain with reliable predictions for T_{max} in lead years 6–9. The confidence intervals calculated at annual scale generally do not contain anomalies with distinct signs in lead years 2–5, 6–9 and 2–9 for any variable. As for PR, the observational values fall outside the confidence intervals in some cases, being more frequent in DJF for T_{max} and T_{mean}. The impact of the spin-up biases on these results is more important for the NSAT variables than for PR, since the spin-up period is longer for the former.
- **The differences between the WRF-DPLE$_4$ and WRF-DPLE$_{10}$ predicted anomalies are generally small at annual scale.** The areas where WRF-DPLE$_{10}$ outperforms WRF-DPLE$_4$ are commonly larger than those showing the opposite in lead years 1 and 2–5. Larger differences between the ensembles can be found at seasonal scale, with some locations showing anomalies with different sings depending on the ensemble size.
- **The predicted spatially averaged NSAT anomalies obtained at annual scale summarize for each region the results observed from a grid-point perspective**. Positive anomalies have been found for the three NSAT variables, regardless of the region, the lead time and the ensemble size. The anomalies are typically higher than 0.5 K, and they often reach values higher than 1 K. The confidence intervals do not commonly contain values with different signs, excepting those obtained in lead year 1. In general, the accuracy of the WRF-DPLE$_{10}$ predictions is better than that of WRF-DPLE$_4$ in lead years 1 and 2–5, with a few exceptions.

8

Drift correction techniques for decadal climate predictions

The contents of this Chapter are based on the study carried out in Rosa-Cánovas et al. (2023), which explores the ability of several drift correction methods to adjust the drift in DCPs and provides guidance to select a subensemble of decadal experiments for DD applications. In the following section, the main purpose of this study has been described. Then, the methodology applied to correct the drift, evaluate the performance of the methods and select the subensemble has been detailed. Finally, the results and the main concluding remarks have been presented.

8.1. The need for a skilful adjustment of the drift in decadal climate predictions

The concept of climate drift in DCPs and the importance of adjusting it to generate skilful climate predictions have been previously addressed in Sections 1.1.2 and 3.3. After the initialization of a decadal experiment, the predicted climate progressively drifts away from the initial state determined by the observations towards the imperfect model climatology (Meehl et al., 2009, 2014). This drift produces lead time-dependent biases in forecasts which must be addressed to properly evaluate their predictive skill or to use them as input information in DD simulations. An illustrative example which shows how the drift correction works has been depicted in Figure 8.1. This Figure shows the evolution of the global mean SST predicted by the 40-member CESM-DPLE ensemble mean for a set of raw (i.e., uncorrected) and drift-corrected decadal experiments (coloured dashed and solid thin lines, respectively), the same variable produced by the raw 40-member CESM-LE ensemble mean (black dotted line) and the observational information provided by ERSST5 (black solid thick line). The raw decadal experiments approach or drift to the path followed by the uninitialized

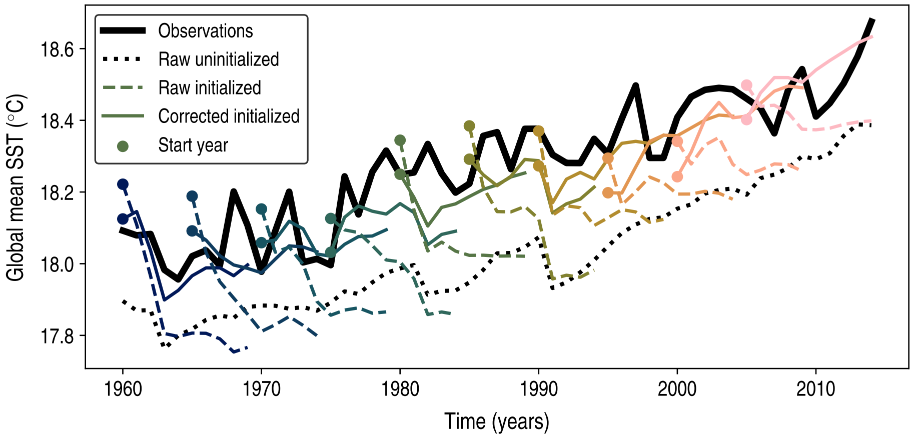

Figure 8.1: Example of the drift correction for the 40-member CESM-DPLE ensemble mean. The global mean SST is depicted for the raw (i.e., uncorrected) CESM-DPLE ensemble mean (coloured dashed thin lines), the CESM-DPLE ensemble mean corrected with the MDC (coloured solid thin lines), the raw CESM-LE ensemble mean (black dotted line) and the observational information provided by ERSST5 (black solid thick line).

experiments as the lead time increases. By contrast, the drift-corrected experiments keep closer to the observed SST.

As stated in Section 3.3, the MDC (Boer et al., 2016; CLIVAR, 2011) has been used in this Thesis to adjust the drift in the 4-member CESM-DPLE subensemble which provided the ICs and LBCs used to produce the WRF-DPLE experiments. Although this method contributes to reducing the mean lead time-dependent bias in CESM-DPLE, it does not account for higher-order biases, such as those observed in the representation of trends (see Section 5.4). Therefore, additional correction techniques have been evaluated in this Chapter to explore alternative approaches to MDC with the aim of improving as much as possible the predictive skill of the input data in DD simulations, which would lead to a more skilful downscaled product in potential future experiments. Taking into account that a reduction of the ensemble size inevitably leads to the loss of predictive skill, this analysis also assesses the impact of this reduction in a context of limited access to computing resources to conduct the DD simulations. This evaluation does not include the DeFoReSt approach (see Section 3.5; Pasternack et al., 2018, 2021) because it is not examined in the original work presented in Rosa-Cánovas et al. (2023). This method accounts not only for unconditional and conditional biases but also for the misrepresentation of the ensemble spread, so it may be considered as a potential candidate to correct the input information for DD simulations, and it should be included in future drift correction assessments.

The main aim of this analysis is to provide guidance for the selection of a drift-corrected 3-member subensemble (ENS3) from CESM-DPLE which is representative of the performance of the whole 40-member ensemble (ENS40) to some extent for several areas of interest. The selected subensemble could be used to conduct DD simulations for future studies, minimizing the computing requirements to conduct the DD while reducing as much as possible the loss of predictive skill in the final product and still allowing a representation of the uncertainty in the predictions comparable to that of the full ensemble.

The procedure to follow encompasses two steps:

1) Selection of the most appropriate method of drift correction to minimize the model drift in ENS40.
2) Selection of members to build ENS3 and evaluation of the impact of ensemble size on the subensemble performance.

8.2. Methodology

Several drift correction methods have been used to correct three climate fields: SST, NSAT anomaly and SLP. Since the reference data must provide all fields needed in the correction process, the adjustment of the drift has been conducted with ERA5 as the reference dataset (see Section 2.2.2). Then, the added value to the accuracy of predictions provided by each drift correction approach has been assessed for these three variables and a set of climate indices computed with them. The observational datasets used to evaluate the results of the drift correction have been described in Section 2.3.2. Since the skill of the correction methods could vary depending on the region (Choudhury et al., 2017), the analysis has been carried out by considering several domains of interest in the context of DD. They are the European (EUR), South American (SA) and North American (NA) domains defined by CORDEX (CORDEX, 2015; Giorgi and Gutowski, 2015; Giorgi et al., 2009). The area covered by each domain can be consulted in CORDEX (2015) or Figure B.82 (available in Appendix B.4). All methods and their corresponding evaluations have been applied to full fields.

8.2.1. *Description of the drift correction methods*

In addition to the MDC described and evaluated in Section 3.3, two more correction techniques have been examined. These are the trend-based drift correction (TrDC;

Kharin et al., 2012) and the initial condition-based drift correction (ICDC; Fučkar et al., 2014). Apart from the conventional formulations of these methods, two complementary approaches have been considered for them: the k-nearest neighbours (kNN; Choudhury et al., 2017) and the polynomial fitting (FIT; Gangstø et al., 2013; Kruschke et al., 2016) approaches.

❦ *Trend-based drift correction*

A model which does not properly capture long-term climate trends, such as the global warming, could produce decadal predictions which drift away from the observations with dependency on the initialization time. Since the MDC cannot be suitable for addressing this sort of bias, Kharin et al. (2012) proposed the TrDC method, which also considers climate trends in the assessment of the model drift. Following their approach, the ensemble mean $\{Y'\}_{j\tau}$, calculated by following Eq. [3.5], and the reference data $X'_{j\tau}$ can be written in terms of a first-order approximation with the initial date j as independent variable:

$$\begin{aligned} \{Y'\}_{j\tau} &= \alpha^Y_\tau + \beta^Y_\tau j + \epsilon^Y_{j\tau}\,, \\ X'_{j\tau} &= \alpha^X_\tau + \beta^X_\tau j + \epsilon^X_{j\tau} \end{aligned}$$

where $\alpha^Y_\tau = \{\overline{Y'}\}_\tau - \bar{j}\beta^Y_\tau$ (with $\{\overline{Y'}\}_\tau$ as the ensemble mean calculated from Eq. [3.1]) and $\beta^Y_\tau = \mathrm{Cov}\left[\{Y'\}_{j\tau}, j\right] / \mathrm{Var}\left[j\right]$ denote the intercept and slope coefficients for the decadal prediction, respectively, while $\alpha^X_\tau = \overline{X'}_\tau - \bar{j}\beta^X_\tau$ (with $\overline{X'}_\tau$ given by Eq. [3.2]) and $\beta^X_\tau = \mathrm{Cov}\left[X'_{j\tau}, j\right] / \mathrm{Var}\left[j\right]$ identify such coefficients for reference data. The higher-order errors are represented by $\epsilon^Y_{j\tau}$ and $\epsilon^X_{j\tau}$. Therefore, the model drift is given by

$$d^{\mathrm{TrDC}}_{j\tau} = \alpha^Y_\tau + \beta^Y_\tau j - (\alpha^X_\tau + \beta^X_\tau j) = d^{\mathrm{MDC}}_\tau + (\beta^Y_\tau - \beta^X_\tau)(j - \bar{j}) \qquad [8.1]$$

where d^{MDC}_τ is the model drift calculated in Eq. [3.37]. Note that if there are not differences between the slope coefficients, the trend-based drift is reduced to the mean drift. The corrected forecast for the ensemble member k and initial date j at lead time τ is calculated by removing the drift from $Y'_{kj\tau}$:

$$Y'^{\mathrm{TrDC}}_{kj\tau} = Y'_{kj\tau} - d^{\mathrm{TrDC}}_{j\tau} \qquad [8.2]$$

❦ *Initial condition-based drift correction*

As TrDC, the objective of the ICDC method (Fučkar et al., 2014) is to correct long-term climate trends. This method intends to take advantage of the relevance that an accurate representation of the climate state at the initialization stage has on the drift and predictive skill. Therefore, in this case, the independent variable in the first-order approximation is the climate state in the reference dataset $\eta_j = X'_{j,\tau=1}$ for the start date j. Note that η_j does not necessarily correspond to the same observational initial conditions used to initialize the decadal predictions. Instead, ERA5 provides the climate state at time of initialization used for η_j in all fields. The equations of the ensemble mean $\{Y'\}_{j\tau}$ and reference data $X'_{j\tau}$ can be written as follows:

$$\begin{aligned} \{Y'\}_{j\tau} &= \tilde{\alpha}^Y_\tau + \tilde{\beta}^Y_\tau \eta_j + \tilde{\epsilon}^Y_{j\tau} \\ X'_{j\tau} &= \tilde{\alpha}^X_\tau + \tilde{\beta}^X_\tau \eta_j + \tilde{\epsilon}^X_{j\tau} \end{aligned} \qquad [8.3]$$

where $\tilde{\alpha}^Y_\tau = \{\overline{Y'}\}_\tau - \overline{\eta}\tilde{\beta}^Y_\tau$ and $\tilde{\beta}^Y_\tau = \text{Cov}\left[\{Y'\}_{j\tau}, \eta_j\right] / \text{Var}\left[\eta_j\right]$ are the intercept and slope coefficients for the decadal forecast, $\tilde{\alpha}^X_\tau = \overline{X'}_\tau - \overline{\eta}\tilde{\beta}^X_\tau$ and $\tilde{\beta}^X_\tau = \text{Cov}\left[X'_{j\tau}, \eta_j\right] / \text{Var}\left[\eta_j\right]$ correspond to the reference data, and $\tilde{\epsilon}^Y_{j\tau}$ and $\tilde{\epsilon}^X_{j\tau}$ are the higher-order errors. Proceeding in the same way as in the TrDC case:

$$d^{\text{ICDC}}_{j\tau} = \tilde{\alpha}^Y_\tau + \tilde{\beta}^Y_\tau \eta_j - (\tilde{\alpha}^X_\tau + \tilde{\beta}^X_\tau \eta_j) = d^{\text{MDC}}_\tau + (\tilde{\beta}^Y_\tau - \tilde{\beta}^X_\tau)(\eta_j - \overline{\eta}) \,, \qquad [8.4]$$

where d^{MDC} is the model drift calculated in Eq. [3.37]. The corrected forecast for the ensemble member k, initial date j and at lead time τ is calculated by removing the drift from $Y'_{kj\tau}$:

$$Y'^{\text{ICDC}}_{kj\tau} = Y'_{kj\tau} - d^{\text{ICDC}}_{j\tau} \qquad [8.5]$$

8.2.2. *Complements for the conventional formulations*

❦ *K-nearest neighbours*

In addition to the conventional formulations of the methods described above, there are other approaches which can be used to complement the adjustment of the drift conducted through those methods. The kNN approach was proposed by Choudhury et al. (2017) and, in the same vein as ICDC, aims at taking advantage of the influence the initial state could have on the model drift (Fučkar et al., 2014). The purpose of this method is to improve the drift calculation by using only a selection of initialized experiments. The selected experiments are those whose initial conditions are the most similar to the observed initial state in the corrected decade. The procedure is the following:

1) Let j be the initialization date of the decadal experiment whose drift will be removed. Consider η_j as the initial observed state in j and $\eta_{i\neq j}$ as the initial observed state in the other decades.
2) Select a percentage of the closest $\eta_{i\neq j}$ values to η_j. As in Choudhury et al. (2017), the 60 % of $\eta_{i\neq j}$ values have been chosen.
3) These selected decades constitute the subset used to remove the model drift by using any of the conventional methods described above.

The corrected forecast for the ensemble member k, initial date j and at lead time τ is calculated by removing the drift from $Y'_{kj\tau}$:

$$Y'^{\text{kNN}}_{kj\tau} = Y'_{kj\tau} - d^{\text{kNN}}_{j\tau} \qquad [8.6]$$

❦ *Polynomial fitting*

The FIT approach (Gangstø et al., 2013; Kruschke et al., 2016) attempts to reduce the sampling uncertainty in drift correction and is built upon the idea of a non-monotonous lead time-dependent drift possibly existing in decadal experiments. The drift is given by

$$d^{\text{FIT}}_{j\tau} = g_{0,j\tau} + g_{1,j\tau}\tau + g_{2,j\tau}\tau^2 + g_{3,j\tau}\tau^3 \ , \qquad [8.7]$$

where $a_{i,j\tau}$ are non-stationary coefficients which change over time t (a function of j and τ). A first-order approximation is considered for them:

$$d^{\text{FIT}}_{j\tau} = (h_0 + h_1 t_{j\tau}) + (h_2 + h_3 t_{j\tau})\tau + (h_4 + h_5 t_{j\tau})\tau^2 + (h_6 + h_7 t_{j\tau})\tau^3 \qquad [8.8]$$

The coefficients in Eq. [8.8] are obtained by adjusting the polynomial to the drift calculated through a conventional procedure. The corrected forecast for the ensemble member k, initial date j and at lead time τ is calculated by removing the drift from $Y'_{kj\tau}$:

$$Y'^{\text{FIT}}_{kj\tau} = Y'_{kj\tau} - d^{\text{FIT}}_{j\tau} \qquad [8.9]$$

8.2.3. *Evaluation*

❦ *Evaluation of drift correction methods*

The first part of this study consists in evaluating each drift correction method to determine which gives the best results in terms of the accuracy of the predictions for SST, NSAT anomaly, SLP and several climate indices from ENS40. DCPs are skilful inasmuch as they can reproduce not only the observed climate variability, but also the magnitude of that change. Therefore, the evaluation has been conducted by using the RMSE and ACC metrics defined in Eqs. [3.10] and [3.13], respectively, both calculated with full fields. Several lead time windows spanning 1, 4 and 8 years have been considered in the analysis to evaluate the dependence of the accuracy of predictions on the lead time, as suggested by Goddard et al. (2013). This evaluation has been carried out for the experiments initialized every year from 1960 to 2009 (50 start dates).

After calculating both metrics for each grid point, the spatially weighted averages $\langle$RMSE$\rangle$ and $\langle$ACC$\rangle$ have been computed in each CORDEX domain considered here. The best qualified method per domain has been selected to build ENS3 afterwards. The non-parametric bootstrapping described in Section 3.2.3 has been applied to assess the statistical significance of the results. To maintain the coherence with the work presented in Rosa-Cánovas et al. (2023), the statistical significance has been assessed for the 95 % confidence level, rather than for the 90 % level considered in the previous chapters.

❦ *Climate indices*

In addition to the analysis of SST, NSAT anomaly and SLP, the ACC scores for several climate indices have been also used as a decision factor when assessing the performance of the drift correction methods. The climate indices include in the analysis the representation of large-scale patterns of climate variability which influence on local climate, allowing a broader assessment which is not only constrained to the CORDEX regions. Several El Niño/Southern Oscillation (ENSO) indices have been considered alongside the NAO and Atlantic Mutidecadal Variability (AMV) indices. The ENSO oceanic component is characterized by the emergence of SST anomalies across tropical Pacific and influence on the weather worldwide (e.g., Brönnimann et al., 2006; Infanti and Kirtman, 2016), although its effects are more perceptible in South America (Cai et al., 2020). Since ENSO SST patterns are spatially variant, various ENSO indices calculated over different areas have been considered in this study: the Niño 1+2, Niño 3, Niño 3.4, Niño 4 and Trans-Niño indices (Trenberth and Stepaniak, 2001).

On the other hand, as previously mentioned in Section 4.5, NAO is one of the most important atmospheric circulation modes in the Northern Hemisphere, described by changes in SLP or geopotential height over the action centers located in the Azores and Iceland, with influence on temperature, precipitation and winds along the whole hemisphere (Hurrell et al., 2003; Smith et al., 2019). Finally, AMV consists in a SST variability pattern over the North Atlantic, which has been associated to the formation of hurricanes or changes in rainfall in the Northern Hemisphere (Knight et al., 2006; Smith et al., 2020).

All indices have been calculated with averaged fields in DJF. The SST has been used to compute the AMV and ENSO indices, whereas the SLP has been used to obtain the NAO index. The regions where the indices have been calculated are described in Table 8.1. The Niño 1+2, Niño 3, Niño 3.4 and Niño 4 indices have been computed by calculating the anomalies of the spatially averaged SST in the corresponding region for each lead time series, while Trans-Niño Index (TNI) has been calculated as the standardized Niño 4 index minus the standardized Niño 1+2 index (Trenberth and Stepaniak, 2001). To calculate the NAO index, the anomalies of the spatially averaged SLP in the Iceland region have been subtracted to the anomalies of the spatial average in the Azores region, applying the definition used in other studies which assessed the

Table 8.1: Definition of the regions where the climate indices are calculated.

Index	**Region**
Niño 1+2	0°–10°S, 90°W–80°W
Niño 3	5°N–5°S, 150°W–90°W
Niño 3.4	5°N–5°S, 170°W–120°W
Niño 4	5°N–5°S, 160°E–150°W
Trans-Niño Index (TNI)	0°–10°S, 90°W–80°W (Niño 1+2 region) and 5°N–5°S, 160°E–150°W (Niño 4 region)
North Atlantic Oscillation (NAO)	36°N–40°N, 28°W–20°W (Azores) and 63°N–70°N, 25°W–16°W (Iceland)
Atlantic Multidecadal Variability (AMV)	0°–60°N, 80°W–0° (North Atlantic) and 60°S–60°N, 180°E–180°W (global average)

predictive skill of different DPSs for this variability pattern (Smith et al., 2019, 2020). Following the same approach of Trenberth and Shea (2006) and Smith et al. (2020), the AMV index has been calculated as the difference between the SST averaged in the North Atlantic region and the SST global average.

❦ *Evaluation of single members and subensemble predictive skill*

To select the CESM-DPLE members which have been used to build ENS3, the analysis has been focused on the ⟨ACC⟩ calculated for SST in lead years 2–9 over each domain separately, following a similar approach to that described in Section 3.3.2. This lead time period has been chosen to examine the performance at decadal time scale rather than accounting for the skill arising from interannual variability. Moreover, the skill which arises from seasonal to annual variability is avoided by removing the first year from the assessed period. This analysis has been centered on SST because of the particularly relevant role that the ocean component plays in the predictive skill of DCPs. As stated in Section 2.1.1, only 10 out of the 40 members of CESM-DPLE (ENS10) provide enough data to conduct DD simulations, so the member selection has been constrained to those suitable members.

ENS3 has been constructed with the member showing the largest skill in reproducing the observed SST variability (the "best" member), the member showing the lowest skill (the "worst" member) and a member with an intermediate behaviour. A similar approach was used by Paeth et al. (2017) to carry out their study about the decadal predictability of the West African monsoon. With this strategy, a representative subensemble of the whole ensemble can be constructed. By selecting members with heterogeneous skill levels, part of the spread of ENS40, or of ENS10 at least, is expected to be retained, since these members cover the whole range of possible single performances among the 10 members available for DD.

In the analysis of the ENS3 performance, the following key points have been addressed:

1) How much does the ⟨ACC⟩ for SST depend on ensemble size?
2) Does the confidence intervals of ⟨ACC⟩ for the subensemble contain the result obtained for ENS10 (the maximum ensemble size attainable by dynamically downscaled CESM-DPLE) and ENS40 (the maximum ensemble size attainable by CESM-DPLE)?
3) Is the spread of the members in the subensemble appropriate to quantify the uncertainty in the subensemble predictions?

By addressing the first question, the unavoidable loss of predictive skill, consequence of reducing the ensemble size to three members, can be quantified. At first, $\langle$RMSE$\rangle$ and $\langle$ACC$\rangle$ have been calculated for ensemble sizes ranging from 3 to 10. The confidence intervals of these metrics have been calculated by applying the same bootstrapping used in the evaluation of the correction methods, and described in Section 3.2.3, but considering only the 10 members available for DD. Secondly, the results have been compared with those for ENS3 and ENS40. To conduct the bootstrapping for these ensembles, 3 (the “best”, the “worst” and “intermediate”) and 40 members have been used, respectively.

The answer to the second question shows if the accuracy levels which can be potentially achieved by ENS10 and ENS40 are covered by the confidence intervals obtained for subensembles with different sizes. Again, the bootstrapping strategy has been applied for ensemble sizes ranging from 3 to 10, alongside our ENS3, to calculate $\langle$RMSE$\rangle$ and $\langle$ACC$\rangle$. For every subensemble, the bootstrapping has been repeated 5000 times in order to calculate the percentage of ENS10 and ENS40 score coverage that the confidence intervals get. This methodology was also used by Sienz et al. (2016) to examine the skill of small subensembles, but considering a conceptual model to calculate the skill score which have to be covered by the confidence intervals.

The third question tests if the subensemble spread is appropriate to represent the range of possible individual predictions over time, i.e., how reliable the subensemble predictions are. Following Goddard et al. (2013), the reliability of decadal predictions can be analyzed through the CRPSS, whose calculation and interpretation has been detailed in Section 3.2.2. The optimal value of CRPSS in Eq. [3.29] is CRPSS = 0. It is attained for $\overline{\sigma_Y^2} = \sigma_X^2$ (from Eqs. [3.32] and [3.33], respectively), when the prediction and reference distributions are equal and, therefore, the average ensemble spread is adequate to quantify the uncertainty. The statistical significance of the results obtained for subensembles of different sizes from 3 to 10 members has been assessed by the bootstrapping approach, applied only for start dates, while the members of the subensembles have been randomly selected by combinations without repetitions of 3 members for ENS3 and 10 members for ensembles sizes from 3 to 10. For ENS40, all members of the ensemble have been considered.

8.3. Results

8.3.1. *Evaluation of the drift correction methods*

This Section is devoted to the analysis of the added value of the drift correction methods to the accuracy level of predictions for SST, NSAT anomaly, SLP and several climate indices for ENS40.

❦ *European domain*

The spatially averaged scores for SST over the EUR domain per drift correction method along lead time have been depicted in Figure 8.2. Crosses denote averages, whereas the median of the sample average is represented by a straight line, boxes identify the 50 % confidence interval and whiskers correspond to the 95 % confidence interval. All methods have performed very well in terms of $\langle$RMSE$\rangle$ and none has substantially achieved better results than the others. The highest $\langle$RMSE$\rangle$ values have been found during the first year. $\langle$RMSE$\rangle$ for uncorrected data (RAW) is around 0.76 K, whereas values below 0.45 K are depicted when a drift correction method is applied. In lead years 2–5 and 6–9, a decrement in $\langle$RMSE$\rangle$ can be observed, while the lowest values are shown at decadal scale. In the results for $\langle$ACC$\rangle$, ICDC-like methods have outperformed almost all other methods in lead year 1, with the exception of MDC$_{\text{kNN}}$, which has obtained similar results to those for ICDC$_{\text{FIT}}$. Among ICDC-like methods, the best results have been obtained by ICDC, followed by ICDC$_{\text{kNN}}$ and ICDC$_{\text{FIT}}$. At this time scale, the initialization fingerprint is more prominent than afterwards, so techniques which incorporate information about initial conditions in drift correction are candidates to perform the best. On the other hand, apart from the ICDC-like methods, only MDC$_{\text{kNN}}$ and TrDC$_{\text{kNN}}$ give slightly better results than RAW. Although MDC is not expected to significantly improve or worsen RAW performance in terms of $\langle$ACC$\rangle$, as explained in Section 3.3.2, it is so for TrDC. Nevertheless, the introduction of the trend component leads to slightly poorer scores compared to RAW and MDC. According to Eq. [8.1], the drift in TrDC is calculated as the drift in MDC added to a component which accounts for the differences in trends between model and reference data. Bearing in mind how MDC affects results for $\langle$ACC$\rangle$, the slight decrease in the skill after applying TrDC is entirely caused by this trend component and may be due to minor discrepancies between ERA5 (reference dataset in drift correction) and ERSST5 (reference dataset in skill evaluation) SST trends. At interannual scale, the differences between methods are smaller than in lead year 1. ICDC-like and MDC-like techniques perform very similar to RAW, with

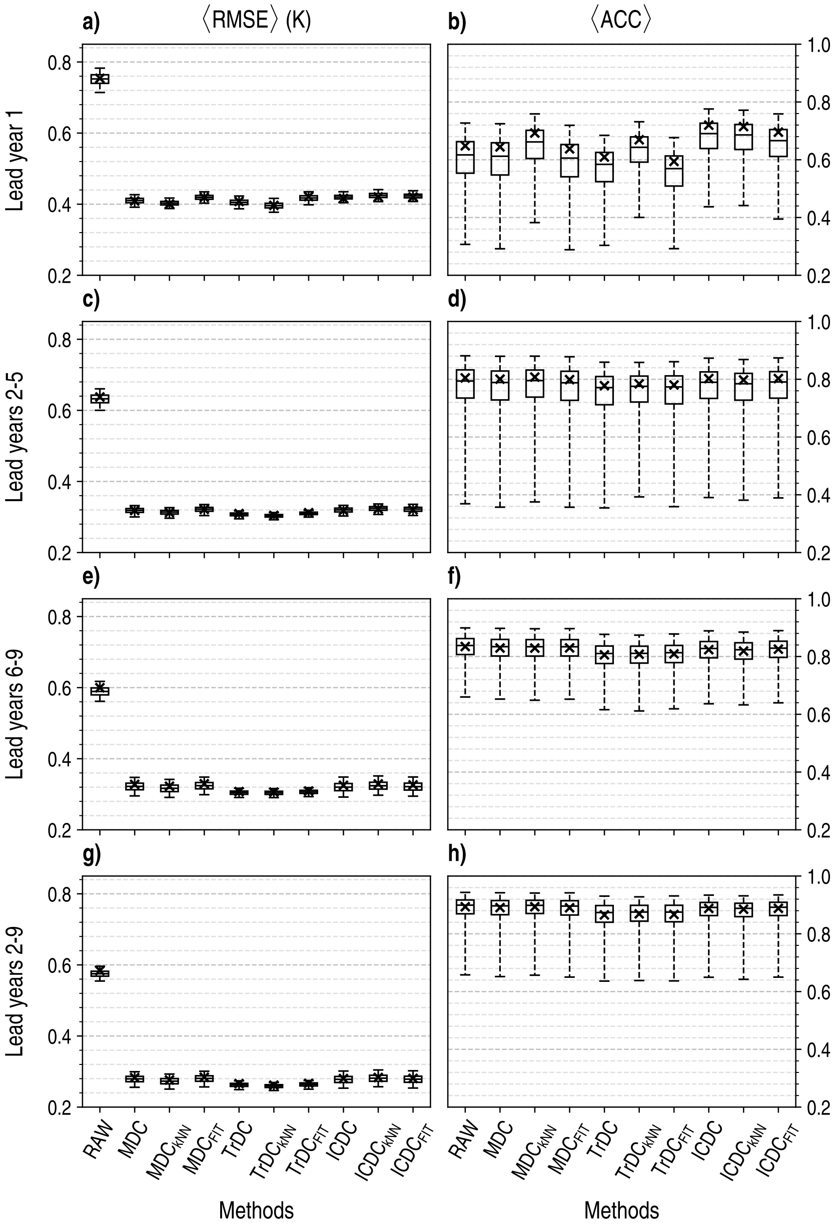

Figure 8.2: Spatially averaged RMSE (⟨RMSE⟩, left column) and ACC (⟨ACC⟩, right column) for the ENS40 SST in lead years 1, 2–5, 6–9 and 2–9 (rows) over the EUR domain. The results are presented for each drift correction method and the uncorrected (raw) data. Crosses denote the spatial averages. Box plots show the results of a bootstrapping (see Section 8.2.3) for which lines indicate the median value and boxes and whiskers enclose the confidence intervals at the 50 % and 95 % levels, respectively.

averages around 0.8 in lead years 2–5 and 0.84 in lead years 6–9. TrDC-like methods give slightly lower ⟨ACC⟩ scores, although the differences are smaller than 0.04. In lead years 2–9, the situation is similar, with ⟨ACC⟩ scores near 0.9 for RAW, MDC-like and ICDC-like methods.

The results for the NSAT anomaly have been depicted in Figure B.74 (available in Appendix B.4). The correction methods do not generally provide an added value to the accuracy, except for lead year 1. For the rest of lead times, the scores for the methods and RAW are very similar. Since the field corrected is an anomaly, the mean field along the analyzed period in the reference dataset for drift correction (ERA5) and hindcasts is the same. Therefore, very similar ⟨RMSE⟩ scores have been found between correction methods and RAW. On the other hand, the ⟨ACC⟩ scores for RAW are very high at interannual and decadal scales, so the added value of correction methods is reduced for this variable. In the analysis of SLP, whose results are depicted in Figure B.75 (also in Appendix B.4), ICDC-like methods clearly give the best results for lead year 1 in terms of ⟨ACC⟩, with correlations around 0.4 for ICDC and ICDC_{kNN}, although they are very close to zero at the other lead times. At interannual and decadal scale, the best results for this score have been found for TrDC-like methods. These scores are below 0.2 at almost all lead times with the exception of lead years 2–9, when spatial averages between 0.2 and 0.28 can be found. Since the climate change trend is stronger for temperature fields than for SLP, a strong decrease in correlation was also expected. On the other hand, the results obtained for ⟨RMSE⟩ are very similar among the different methods.

In Europe, it is also interesting to evaluate how well these techniques perform in the prediction of the NAO and AMV indices. Their corresponding ACC scores, for each drift correction method and several lead times (interannual and decadal), are shown in Table 8.2. With respect to NAO, the best results have been obtained for the TrDC-like methods. At interannual scale, the ability to predict the NAO is higher in the first half of the decade. The three TrDC-like variants show statistically significant positive results in lead years 1–4, with a maximum value of 0.53 for TrDC_{kNN}. In lead years 2–5, the significant results have been found for TrDC_{FIT} and TrDC_{kNN}, with the latter showing the highest score (ACC = 0.49). The correlations are not significant in lead years 5–8, with values equal to 0.35 and 0.39 for TrDC and TrDC_{kNN}, respectively. The strongest ACC values have been found in lead years 2–9, with significant outcomes of 0.67, 0.68 and 0.51 for TrDC, TrDC_{kNN} and TrDC_{FIT}, respectively. The other methods do not give significant results at any lead time. Indeed, the results are often negative (but not significant) in the second half of the decade for such techniques. Even RAW

Table 8.2: ACC calculated for several climate indices from ENS40 along lead time for each drift correction method and the uncorrected (raw) data. The bold formatting in the ACC values indicates statistical significance at the 95 % confidence level.

Index	**Method**	**Lead years**			
		1-4	**2-5**	**5-8**	**2-9**
NAO	RAW	**0.36**	**0.38**	0.01	0.28
	MDC	0.25	0.23	-0.20	0.13
	MDC$_{\text{kNN}}$	0.32	0.30	-0.06	0.33
	MDC$_{\text{FIT}}$	0.33	0.35	-0.05	0.25
	TrDC	**0.46**	0.44	0.35	**0.67**
	TrDC$_{\text{kNN}}$	**0.53**	**0.49**	0.39	**0.68**
	TrDC$_{\text{FIT}}$	**0.45**	**0.41**	0.17	**0.51**
	ICDC	0.16	0.10	-0.17	0.09
	ICDC$_{\text{kNN}}$	0.27	0.24	-0.18	0.30
	ICDC$_{\text{FIT}}$	0.31	0.30	-0.06	0.20
AMV	RAW	**0.72**	**0.83**	**0.84**	**0.91**
	MDC	**0.70**	**0.81**	**0.82**	**0.91**
	MDC$_{\text{kNN}}$	**0.73**	**0.83**	**0.80**	**0.90**
	MDC$_{\text{FIT}}$	**0.71**	**0.82**	**0.82**	**0.91**
	TrDC	**0.68**	**0.78**	**0.80**	**0.89**
	TrDC$_{\text{kNN}}$	**0.73**	**0.81**	**0.79**	**0.89**
	TrDC$_{\text{FIT}}$	**0.68**	**0.77**	**0.81**	**0.89**
	ICDC	**0.76**	**0.83**	**0.75**	**0.89**
	ICDC$_{\text{kNN}}$	**0.76**	**0.83**	**0.74**	**0.89**
	ICDC$_{\text{FIT}}$	**0.76**	**0.83**	**0.77**	**0.89**
TNI	RAW	**0.41**	0.14	-0.13	-0.07
	MDC	**0.36**	0.06	-0.23	-0.13
	MDC$_{\text{kNN}}$	**0.35**	0.03	0.10	0.00
	MDC$_{\text{FIT}}$	**0.36**	0.06	-0.22	-0.13
	TrDC	**0.41**	0.20	-0.10	0.02
	TrDC$_{\text{kNN}}$	**0.37**	0.14	0.16	0.10
	TrDC$_{\text{FIT}}$	**0.39**	0.12	-0.15	-0.07
	ICDC	**0.44**	0.03	0.11	0.04
	ICDC$_{\text{kNN}}$	**0.39**	-0.05	**0.19**	0.05
	ICDC$_{\text{FIT}}$	**0.41**	0.03	-0.02	-0.05

often performs better than them, especially in lead years 1–4 and 2–5, with significant positive results of 0.36 and 0.38, respectively.

The results obtained for TrDC and TrDC$_{\mathrm{kNN}}$ at decadal scale are certainly promising. Smith et al. (2019) found an ACC of 0.49 with a multimodel ensemble composed of 71 members (31 members more than CESM-DPLE) without any post-processing adjustment for the same lead time. On the other hand, an ACC of 0.79 was found by Smith et al. (2020) for an even higher multimodel ensemble, composed of 169 members, after post-processing the NAO time series to rescale the signal. The same ensemble with no post-processing gave a result of 0.48. In this context, TrDC$_{\mathrm{kNN}}$ and TrDC methods constitute a relatively simple approach to improve the ability to predict the NAO variability for the 40-member CESM-DPLE ensemble. The contrast between their high ACC scores for NAO and the low $\langle$ACC$\rangle$ scores for SLP is due to the fact that $\langle$ACC$\rangle$ for SLP has been calculated with annual averages, whereas ACC for NAO has been calculated in DJF, when results for SLP are slightly more optimistic in lead years 2–9 (see Figure B.76 in Appendix B.4).

Much higher correlations have been obtained in the analysis of AMV, as expected given that this index entirely depends on SST. In general, ACC is above 0.70 for almost all methods and lead times, with the exception of TrDC and TrDC$_{\mathrm{kNN}}$ in lead years 1–4. In this case, the MDC-like methods perform slightly better than the others, reaching results equal to or larger than 0.9 in lead years 2–9. They are followed by TrDC-like and ICDC-like methods, in that order. In the evaluation of the prediction for AMV, there is no added value over what has been obtained by RAW.

Since the differences between the performance of the methods in the prediction for SST and NSAT anomaly are very small, the selection of the most appropriate technique for the EUR domain has been done in terms of the results obtained for SLP and NAO. Although ICDC-like methods can provide an added value to the accuracy of predictions in lead year 1 in the analysis of SLP, the TrDC-like methods provide the best results at decadal scale, especially TrDC and TrDC$_{\mathrm{kNN}}$ for NAO. Since the latter method slightly outperforms the former at all lead times in the analysis of SLP, TrDC$_{\mathrm{kNN}}$ may be the preferred choice in this area.

❦ *South American domain*

The results for the averaged scores over the SA domain for SST have been depicted in Figure 8.3. As for the EUR domain, drift correction always improves $\langle$RMSE$\rangle$ compared to RAW, regardless of the correction method. In lead year 1, the maximum

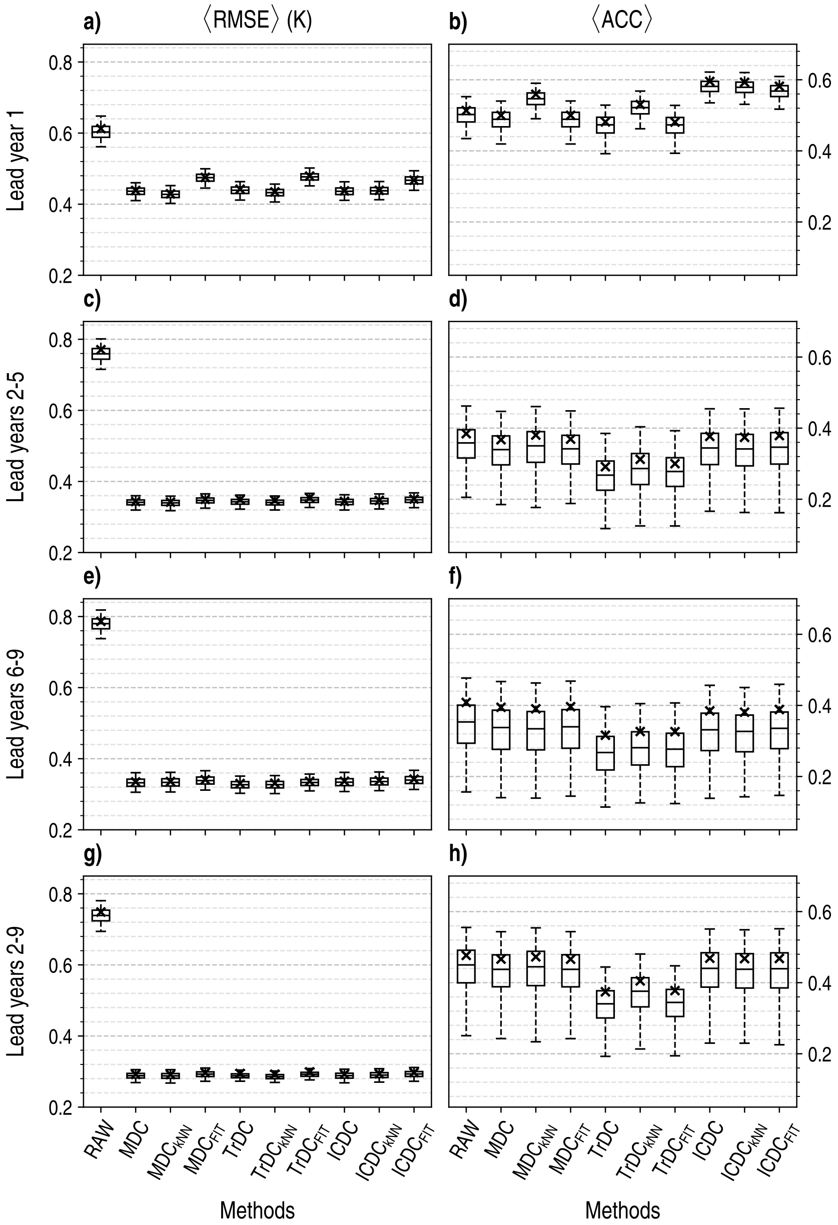

Figure 8.3: As Figure 8.2 but for the SA domain.

improvements are around 0.24 K, with techniques using polynomial fitting giving the highest ⟨RMSE⟩. Differences among methods are less evident at the other lead times, when reductions in ⟨RMSE⟩ from approximately 0.80 K to 0.30 K are found, compared to RAW. Relevant disparities among method performances are observed for ⟨ACC⟩ scores. As expected, ICDC-like procedures contribute to better capturing the climate variability in lead year 1, with outcomes around 0.6. MDC-like and TrDC-like methods achieve similar results among them, with ⟨ACC⟩ ranging from 0.48 to 0.56 and with kNN approaches getting higher scores than the others. In lead years 2–5 and 6–9, the ⟨ACC⟩ scores are lower than in lead year 1. At this lead time, MDC-like and ICDC-like methods perform similar to each other, with averaged scores between 0.36 and 0.4, getting better results than TrDC-like methods. Additionally, the values for TrDC-like methods are certainly lower than for RAW, as happened for the EUR domain. The situation is similar at decadal scale, with higher ⟨ACC⟩ scores than at interannual scale but lower than for the first year.

In the analysis of the accuracy of predictions for the NSAT anomaly (Figure B.77 in Appendix B.4), all methods give outcomes similar to RAW in terms of ⟨RMSE⟩. Maximum values around 0.5 K in lead year 1 and minimum around 0.22 K in lead years 2–9 are observed. The performance among methods in terms of ⟨ACC⟩ is also very similar for all lead times, excepting in lead year 1, when ICDC-like methods get the highest correlations around 0.52. At the other lead times, the correlations are higher but the differences among methods are hardly appreciable. After examining the results for SLP (Figure B.78 in Appendix B.4), similar conclusions have been obtained, although ICDC-like methods perform slightly worse than the others in lead years 6–9 and 2–9. Nevertheless, the correlations are poor in general for all methods and lower than for temperature variables, as happened in the EUR domain. The only exception is found for ICDC and ICDC$_{\mathrm{kNN}}$ in lead year 1, with ACC values around 0.4.

The skill to capture the variability of some ENSO indices, described in Section 8.2.3, has also been assessed for each drift correction method and several lead times, being the most relevant results, obtained for TNI, shown in Table 8.2. In general, for the Niño 1+2, 3, 3-4 and 4 indices, there is a lack of statistical significance in the ACC values found for all methods at almost all lead times. This also happens for TNI, with the exception of the significant positive results found in lead year 1 for all methods and in lead years 5–8 only for ICDC$_{\mathrm{kNN}}$. While the results are commonly positive at the beginning of the decade, there is a decrease of ACC along lead time for all ENSO indices, although part of the statistical significance is retained by ICDC$_{\mathrm{kNN}}$ for TNI in the second half of the decade. The ICDC-like methods perform slightly better than their counterparts in lead years 1–4 and 5–8. The decrease of ACC for all ENSO indices during the first years of the decade is consistent with the results obtained by Gonzalez and Goddard (2016) for the Niño 3.4 index. The authors attribute the differences between the modeled and observed ENSO, in part, to biases in the location of the maximum SST anomalies, which in turn affect the location of the coupling between SST and the atmosphere. The spatial distribution of ACC for the ENS40 SST, drift-corrected with ICDC$_{\mathrm{kNN}}$ in DJF, has been depicted in Figure B.79 (available in Appendix B.4). The ACC outcomes are very low in the Pacific, with not significant results over large areas inside the ENSO regions, partly explaining the correlations observed for these indices.

The added value of the ICDC-like methods to the accuracy of predictions in lead year 1 is higher than that of other methods for SLP and, to a lesser extent, for SST. In the analysis of the ENSO indices, their performance is, in general, slightly better than that provided by the other methods, especially with respect to the TNI. For these reasons, the ICDC-like methods may be the most suitable correction techniques for the SA domain. Since the ICDC$_{\mathrm{kNN}}$ is the only method which retains some significant skill for TNI in the second half of the decade, it has been chosen for the second part of the analysis in Section 8.3.2. If a more straightforward and less computationally demanding technique is required, the ICDC method may also be a good option, as it additionally performs slightly better than ICDC$_{\mathrm{kNN}}$ in terms of the TNI at the beginning of the decade.

❦ *North American domain*

The spatially averaged scores for SST in the NA domain have been depicted in Figure 8.4. The performance in terms of ⟨RMSE⟩ is similar to that within the other domains in general. All methods significantly contribute to reducing the bias in RAW and there are not big differences among them. ⟨RMSE⟩ for corrected data is higher in lead year 1 with values around 0.46 K, whereas the lowest values are found in lead years 2–9, ranging from 0.32 K to 0.36 K, approximately. Again, the analysis of the ⟨ACC⟩ scores is needed to choose the most appropriate correction procedure for this region. In lead year 1, ICDC-like methods outperform the rest, with values near 0.68. There are not robust differences between MDC-like and TrDC-like methods, although kNN gives slightly better results than the conventional and polynomial approaches. In lead years 2–5, the gap between ICDC-like and the other methods is smaller, showing a performance similar to RAW and MDC-like methods, with spatial averages between 0.44 and 0.48. In lead years 6–9, the differences among methods are hardly appreciable. All perform similar to RAW, with ⟨ACC⟩ between 0.52 and

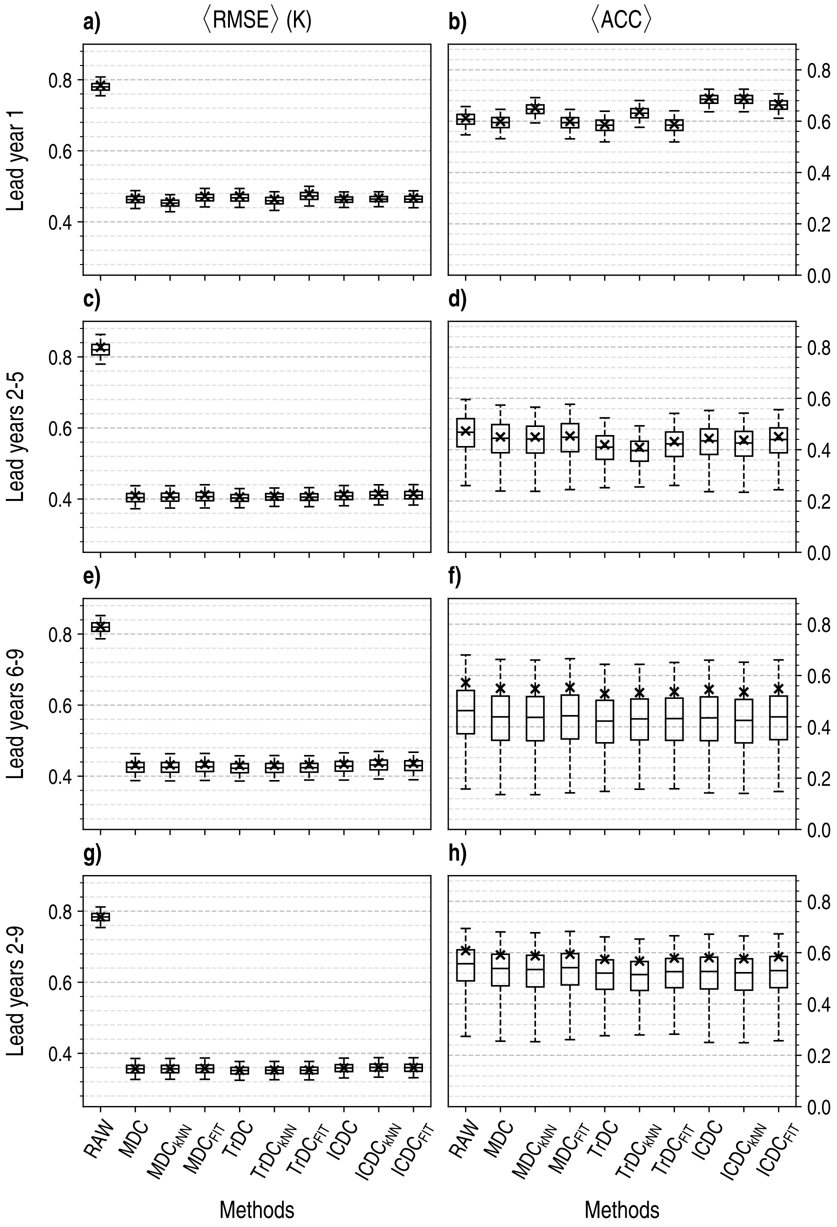

Figure 8.4: As Figure 8.2 but for the NA domain.

0.56. The situation is the same in lead years 2–9, but with higher scores close to 0.6.

In the analysis of the NSAT anomaly (Figure B.80 in Appendix B.4), the results are similar to those for the EUR and SA domains. The methods generally give the same outcomes for $\langle$RMSE$\rangle$, although ICDC-like techniques perform slightly worse than the others in lead year 1. While the highest errors have been found at this lead time, generally between 1 K and 1.04 K, the lowest are around 0.35 K at decadal scale. In regard to $\langle$ACC$\rangle$, ICDC-like methods slightly outperform the others again in lead year 1, with correlations between 0.4 and 0.48, but the differences among them are smaller at the other lead times. The highest $\langle$ACC$\rangle$ scores are shown in lead years 2–9, with values varying from 0.8 to 0.84 depending on the method. In terms of SLP (Figure B.81 in Appendix B.4), ICDC-like methods give the best results for $\langle$ACC$\rangle$ in lead year 1, when the largest scores around 0.46 are observed. At the other lead times, the results are very similar among methods. The TrDC$_{\text{kNN}}$ method performs slightly better than the others and provides a small added value over RAW, although the outcomes are always below 0.1. On the other hand, the lowest correlations have been found at decadal scale, with scores around 0.2. The other methods do not contribute to the improvement of RAW performance at any lead time.

Although ICDC-like methods can provide an added value to the accuracy of predictions over the other methods and RAW in lead year 1 for all variables (especially for SLP), the method performances are very similar to each other at the other lead times for the three variables. The good results obtained in lead years 2–9 for the NAO index lead to consider the TrDC$_{\text{kNN}}$ method as one of the most suitable correction techniques to adjust the drift with focus on the decadal scale, just as for the EUR domain.

Finally, the RMSE and ACC spatial distributions for SST, NSAT anomaly and SLP in lead years 2–9, with TrDC$_{\text{kNN}}$ and ICDC$_{\text{kNN}}$ used as the drift correction methods, as well as for RAW, can be consulted in Figures B.82 to B.84, respectively (available in Appendix B.4). The results obtained for the three domains considered in the analysis reflect what has been found for the spatial averages. In terms of RMSE, the correction methods help to improve the accuracy of the RAW experiments in all domains. On the other hand, TrDC$_{\text{kNN}}$ improves the spatial distribution of ACC compared to RAW in the EUR domain (compare Figures B.82f and B.84f). In the same line, this method also improves the correlations obtained in the northern and eastern parts of North America (compare Figures B.82f and B.84f again), contributing to slightly increasing the averaged ACC in the NA domain over RAW (see Figure 8.4h). The ICDC$_{\text{kNN}}$ method also provides an added value to the ACC in the subtropical latitudes close to the SA domain in the Pacific over RAW, increasing the statistical significance of the positive ACC scores (compare Figures B.83b and B.84b).

8.3.2. *Evaluation of single members and subensemble skill*

❦ *Selection of single members to build ENS3*

The predictive skill of the individual members usually varies depending on the field and domain, the lead time and even the skill score under analysis, so the selection of the members to build ENS3 is not straightforward. Since the ocean is the main reservoir of memory in the climate system, the selection has been done by considering the results obtained in terms of the ⟨ACC⟩ for SST in lead years 2–9. Figure 8.5 shows the ⟨ACC⟩ scores for SST in lead years 2–9 over the EUR, SA and NA domains for ENS40, ENS3 and the 10 single members available for DD. While SST has been corrected with TrDC_{kNN} in the EUR and NA domains, ICDC_{kNN} has been used in the SA domain. As stated in Section 8.2.3, the member with the highest skill (the “best” member), the member with the lowest skill (the “worst” member) and another member with an intermediate behaviour (“intermediate” member) have been chosen to build ENS3 in each domain.

In the EUR domain, ENS3 encompasses member 2 (the “best” member), member 8 (the “worst” member) and member 7 (“intermediate” member). The results for the single members range from 0.74 to 0.82, approximately. The highest correlations are

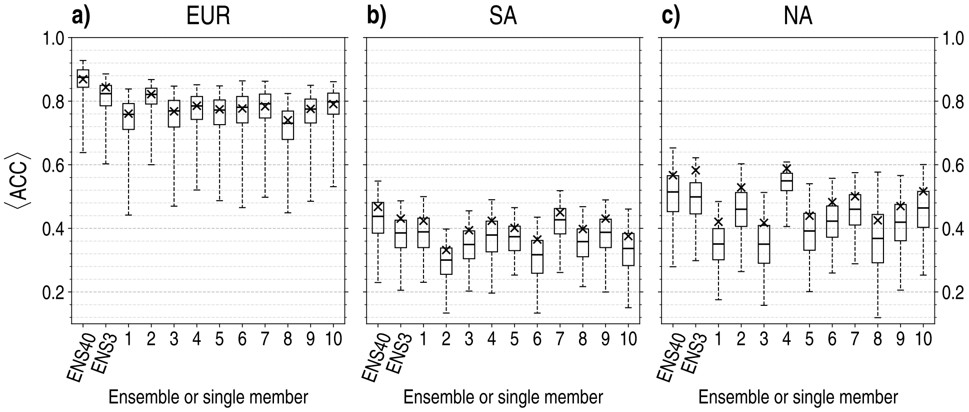

Figure 8.5: Spatially averaged ACC (⟨ACC⟩) for SST in lead years 2–9 over the **a**) EUR domain, **b**) SA domain and **c**) NA domain. In the EUR and NA domains, SST has been corrected with TrDC_{kNN}, whereas ICDC_{kNN} has been used in the SA domain. The results are depicted for ENS40, ENS3 and each single member. Crosses denote the spatial averages. Box plots show the results of a bootstrapping (see Section 8.2.3) for which lines indicate the median value and boxes and whiskers enclose the confidence intervals at the 50 % and 95 % levels, respectively.

observed for ENS40 and ENS3, with values around 0.88 and 0.84, respectively. Only a difference of nearly 0.02 is observed between ENS3 and member 2. With respect to the SA domain, member 7 has been chosen as the "best" member, member 2 as the "worst" and member 3 representing an intermediate behaviour. The differences in terms of accuracy are slightly larger than for the EUR domain, with averaged correlations ranging from about 0.32 to 0.44, approximately. Member 7 performs sligthly better than ENS3, getting an ⟨ACC⟩ around 0.44, whereas ENS40 shows a value close to 0.48. Finally, member 4 has been chosen as the "best" member, member 3 as the "worst" and member 10 as a member with intermediate skill in the NA domain. Likewise in the SA domain, slightly larger differences between single members are observed compared to the EUR domain, with averages scores ranging from 0.4 to 0.58, approximately. In this case, ENS40 is outperformed by ENS3, which shows an ⟨ACC⟩ around 0.58. For this domain, member 4 has obtained similar results to those for ENS3 and, therefore, better than those for ENS40.

❦ *Dependence of SST predictive skill on ensemble size*

The reduction of the ensemble size leads to an inevitably loss of predictive skill compared to the full ensemble. However, given the huge amount of computing resources needed to dynamically downscale an ensemble of decadal predictions, some concessions may be made in relation to the number of members considered for such simulations. In this Section, the impact of ensemble size on the accuracy of predictions for SST has been addressed. Figure 8.6 depicts how ⟨RMSE⟩ and ⟨ACC⟩ for SST vary with ensemble sizes from 3 to 10 members over the EUR, SA and NA domains in lead years 2–9. The ensemble size 3 with the symbol "*" represents the ENS3 selected above, while ensemble sizes without that symbol represent subensembles built with members randomly selected (see Section 8.2.3). The results for ENS40, which show the potential accuracy that can be achieved by CESM-DPLE, have also been included in the plots for comparison purposes, although only 10 CESM-DPLE members are available for DD. In addition, the total number of years which must be simulated depending on the ensemble size to generate the set of decadal experiments has also been included in the plots. This number has been calculated as:

$$N_{\text{sim}} = N_y \times N_d \times N_{\text{ens}} = 10 \text{ years/date} \times 50 \text{ dates} \times N_{\text{ens}} = 500 \text{ years} \times N_{\text{ens}}$$

where N_y = 10 years/date is the number of years in a decadal experiment, N_d = 50 dates is the number of initial dates and N_{ens} is the ensemble size.

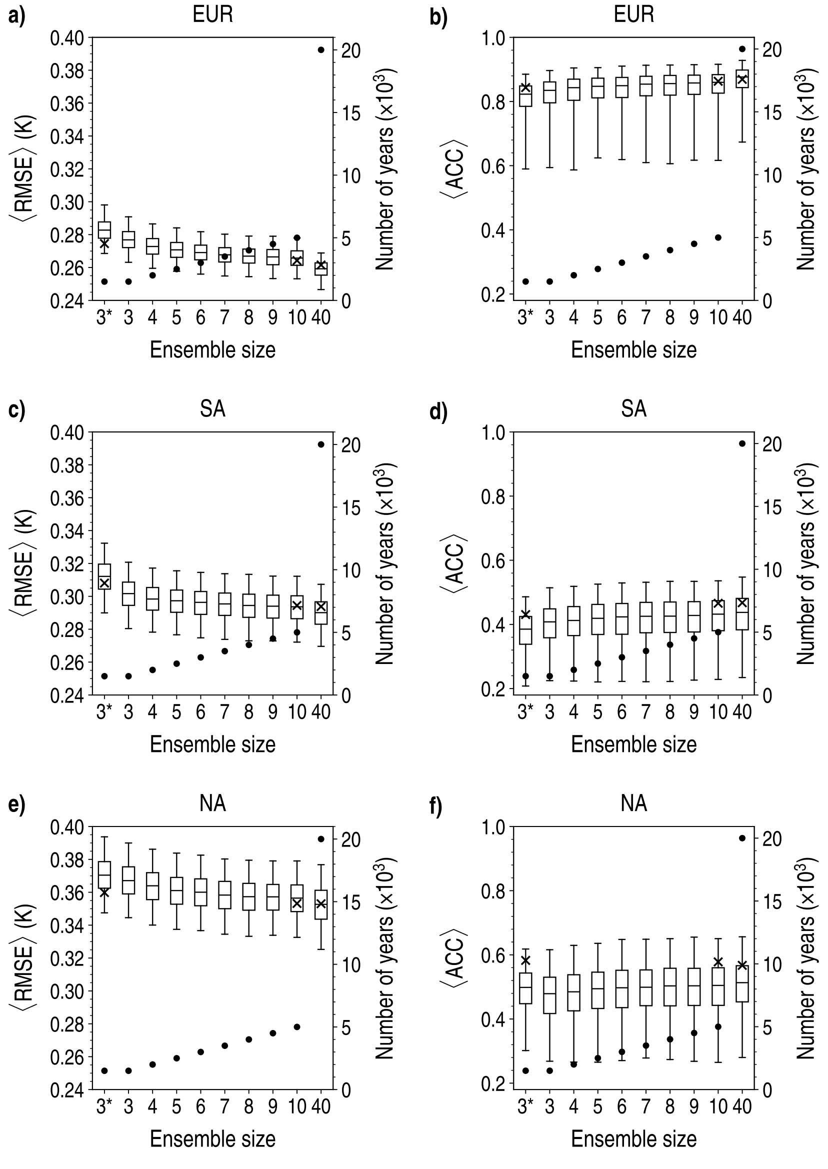

Figure 8.6: On the left axis, the dependence of the spatially averaged RMSE (⟨RMSE⟩, left column) and ACC (⟨ACC⟩, right column) for SST on the ensemble size over the EUR, SA and NA domains is represented. While SST has been corrected with TrDC$_{\text{kNN}}$ in the EUR and NA domains, ICDC$_{\text{kNN}}$ has been used in the SA domain. Crosses denote the spatial averages for ENS3, ENS10 and ENS40. Box plots show the results of a bootstrapping (see Section 8.2.3) for which lines denote the median value and boxes and whiskers enclose the confidence intervals at the 50 % and 95 % levels, respectively. On the right axis, the number of years to be simulated per ensemble size is represented by dots. The ensemble size 3 with the symbol "*" denotes the manually selected ENS3.

In general, there is an improvement of the accuracy with the increase of the ensemble size, as expected. There is not any stabilisation in the median of the scores for an ensemble size of up to 10 members. The median scores in the ensemble size of 40 members, the maximum attainable for CESM-DPLE, get always the best results. However, these differences in ⟨RMSE⟩ and ⟨ACC⟩ between different ensemble sizes are very low, and the width of the confidence intervals does not decrease by increasing the ensemble size. The largest disparities are observed between ENS3 and ENS40, as expected, but they do not surpass 0.03 K for ⟨RMSE⟩ and 0.1 for ⟨ACC⟩ for any domain. When comparing the averaged scores for ENS3 and ENS10, the differences are also very small. There is a large contrast between the gain of accuracy by increasing the ensemble size and the increase of the number of years which have to be simulated. For example, the added value to ⟨ACC⟩ of using ENS10 instead of ENS3 is about 0.02 and 0.04 for the EUR and SA domains, respectively, whereas there is a hardly appreciable better performance for ENS3 in the NA domain. On the other hand, the number of years to be simulated is multiplied by 3.33. These findings show that a large investment in computing resources is needed to only get small improvements, if any, in the accuracy of SST. Similar outcomes were found by Sienz et al. (2016) in correlations for the North Atlantic SST in lead years 2–5. In the cited study, the added value of increasing the ensemble size was higher for Central Europe summer temperature, where improvements around 0.1 were found between a 10-member and a 3-member ensembles. Reyers et al. (2019), using dynamically downscaled decadal predictions over Europe, also analyzed the dependence of the accuracy of predictions on ensemble size. They found a small improvement in correlation for air temperature in Europe by increasing the ensemble size from 3 to 10 members with a value near 0.03 in lead years 1–5. However, the improvement in correlation for PR because of increasing the ensemble size were much larger, with differences of almost 0.4 between 10-member and 3-member ensembles. The added value over uninitialized simulations was observed in the case of precipitation only for ensemble sizes equal to 7 members and above, whereas an ensemble of 3 members was enough for temperature.

The performance of ENS3 in comparison to a 3-member subensemble randomly selected can also be analyzed by examining Figure 8.6. The median values for ENS3 are above and below those for a random 3-member subensemble in ⟨RMSE⟩ and ⟨ACC⟩ plots, respectively, in the three domains. This can be explained not only by the fact that ENS3 encompasses the member with the lowest skill, but also by the number of members considered in the bootstraping (3 for ENS3 and 10 for the 3-member subensemble). However, the subensemble averaged scores for ENS3 outperforms almost always the median scores of the random subensemble. In other words, this simple member selection carried out to build ENS3 performs better than at least the 50 % of the random 3-member subensembles generated for all domains, except for ⟨RMSE⟩ in the SA domain.

❦ *Dependence of skill score coverage on ensemble size*

The bootstrapping conducted above has been repeated 5000 times in order to get the coverage percentage of ENS10 and ENS40 scores by the confidence intervals of subensembles with different sizes. As can be seen in Figure 8.7, the confidence intervals get a 100 % coverage of ENS10 and ENS40 scores in almost all occasions, with the exception of the ENS3 and the random 3-member ensemble for ⟨RMSE⟩ in the EUR domain. In the case of ENS3, this is partially due to the fact that the

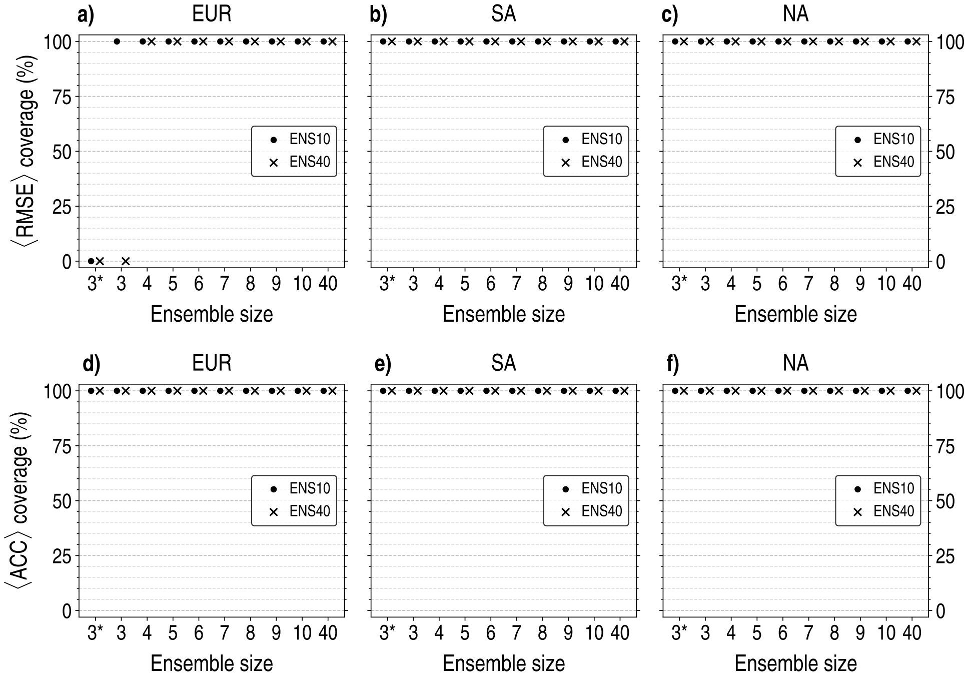

Figure 8.7: Percentage of score coverage by confidence intervals for different ensemble sizes. This scores have been calculated for the CESM-DPLE SST. While SST has been corrected with TrDC_{kNN} in the EUR and NA domains, ICDC_{kNN} has been used in the SA domain. Dots correspond to the coverage of the ENS10 scores, whereas the results found for ENS40 are denoted by crosses. The ensemble size 3 with the symbol "*" represents the manually selected ENS3.

bootstrapping has been conducted only for 3 members, as opposed to the 10 members included in the process for the randomly selected subensembles. In consequence, there is a slight offset of the ENS3 median and confidence interval limits compared to those for the random 3-member subensemble, as mentioned in the previous section. When the average scores of ENS10 or ENS40 are very close to the upper boundary of the ENS3 confidence intervals in Figure 8.6, the probability of that those averages are not covered by the ENS3 confidence intervals in a bootstrapping iteration is very large, so a 0 % coverage has been obtained for ENS10 and ENS40 ⟨RMSE⟩. It also occurs in the case of the ENS10 ⟨RMSE⟩ for the random 3-member ensemble. Nevertheless, there is a full score coverage by both confidence intervals in terms of ⟨ACC⟩. For the other domains, the confidence intervals of the scores calculated for ENS3 contain the results of ENS10 and ENS40 with a coverage percentage of 100 %. For ensemble sizes equal to 4 or higher, the score coverage is always of 100 % for the three domains.

❦ *Dependence of CRPSS on ensemble size*

The results for the spatially averaged CRPSS of SST in lead years 2–9 in the EUR, SA and NA domains have been depicted in Figure 8.8. The best results are found for the EUR domain, where the closest median values to zero are shown, ranging from -0.084 to -0.078. On the other hand, the most pessimistic values are found for the SA domain, where the median values are enclosed in the interval from -0.126 to -0.12. There is not a pronounced dependence of ⟨CRPSS⟩ on ensemble size in any domain. The median ⟨CRPSS⟩ of ENS3 is always above the median of the randomly selected 3-member subensemble. Its value in the EUR domain, about -0.08, is better than that for ENS10 and comparable to that for ENS40. In the SA domain, the ⟨CRPSS⟩ of ENS3 is lower than that of ENS10 but higher than that of ENS40, while ENS3 performs better than ENS10 and ENS40 in the NA domain. Nevertheless, as in the EUR domain, there are no large differences between the performances of ENS3, ENS10 and ENS40 in the other domains. Finally, even in the case of the EUR domain, the ⟨CRPSS⟩ values are significantly negative for all ensemble sizes at the 95 % level (note that 95 % confidence intervals are completely below 0) and, as consequence, different from 0. As mentioned in Section 3.2.2, the optimal value for CRPSS is 0. This would be attained for $\overline{\sigma_Y^2} = \sigma_X^2$, from Eqs. [3.32] and [3.33], respectively. If the standard error σ_X^2 is equal to the average spread $\overline{\sigma_Y^2}$, the spread of the members will represent the true range of possibilities for the predicted climate (Goddard et al., 2013). There is a statistically significant evidence that the results for the spatially averaged CRPSS depicted in Figure 8.8 do not satisfy that desired condition. Therefore, neither the

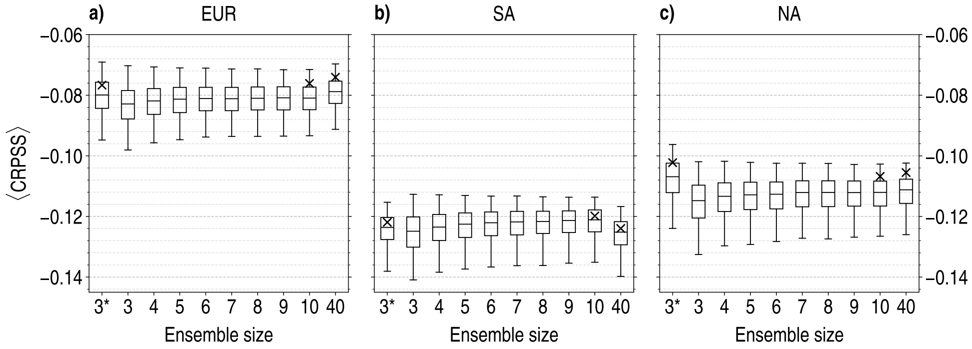

Figure 8.8: Spatially averaged CRPSS (⟨CRPSS⟩) of SST in lead years 2-9 for different ensemble sizes in the **a**) EUR domain, **b**) SA domain and **c**) NA domain. While $\mathrm{TrDC}_{\mathrm{kNN}}$ has been used to correct the drift in the EUR and NA domains, $\mathrm{ICDC}_{\mathrm{kNN}}$ has been considered for the SA domain. Crosses denote the averages for ENS3, ENS10 and ENS40. Box plots show the results of a bootstrapping for which lines identify the median and boxes and whiskers enclose the confidence intervals at the 50 % and 95 % levels, respectively. The ensemble size 3 with the symbol "*" represents the manually selected ENS3.

subensemble nor full ensemble spreads are good representations of the prediction uncertainty on average for SST over any domain.

Note that, likewise in Section 3.3.2, the conditional bias of the CESM-DPLE subensembles has not been explicitly removed. Goddard et al. (2013) suggest correcting the conditional bias together with the mean drift to estimate CRPSS because of the negative influence these biases have on the reliability. However, the conditional bias has not been explicitly removed here because the purpose of this analysis was to evaluate the performance of the drift correction methods described above, without additional enhancements, and the performance of the drift-corrected subensembles which may be potentially used as input information in DD simulations. The use of the DeFoReSt approach instead of the methods analyzed here might contribute to improving the results obtained for CRPSS, since it explicitly includes a correction of both mean and conditional biases alongside an adjustment of the ensemble spread.

8.4. Concluding remarks

The purpose of this Chapter has been to explore some alternative drift correction methods to the MDC approach to correct the lead time-dependent drift in CESM-DPLE, as well as to provide guidance to select a representative 3-member subensemble to conduct future potential DD simulations in a context of limited access to computing resources. In the first part of this Chapter, the most skilful correction technique over some domains of special interest in the context of DD (the EUR, SA and NA domains) has been selected to correct the CESM-DPLE data. Several variables and climate indices have been evaluated to make the decision. In the second part, the drift-corrected SST has been analyzed to carry out the member selection and build a subensemble for each domain. These subensembles have been composed by the "best", the "worst" and an "intermediate" member in terms of their ability to reproduce the SST variability. By selecting members with heterogeneous skill levels, part of the spread of ENS40, or of ENS10 at least, is expected to be retained, since these members cover the whole range of possible single performances among the 10 members available for DD.

All methods contribute to reducing the RMSE observed in the uncorrected CESM-DPLE experiments for all variables. In terms of ACC, there are differences between the methods, and their performances sometimes vary depending on the analyzed field and the lead time. The $\mathrm{TrDC_{kNN}}$ method has been chosen as one of the most adequate techniques for drift correction in the EUR and NA domains. The TrDC-like methods have shown promising results in the prediction of NAO, obtaining significant positive correlations for the NAO index in lead years 2–9, with higher values than those obtained by Smith et al. (2019, 2020) using larger but uncorrected ensembles. Among these methods, the $\mathrm{TrDC_{kNN}}$ has obtained the maximum ACC, with a value of 0.68 in lead years 2–9. Additionally, significant positive correlations have also been obtained in the first half of the decade for this method, with values of 0.53 and 0.49 in lead years 1–4 and 2–5, respectively. The ICDC-like methods usually get better results in lead year 1 for all variables in the three domains. They slightly improve the representation of the subtropical Pacific SST in terms of ACC in lead years 2–9, showing performances which are in general slightly better than those provided by the other methods. The $\mathrm{ICDC_{kNN}}$ has been chosen to correct the CESM-DPLE data for the SA domain. It provides a small added value to the predictive skill for TNI over ICDC and $\mathrm{ICDC_{FIT}}$ in the second half of the decade. If a more straightforward and less computationally demanding technique is required, the ICDC method may also be a good option, as it additionally performs slightly better than $\mathrm{ICDC_{kNN}}$ in terms of the TNI at the beginning of the decade.

In a context of limited access to computing resources, the modest ENS3 could be a good alternative to very large ensembles for some specific applications. The added value of increasing the ensemble size to the predictive skill of SST is very small in comparison with the increase of the computing requirements to conduct the DD simulations. This behaviour is shared with other temperature variables, as shown by Reyers et al. (2019) and Sienz et al. (2016), and discussed in Chapter 5.

However, some fields such as precipitation clearly benefit from using larger ensemble sizes (Reyers et al., 2019). In any case, the corrected predictions of SST lack of reliability on average over all domains, regardless of the ensemble size. The results obtained for the $\langle$CRPSS$\rangle$ of SST show statistically significant negative results, so the average ensemble spread is not appropriate to quantify its forecast uncertainty. Other methods considering an explicit correction of both mean and conditional biases, such as the DeFoReSt approach (Pasternack et al., 2018, 2021), may help to improve the probabilistic skill of the predictions.

9

Conclusions

The main purpose of this Thesis has been to generate a collection of dynamically downscaled DCPs over the IP and evaluate their accuracy and reliability, as well as their predictive skill compared to the global DCPs and a set of dynamically downscaled uninitialized experiments. This evaluation has been carried out for PR, T_{max}, T_{min} and T_{mean}. The DD simulations were conducted with the WRF model in two nested domains. The coarse-grid domain was defined to cover the EURO-CORDEX region with an horizontal resolution around 50 km, whereas the fine-grid domain, with an approximate resolution of 10 km, was centered in the IP. To the best of my knowledge, the research presented here constitutes the first study which comprehensively assesses the performance of a dynamically downscaled DPS at an horizontal resolution of 10 km, becoming the maximum resolution attained in this branch of the climate prediction.

At the time of writing this dissertation, the only DPS which publicly provides all the fields required to run WRF is the CESM-DPLE. Thus, it supplied the ICs and LBCs to conduct the DD decadal experiments. Additionally, the CESM-LE provided the information needed to generate the dynamically downscaled uninitialized experiments which have been considered in part of the evaluation of the predictive skill. In spite of the huge development achieved in climate modeling during the last three decades, models are intrinsically based on approximations and, consequently, contain biases which arise from different sources. Therefore, a bias correction was applied to CESM-DPLE and CESM-LE data before using their fields as input information for WRF simulations with the aim of reducing the potentially negative impact that those biases may have on the downscaled product. Since these simulations were conducted in a context of limited access to computing resources, a selection of suitable members for DD was carried out for both CESM-DPLE and CESM-LE ensembles. A subensemble composed by 4 members was built from each global product, focusing on the ability of the single members to represent the SST on average over the EURO-CORDEX domain in lead years 2–9. The 4-member subensembles were composed of the two members showing the best performance (the "best" members), the member showing the worst performance (the "worst" member) and a member with an "intermediate" behaviour. With this strategy, a representative subensemble of each global product was built. The results obtained for the bias correction and subensemble selection have been presented in Chapter 3. The main conclusions extracted from this analysis are summarized in the following:

- **The correction of the lead time-dependent mean drift of CESM-DPLE contributes to robustly improving the representation of the predicted SST on average.** This improvement has been mainly observed in terms of a reduction of the ⟨RMSE⟩ compared to the uncorrected predictions, whereas there is not a significant increase of the ⟨ACC⟩. The 4-member CESM-DPLE subensemble has obtained a higher accuracy than the individual members, but slightly lower than the 10-member subensemble available for DD. Nevertheless, the added value to the SST predictive skill of increasing the ensemble size is very low. Both ensembles perform similarly in terms of reliability, with a slightly better result for the 4-member ensemble. However, the MDC method is not enough to get reliable predictions for the SST on average over the EURO-CORDEX domain, neither for the 4-member ensemble, nor for the 10-member ensemble, nor for the 40-member ensemble. An additional correction of the conditional bias may help to improve these results.
- **The correction of the mean bias of CESM-LE helps to consistently reduce the errors in the representation of SST on average by the uninitialized experiments compared to the biased product.** As for the initialized experiments, the corrected 10-member CESM-LE subensemble performs slightly better than the 4-member subensemble, which in turn performs better than the individual members. The differences between the 10-member and the 4-member subensembles are not very pronounced (the differences between both ⟨RMSE⟩ are lower than 0.1 K) compared to the cost of increasing the ensemble size from 4 to 10 members.

The evaluation of the WRF-DPLE downscaled hindcasts for PR has been conducted in Chapter 4. This evaluation encompasses an assessment of the accuracy and reliability of the hindcasts alongside the comparison with the global CESM-DPLE and the WRF-LE uninitialized experiments. Before the analysis, the WRF-DPLE experiments were recalibrated by applying the DeFoReSt approach to reduce the unconditional and conditional biases and adjust the ensemble spread. The most important results are summarized in the following:

- **The signal-to-noise paradox affects the WRF-DPLE hindcasts for PR.** It is also present in the hindcasts from the 4-member CESM-DPLE subensemble. The results obtained for the spatial distribution of the RPC in the IP indicate that the predictive skill for PR would clearly benefit from the addition of new members to the ensemble. These results are consistent with previous studies available in literature.

- **In the analysis of the WRF-DPLE hindcasts for PR at annual scale, the IP is predominantly covered by positive ACC at all lead times.** However, these results are not robust enough to indicate statistical significance over most part of the domain. The most promising outcomes have been found in the northwestern sector of the IP in lead year 1. On the other hand, the spatial distributions of $\mathrm{RMSE_R}$ show the lowest errors in the northern regions of the domain at all lead times, since the highest PR rates are commonly observed there.

- **At seasonal scale, some of the best results obtained for PR in terms of ACC have been found in JJA for lead years 1 and 2–5.** These spatial distributions show generalized positive results in the IP, with the statistical significance constrained to very specific areas depending on the lead time. Relatively similar outcomes are also observed at certain lead times in DJF and MAM. The lowest $\mathrm{RMSE_R}$ values have been found in MAM, whereas the highest scores are shown mainly across the southern regions in JJA. These high relative errors have been mainly motivated by the low PR rates observed in this part of the domain during this season.

- **The WRF-DPLE predictive skill for PR, with climatology as reference, is limited.** At annual scale, the spatial distributions of $\mathrm{MSSS_C}$ show generalized negative results over the IP at almost all lead times. Small regions with positive scores have been also found, especially in lead year 1, but the results lack of statistical significance. These outcomes are caused by two concurring factors: the aforementioned low ACC and the large absolute CB values. Even for $\mathrm{MSSS_{CBA}}$ ($\mathrm{MSSS_C}$ for $\mathrm{CB} = 0$), the predictive skill is limited because of the low ACC. Similar results have been obtained at seasonal scale.

- **The WRF-DPLE hindcasts for PR are reliable over large areas of the domain.** Therefore, the ensemble spread can be used to quantify the uncertainty of the predictions in those regions. At annual scale, the hindcasts are reliable over almost the whole domain in lead years 2–9, with the areas of not significant CRPSS results being smaller at other lead times. The regions in the northern half of the IP generally show the best results. This reliability is motivated by the not significant results obtained for LESS, indicating that there are not significant differences between the average ensemble spread and the squared standard error in those locations. At seasonal scale, the hindcasts are reliable over almost the whole domain for all lead times in MAM. Very good results have been obtained also in JJA and, to a lesser extent, in SON.

- **In the analysis of PR at annual scale, the best results for the WRF-DPLE predictive skill, with CESM-DPLE as reference, have been found in lead years 6–9.** Large areas show positive MSSS_G values, as consequence of the generalized positive $\Delta\text{ACC}_\text{G}$ results obtained at this lead time alongside the improvement also observed in terms of ΔCB_G in some regions. However, the statistical significance of the MSSS_G results is restricted to small regions in the central eastern part of the domain. This lack of statistical significance is also observed at the other lead times, but with smaller areas showing positive scores. The northern, northwestern and central eastern regions of the domain have generally obtained the best results in lead years 2–5, 6–9 and 2–9. In lead year 1, they are fundamentally observed along the Mediterranean coast and part of the Northern Subplateau.
- **The highest WRF-DPLE predictive skill, with CESM-DPLE as reference, for the predicted PR at seasonal scale has been obtained in JJA.** Generalized positive results for MSSS_G have been obtained in lead years 2–5, 6–9 and, especially, in lead years 2–9. The statistical significance, however, is constrained to small regions. These outcomes are a consequence of the joint action of the positive results obtained for $\Delta\text{ACC}_\text{G}$ and the significant positive results found for ΔCB_G.
- **The reliability of the PR hindcasts is higher for WRF-DPLE than for CESM-DPLE.** Although these results lack of statistical significance, positive $\Delta\text{CRPSS}_\text{G}$ values have generally been found over the whole domain in lead years 1, 2–5 and 6–9 at annual scale, being predominant also in the coastal and some inner regions for lead years 2–9. These results have been produced by a large improvement in the representation of the ensemble spread for WRF-DPLE in comparison to CESM-DPLE, as the results obtained for LESSS_G indicate. For the same reasons, the highest improvements in terms of reliability at seasonal

scale have been obtained in DJF and SON.

- **Compared to the WRF-LE uninitialized experiments, the added value to the predictive skill provided by WRF-DPLE in terms of $\mathrm{MSSS_U}$ has been mainly found in lead year 1 over the western part of the domain.** The positive values obtained for $\Delta\mathrm{ACC_U}$ and $\Delta\mathrm{CB_U}$ in those regions have led to such results. On the other hand, the added value quantified by $\mathrm{MSSS_U}$ is fundamentally restricted to northern and southern regions in lead years 2–5. At seasonal scale, the results for $\mathrm{MSSS_U}$ are more promising in JJA and SON, when an added value of WRF-DPLE is observed over large areas in the domain, especially in lead year 1, although the statistically significant results are shown only in small regions. Since this evaluation has been focused on the period 1990–2005 due to the limited access to CESM-LE data, these results should be taken with caution.
- **In the analysis of the spatially averaged WRF-DPLE hindcasts for PR, some of the best results have been obtained in the NW region, as well as in the CN region to a lesser extent.** There, the hindcasts time series are able to reproduce part of the relative maximums and minimums of the observational time series. In general, however, there is a poor representation of the PR in all lead times in terms of accuracy. The magnitude of the ensemble mean signal is very low compared to that from the observational time series and the width of the confidence intervals, as consequence of the signal-to-noise paradox. On the other hand, there are many regions where the hindcasts perform well in terms of reliability, particularly in lead years 2–9.
- **The results obtained for WRF-DPLE could have been partially influenced by the limited ability of the 4-member CESM-DPLE subensemble to represent the spatio-temporal variability of the SLP.** Although the CESM-DPLE hindcasts can partially reproduce the SLP variability, they cannot clearly capture most part of the main spatio-temporal variability modes extracted from the PCA computed with the ERA5 SLP. Neither the 4-member nor the 10-member CESM-DPLE ensemble means are able to consistently simulate the NAO. Further improvements in this line could help to enhance the predictive skill of the downscaled product.

The evaluation of T_{max}, T_{min} and T_{mean} has been presented in Chapter 5. The applied procedure is similar to that followed in the analysis of PR in Chapter 4. The most remarkable results obtained from this evaluation are the following:

- **The signal-to-noise paradox is also present in the WRF-DPLE hindcasts for the NSAT variables.** However, the results for the RPC indicate that this presence is weaker than for PR. Although the addition of new members to the downscaled ensemble would contribute to increasing the predictive skill, the improvement is expected to be lower than it would be for the same number of members in the case of PR, as shown by previous studies.

- **The WRF-DPLE hindcasts for the NSAT variables show generalized positive ACC values over the whole domain at annual scale.** The statistical significance of these outcomes depends on the NSAT variable and the lead time. The best represented field in terms of accuracy is T_{min}, with predominant significant positive ACC values at all lead times. The highest RMSE and less generalized significant ACC results have been obtained for T_{max}, although there are still large areas with significant positive ACC values in lead years 6–9 and 2–9. The results for T_{mean} represent a midpoint between those from the other NSAT variables.

- **At seasonal scale, the highest ACC scores for the WRF-DPLE NSAT hindcasts have been obtained in MAM and JJA.** For T_{min} and T_{mean}, the significant positive results span large areas in lead years 2–5, 6–9 and 2–9 in both seasons. Similar outcomes are shown for T_{max}, but with less statistical significance in general. In lead year 1, the results are predominantly positive in MAM and, to a lesser extent, in JJA, but they lack of statistical significance.

- **The highest predictive skill of the downscaled hindcasts, with climatology as reference, is observed for T_{min}, followed by T_{mean} and T_{max}, in that order.** The areas covered by positive $\mathrm{MSSS_C}$ values are predominant at all lead times for T_{min} and T_{mean}, whereas some regions with negative results cover part of the domain for T_{max}. These results are a consequence of the spatial distributions obtained for ACC and CB. While the CB is close to zero and even not significant over wide regions for T_{min} and, to a lesser extent, for T_{mean}, the CB is generally significant and negative for T_{max} over almost the whole domain at all lead times.

- **At seasonal scale, the assessment of the predictive skill for the NSAT variables, with climatology as reference, has given the best results in JJA and MAM.** The spatial distributions of $\mathrm{MSSS_C}$ obtained for the three NSAT variables resemble to those obtained for ACC at seasonal scale. The spatial distributions of CB show large areas with not statistically significant results, mainly in MAM for the three NSAT variables in lead years 2–5 and 2–9, and also in lead years 6–9 for T_{min}.

- **At annual scale, the WRF-DPLE hindcasts for T_{max} and T_{mean} are reliable over almost the whole domain in lead years 2–5, 6–9 and 2–9.** Nevertheless, in the case of T_{min}, the downscaled hindcasts are reliable mainly in lead year 1 over the northern regions of the IP. The results obtained for the CRPSS have been determined by the outcomes for LESS. For T_{min}, there is a general significant ensemble underdispersion at all lead times, whereas not significant LESS values have been found mainly in lead years 6–9 and 2–9 for T_{mean} and, especially, for T_{max}. At seasonal scale, the results vary depending on the season and the variable.

- **The most robust added value of WRF-DPLE to the predictive skill at annual scale, with CESM-DPLE as reference, has been found in lead years 6–9 for T_{max} and T_{mean}, whereas it is shown in lead year 1 for T_{min}.** In these cases, the positive $\mathrm{MSSS}_{\mathrm{G}}$ values cover most part of the domain, with very large areas showing statistical significance for T_{max} and T_{min}. Wide regions with positive $\mathrm{MSSS}_{\mathrm{G}}$ results can be generally found in lead years 2–9 for the three NSAT variables. Since the results obtained for $\Delta\mathrm{ACC}_{\mathrm{G}}$ are mainly not significant, the high $\mathrm{MSSS}_{\mathrm{G}}$ scores have been motivated by the very good results obtained in terms of $\Delta\mathrm{CB}_{\mathrm{G}}$.

- **At seasonal scale, the WRF-DPLE predictive skill for the NSAT variables, with CESM-DPLE as reference, depends on the season and tends to increase when the performance of the global product is limited.** For instance, the highest $\mathrm{MSSS}_{\mathrm{G}}$ scores for T_{max} have been observed for lead years 6–9 and 2–9 in SON, when significant positive values are widespread along the IP. A similar situation is presented for T_{mean}. For T_{min}, positive $\mathrm{MSSS}_{\mathrm{G}}$ outcomes are predominant at all lead times in DJF, although not always showing statistical significance. As at annual scale, the role of the improvement in terms of $\Delta\mathrm{CB}_{\mathrm{G}}$ is more important than that for $\Delta\mathrm{ACC}_{\mathrm{G}}$ to achieve the positive $\mathrm{MSSS}_{\mathrm{G}}$ scores.

- **There are not significant differences between the reliability of WRF-DPLE and CESM-DPLE hindcasts for the NSAT variables.** The $\Delta\mathrm{CRPSS}_{\mathrm{G}}$ are not large enough in absolute value to show statistical significance for any NSAT variable. The regions which show an improvement or deterioration in the representation of the ensemble spread depend on the variable, the lead time and the time scale of the analysis (annual or seasonal).

- **The results indicating an added value to predictive skill of the WRF-DPLE hindcasts over the WRF-LE uninitialized experiments at annual scale have**

been found mainly in lead year 1 for the three NSAT variables. The positive MSSS_U results cover most part of the domain for T_max and T_mean at this lead time, whereas they are mainly constrained to the northern regions and the southern half of the domain for T_min. However, the statistical significance of the results is shown only for a few specific locations. At seasonal scale, the most promising MSSS_U results have been found in MAM, with positive and significant scores covering large areas of the domain in lead years 1 and 2–5 for the three variables. The contribution of $\Delta\text{ACC}_\text{U}$ to MSSS_U is often comparable to that of ΔCB_U at both annual and seasonal scales.

- **A generalized overestimation of the observed NSAT anomalies at the beginning of the control period has been found in the analysis of the spatially averaged lead time series.** This overestimation has contributed to enhance the differences between the trends of the WRF-DPLE and the observational time series for the three variables. These errors in the representation of the trends were transferred to the WRF-DPLE experiments by CESM-DPLE during the DD simulations, since they are also present in the global product.

The sensitivity of WRF simulations to extreme ICs of soil moisture has been examined in Chapter 6, focusing on the analysis of the spin-up time required by several variables to reach a dynamical equilibrium. The simulations were initialized in two different dates, 1990-01-01 and 1990-07-01, with ERA-Interim providing the ICs and LBCs for all variables with the exception of soil moisture. The ICs of soil moisture were calculated by combining the SMI with some physical soil properties which depend on the soil textures. Three different ICs were considered to represent a wet, a dry and a very dry soil. The main findings of this analysis are summarized in the following:

- **The spin-up time required by soil moisture to guarantee the dynamical equilibrium in all soil layers after starting from extreme ICs in the IP is 8 years**. This is the maximum length of the spin-up period obtained for some locations, mainly placed in the southern half of the domain and in the Ebro Valley, but shorter values have been found in other regions depending on the soil moisture ICs, the depth of the soil layer, the initialization date, the atmospheric conditions, the soil texture and the land class.
- **The longest spin-up periods of soil moisture are observed for the deepest layer.** The upper layers are subjected to a higher variability because their

interactions with the atmosphere are more immediate than for the deeper layers. The water coming from precipitation needs more time to reach the deeper layers, making the soil stated defined by the ICs more persistent and, therefore, leading to longer spin-up times.

- **Drier conditions contribute to increasing the spin-up time for soil moisture.** The capacity of holding water is higher for drier soils, so the absence of soil water slows down the transport of water coming from precipitation through the soil layers. Thus, soil moisture tends to reach the dynamical equilibrium later in experiments with drier ICs.

- **The initialization date also affects the length of the spin-up period for soil moisture, although its effects depend on the ICs.** The wet ICs in January are closer to the control state than in July. On the other hand, the deviation from the control state for the dry and very dry experiments in July is lower than in January. The shock produced by initializing the experiments from states more different from that established by the control simulation often leads to longer times to reach the dynamical equilibrium, resulting in longer spin-up periods. Those differences between January and July control soil moistures are partly caused by the differences between the meteorological conditions of the preceding months (lower temperatures and higher PR rates in December, whereas the opposite situation is observed in June).

- **The hydraulic conductivity associated to the texture class influences on the spin-up time of soil moisture.** The soils with low hydraulic conductivity, such as clay and clay loam soils, are expected to have a higher climate persistence because more time than for other soil types might be required to observe changes in soil moisture content. Although this phenomenon can be modulated by the atmospheric conditions, the longest spin-up times are often observed in regions characterized by these texture classes (e.g., the Ebro and Guadalquivir Valleys).

- **The stomatal resistance, whose value depends on the land class (among other factors), affects the soil moisture content and, therefore, influences on the spin-up time of soil moisture.** Low stomatal resistance values, typical of cropland areas, favour high evapotranspiration rates from vegetation which impact on the soil moisture content. In regions habituated to exhibit minor resistance to this type of evapotranspiration, the soil state in dry and very dry scenarios generally tends to be closer to the control state than in the case of regions where the stomatal resistance is higher, leading to shorter spin-up times in the former. This phenomenon has been only observed in the upper three layers, which comprise the root zone of the land class.

- **The initialization with extreme soil moisture ICs has an impact on the spin-up time required by the atmospheric variables. PR, with spin-up times mostly below 10 months, is the least affected variable.** There are not big differences between the three soil scenarios, suggesting that the role of the model internal variability prevails over the imposed ICs of soil moisture in determining the evolution of PR. However, there are differences between the experiments with different initialization dates, especially for the wet simulations, influenced by the magnitude of the deviation of soil moisture ICs from the control state.
- **The spin-up time for T_{max}, the most affected variable by the soil moisture ICs, reachs values even higher than 36 months (3 years) in some locations.** There are similarities between the spatial distributions of the spin-up period required by T_{max} and soil moisture in the upper three layers, highlighting the existence of land surface-atmosphere coupling processes involving these two variables, in accordance with previous studies. This relationship, although still existing to some extent, is not as marked for T_{min} and T_{mean}. After PR, T_{min} is the least affected variable by the soil moisture ICs. In general, the spin-up periods of T_{min} show lengths shorter than 12 months. On the other hand, longer periods are required for T_{mean}, with maximum lengths starting from 24 up to 32 months in some locations.

The results found in Chapter 6 led to consider the soil initialization in the DD simulations presented in Chapter 7, devoted to analyze the WRF-DPLE experiments for the decade 2015–2025. Since no spin-up time was considered in the analyses conducted in Chapters 4 and 5 (it would have implied the loss of the first simulated years), the predictive skill might have experienced some deterioration in presence of spin-up biases, at least during the first years of the simulations. Although the spin-up time required for these variables may be shorter than that indicated in Chapter 6 under normal ICs of soil moisture, a dynamically equilibrated soil state, taken from the WRF control simulation, was used to initialize the simulations examined in Chapter 7 to improve as much as possible the predictive skill of the downscaled predictions.

The results examined in Chapter 7 correspond to the predictions for the same variables analyzed in the control period: PR, T_{max}, T_{min} and T_{mean}. Since the observational information is available up to 2022, the predicted variables have been compared with the observational values in lead years 1 and 2–5. The most relevant findings are summarized in the following:

- **In regions with reliable predictions for PR, the WRF-DPLE$_4$ predicted anomalies for this variable are generally positive at the beginning of the decade and turn into negative during the second half of the decade at annual scale.** The prediction for the lead years 2–9 shows generalized negative anomalies over almost the whole domain. In both lead years 6–9 and 2–9, the strongest negative anomalies have been found in the northwestern regions, the Central System and the Pyrenees, with values below -12 mm/month in some locations. However, at all lead times, the width of the confidence intervals is much higher than the magnitude of the anomaly and, therefore, they usually encompass anomaly values with opposite signs. At seasonal scale, the largest areas with reliable negative anomalies for lead years 2–9 are shown in SON. In MAM, there are many regions which show observational values outside the confidence intervals in spite of predictions are reliable there. Although there is a probability of 10 % associated to these occurrences, they may also be partially due to the gap between the control period and the decade 2015–2025 and the fact that the length of the control period relative to this gap is not long enough.

- **The WRF-DPLE$_{10}$ predictions for PR are qualitatively similar to those from WRF-DPLE$_4$. However, with a few exceptions, WRF-DPLE$_{10}$ predictions generally get lower errors than WRF-DPLE$_4$ predictions at annual scale for lead years 1 and 2–5.** Additionally, the WRF-DPLE$_{10}$ predictions show slightly more moderate anomalies than WRF-DPLE$_4$ at annual scale. At seasonal scale, the areas covered by lower WRF-DPLE$_{10}$ errors than those for WRF-DPLE$_4$ are generally larger than those with better WRF-DPLE$_4$ accuracy, with the clear exceptions of MAM and JJA in lead year 1, when WRF-DPLE$_4$ performs better over most part of the domain.

- **The results obtained for the spatially averaged anomalies of PR reproduce what has been observed from the grid-point perspective**. In regions with reliable predictions, the strongest anomalies are negative and have been found in the CN and NW regions in lead years 6–9 and 2–9 for WRF-DPLE$_4$. The results obtained for WRF-DPLE$_{10}$ are qualitatively similar, but generally getting lower errors than WRF-DPLE$_4$ in lead years 1 and 2–5.

- **The WRF-DPLE$_4$ predictions for the NSAT variables show positive anomalies at annual scale for all lead times over the whole domain**. In regions with reliable predictions, the highest anomalies have been found in lead years 2–5 for the three NSAT variables. These anomalies are generally higher than 1 K at this lead time, reaching the maximum values between 1.5 K and 1.75 K for T_{max} in regions over the Iberian System. At seasonal scale, the anomalies are even more exacerbated. The highest predicted anomalies have been found in JJA for the three variables, with maximum outcomes up to 2 K in large areas of the domain with reliable predictions for T_{max} in lead years 6–9 over several southeastern and northeastern regions. At annual scale, the confidence intervals commonly do not contain values with distinct signs in lead years 2–5, 6–9 and 2–9 for the three variables. Similarly to PR, observational values fall outside the confidence intervals in some cases, being more frequent in DJF for T_{max} and T_{mean}. In addition to the sample size and the gap between the control period and the decade 2015–2025, this may be partially due to the fact that these predictions were initialized from a dynamically equilibrated soil state, thus reducing the spin-up time required for them, unlike the hindcasts which were used to compute those intervals.

- **For WRF-DPLE$_{10}$, the NSAT anomalies are qualitatively similar to those obtained for WRF-DPLE$_4$ at all lead times, in general with minor differences in the magnitude of the anomalies at annual scale.** The areas where the accuracy of WRF-DPLE$_{10}$ is higher than that of WRF-DPLE$_4$ in lead years 1 and 2–5 are generally larger than those areas where WRF-DPLE$_4$ outperforms WRF-DPLE$_{10}$ for the three NSAT variables. At seasonal scale, the differences between WRF-DPLE$_{10}$ and WRF-DPLE$_4$ predictons are slightly higher than at annual scale, with some locations getting anomalies of distinct signs depending on the ensemble size.

- **At annual scale, the results obtained for the spatially averaged predictions of the NSAT variables summarize, for each region, what has been obtained at a grid-point scale.** All regions show positive anomalies for all NSAT variables at all lead times for both ensemble sizes. The anomalies are commonly above 0.5 K, often reaching anomalies higher than 1 K. With the exception of lead year 1, the confidence intervals do not contain values with different signs for any variable in any region. In general, the error made by WRF-DPLE$_{10}$ is lower than that made by WRF-DPLE$_4$ in lead years 1 and 2–5, with a few exceptions.

Finally, a set of drift correction methods has been examined in Chapter 8. The MDC method contributes to reducing the mean lead time-dependent bias in the ICs and LBCs provided by CESM-DPLE, but it does not account for higher-order biases, such as those observed in the representation of trends. Thus, additional correction techniques have been evaluated to explore alternatives to the MDC to improve as much as possible the predictive skill of the input data in DD simulations. The most skilful correction method over several domains of special interest for DD purposes (the EUR, SA and NA domains) has been used to correct the CESM-DPLE data and select a subensemble of 3 members for potential future DD experiments. The main conclusions extracted from this analysis are:

- **The TrDC_{kNN} method has been chosen as one of the most adequate methods for drift correction in the EUR and NA domains, whereas ICDC_{kNN} may be suitable for the SA domain.** The TrDC-like methods show promising results in the prediction of NAO, obtaining significant positive correlations for the NAO index in lead years 2–9, with higher values than those obtained by other studies which considered larger but uncorrected ensembles. The ICDC-like methods generally get better results in lead year 1 for all variables in the three domains. Moreover, ICDC_{kNN}, in particular, slightly improves the representation of the subtropical Pacific SST in terms of ACC in lead years 2–9, additionally providing a small added value to the representation of TNI over ICDC and ICDC_{FIT} in the second half of the decade. If a more straightforward and less computationally demanding technique is required, the ICDC method may also be a good option, as it additionally performs slightly better than ICDC_{kNN} with respect to the TNI at the beginning of the decade.

- **The modest ENS3 could be a good alternative to very large ensembles in a context of limited computing resources for some specific applications**. The added value of increasing the ensemble size to the predictive skill of SST is very small in comparison to the increase of the computing requirements to conduct the DD simulations. This behaviour is shared by other temperature variables, as shown in other studies and discussed in Chapter 5. However, some fields such us PR would clearly benefit from using larger ensemble sizes to generate the downscaled predictions.

- **The corrected predictions for SST lack of reliability, regardless of the ensemble size.** The results obtained for the ⟨CRPSS⟩ of SST show statistically significant negative results, so the average ensemble spread is not appropriate to quantify the forecast uncertainty. Other methods which consider an explicit correction of both unconditional and conditional biases, such as the DeFoReSt approach, may contribute to improving the probabilistic skill of the predictions.

Potential future works

The research presented in this Thesis evidences the valuable role that WRF can play for the generation of high resolution DCPs over the IP. Significant improvements over the global CESM-DPLE and the uninitialized WRF-LE experiments have been found at both annual and seasonal scales for NSAT variables, whereas the added value of WRF-DPLE to the predictive skill for PR is more limited. Nevertheless, since the signal-to-noise paradox is so strong in the case of PR, as revealed by the RPCs calculated in Section 4.1 and showed by Smith et al. (2019) for a larger ensemble composed of different DPSs, these results could be highly improved by adding new members to the downscaled ensemble. Reyers et al. (2019) showed how much the predictive skill for PR and, to a lesser degree, NSAT can be improved by increasing the ensemble size. Additionally, the same authors, as well as Sienz et al. (2016), also showed that the increase of the sample size (i.e., the number of start dates) positively influences on the robustness of the predictions. Therefore, the ideal next step would consist in increasing the downscaled ensemble size up to 10 members (the maximum size attainable with CESM-DPLE providing the input information) and progressively adding new start dates from 1969 backwards, to enhance as much as possible the predictive skill of the downscaled product. However, this would only be possible with access to enough computing resources to conduct the simulations.

The analysis of the downscaled DCPs could continue with an assessment of the temperature and precipitation extremes, since their frequency and intensity may increase under conditions of climate change in the IP in the next decades (Cardoso Pereira et al., 2020; Lorenzo and Alvarez, 2022, 2020). Additionally, as one of the main advantages of DD over other downscaling approaches is that the RCM is able to produce a wide range of variables for each downscaled experiment, this analysis could also be extended to other phenomena of special interest in the IP, such as drought and wildfires, which commonly involve the interplay of multiple climate variables in their assessments. Not only the IP is currently vulnerable to these natural hazards for human, environmental and economic reasons (Cammalleri et al., 2020; San-Miguel-Ayanz et al., 2018), but also they are expected to have a more prominent presence in the future (García-Valdecasas et al., 2021; Turco et al., 2018).

Another interesting line of research would be to further explore the added value that soil initialization can provide to the predictive skill of the downscaled ensemble.

Kothe et al. (2016) studied different soil initialization strategies which are potentially applicable for decadal DD simulations. The initialization from a dynamically equilibrated soil state, reviewed in that study, has been used in this Thesis. However, Kothe et al. (2016) also examined more complex approaches which could lead to better results, such as the generation of the initial soil state by running the LSM of the RCM in a standalone mode or by implementing data assimilation techniques.

The multiple applications of DD in the branch of the DCP and their potential ramifications open a vast field of research which could be explored in future works by taking the study presented in this Thesis as a solid starting point.

Conclusiones

El propósito principal de esta Tesis ha sido generar una colección de DCPs de alta resolución con simulaciones DD en la IP y evaluar su precisión y fiabilidad, así como su habilidad predictora frente a DCPs globales y experimentos no inicializados de alta resolución. Esta evaluación se ha llevado a cabo para las variables PR, T_{max}, T_{min} y T_{mean}. Las simulaciones DD fueron realizadas en dos dominios anidados. El dominio mayor se definió para cubrir la región de EURO-CORDEX con una resolución espacial alrededor de 50 km, mientras que el dominio menor, con una resolución aproximada de 10 km, se centró en la IP. Hasta donde alcanza mi conocimiento, la investigación presentada aquí constituye el primer estudio que evalúa en profundidad el desempeño de un DPS generado mediante el método de reducción dinámica de escala a una resolución de 10 km, convirtiéndose en la máxima resolución espacial jamás lograda en esta rama de la predicción del clima.

En el momento de redacción de esta Tesis, el único DPS que proporciona públicamente todos los campos requeridos para ejecutar WRF es el CESM-DPLE. Por tanto, fue éste el utilizado para suministrar las ICs y las LBCs para llevar a cabo los experimentos decenales de DD. Adicionalmente, el CESM-LE proporcionó la información necesaria para generar los experimentos no inicializados de alta resolución que han sido utilizados en una parte de la evaluación. A pesar del gran desarrollo logrado en la modelización del clima durante las últimas tres décadas, los modelos están intrínsecamente basados en aproximaciones y, en consecuencia, contienen sesgos que surgen de fuentes diversas. Por tanto, se aplicó una corrección de sesgo a los datos del CESM-DPLE y el CESM-LE antes de usar sus variables como información de entrada en las simulaciones con WRF para así reducir el impacto potencialmente negativo que estos sesgos pudieran tener en el producto final. Como estas simulaciones se llevaron a cabo en un contexto de acceso limitado a recursos computacionales, se realizó una selección de los miembros de los conjuntos de CESM-DPLE y CESM-LE disponibles para las simulaciones DD. Se construyó un subconjunto compuesto por 4 miembros por cada producto global en función de los resultados obtenidos en relación a la capacidad de los miembros individuales para representar la SST en promedio sobre el dominio de EURO-CORDEX en el rango de predicción de 2–9 años.

Cada subconjunto de 4 miembros fue compuesto por los dos que mostraron un mejor desempeño (los "mejores"), el miembro con el peor desempeño (el "peor") y un miembro con un comportamiento "intermedio". Con esta estrategia, se construyó un subconjunto representativo de cada producto global. Los resultados obtenidos para la corrección del sesgo y la selección de los subconjuntos han sido presentados en el Capítulo 3. Las conclusiones principales extraídas de estos análisis se resumen a continuación:

- **La corrección de la deriva promedio dependiente del rango de predicción, aplicada sobre los datos del CESM-DPLE, contribuye a mejorar consistentemente la representación de los pronósticos de SST en promedio.** Esta mejora ha sido observada principalmente a través de la reducción del ⟨RMSE⟩ frente a las predicciones sin corregir. Sin embargo, no ha habido una mejora significativa del ⟨ACC⟩. El subconjunto de 4 miembros del CESM-DPLE ha obtenido una precisión mayor que la de los miembros individuales, pero ligeramente más baja que la del subconjunto de los 10 miembros disponibles para simulaciones DD. Aún así, el valor añadido a la habilidad predictora de la SST por el incremento del tamaño del conjunto es muy bajo. Ambos subconjuntos han obtenido resultados similares en lo que respecta a la fiabilidad de las predicciones, con el subconjunto de 4 miembros obteniendo un resultado ligeramente mejor. No obstante, el uso de la técnica MDC no ha sido suficiente para conseguir predicciones fiables de la SST en promedio en el dominio de EURO-CORDEX, ni para el subconjunto de 4-miembros, ni para el de 10, ni para el conjunto total de los 40 miembros. La incorporación de una corrección del sesgo condicional podría ayudar a mejorar estos resultados.
- **La corrección del sesgo promedio del CESM-LE provoca una reducción notable de los errores en la representación de la SST en promedio para los experimentos no inicializados frente al producto sin corregir.** Del mismo modo que para los experimentos inicializados, el subconjunto de 10 miembros del CESM-LE tiene un desempeño ligeramente mejor que el del subconjunto de 4 miembros, que a su vez ha obtenido mejores resultados que los miembros individuales. Las diferencias entre ambos subconjuntos no son muy pronunciadas (las diferencias entre valores del ⟨RMSE⟩ son menores a 0.1 K) en comparación con el coste de incrementar el tamaño del conjunto de 4 a 10 miembros.

La evaluación de las retropredicciones del WRF-DPLE para PR se ha llevado a cabo en el Capítulo 4. Esta evaluación ha consistido en un análisis de la precisión y la fiabilidad de las retropredicciones y en su comparación con los experimentos globales del CESM-DPLE y con los no inicializados del WRF-LE. Los experimentos del WRF-DPLE fueron previamente recalibrados siguiendo el método DeFoReSt para reducir los sesgos incondicionales y condicionales y para ajustar la dispersión de los miembros del conjunto. Los resultados más importantes se resumen a continuación:

- **La paradoja señal-ruido afecta a las retropredicciones del WRF-DPLE para PR**. También está presente en el subconjunto de 4 miembros del CESM-DPLE. Los resultados obtenidos para la distribución espacial de RPC en la IP indican que la habilidad predictora para PR se podría beneficiar claramente del incremento del número de miembros del conjunto. Estos resultados son consistentes con estudios previos disponibles en la literatura.

- **En el análisis a escala anual de las retropredicciones del WRF-DPLE para PR, los resultados positivos de ACC predominan en la IP en todos los rangos de predicción.** Sin embargo, estos resultados no son lo suficientemente robustos como para obtener significación estadística sobre la mayor parte del dominio. Los resultados más prometedores han sido encontrados en el sector noroeste de la IP en el rango de predicción de 1 año. Por otro lado, las distribuciones espaciales de $\mathrm{RMSE_R}$ muestran los errores más bajos en las regiones del norte del dominio para todos los rangos de predicción, ya que es allí donde comúnmente se observan las mayores tasas de PR.

- **A escala estacional, algunos de los mejores resultados obtenidos para PR en términos de ACC han sido encontrados en JJA para los rangos de predicción de 1 y 2–5 años.** Estas distribuciones espaciales muestran resultados positivos generalizados en la IP, con significación estadística limitada a áreas muy específicas en función del rango de predicción. Se han obtenido resultados relativamente similares para ciertos rangos de predicción en DJF y MAM. Los valores más bajos de $\mathrm{RMSE_R}$ han sido encontrados en MAM, mientras que los resultados más elevados se muestran principalmente sobre las regiones del sur en JJA. Estos errores relativos tan altos son motivados principalmente por las bajas tasas de PR que se observan en esta parte del dominio durante esta estación.

- **La habilidad predictora del WRF-DPLE para PR, tomando la climatología como referencia, es limitada.** Las distribuciones espaciales de $\mathrm{MSSS_C}$ muestran resultados negativos generalizados sobre la IP en casi todos los rangos de predicción. También se han encontrado resultados positivos en algunas regiones pequeñas, especialmente en el rango de predicción de 1 año, pero carecen de significación estadística. Estos resultados son causados por dos factores concurrentes: los ya mencionados valores bajos de ACC y los grandes valores absolutos de CB. Incluso en términos de MSSS_{CBA} (MSSS_{C} con CB $= 0$), la habilidad predictora es limitada debido a las bajas correlaciones. Se han obtenido resultados similares a escala estacional.

- **Las retropredicciones del WRF-DPLE para PR son fiables sobre grandes áreas del dominio.** Por tanto, la dispersión de los miembros del conjunto puede ser utilizada para cuantificar la incertidumbre de las predicciones en esas regiones. En escala anual, las retropredicciones son fiables sobre casi todo el dominio para el rango de predicción de 2–9 años, con áreas de resultados no significativos para CRPSS más pequeñas en otros rangos de predicción. Las regiones de la mitad norte de la IP muestran generalmente los mejores resultados. Esta fiabilidad se debe a los resultados no significativos obtenidos para LESS, que indican que no existen diferencias significativas entre la dispersión promedio del conjunto y el error cuadrático estándar en esas localizaciones. En escala estacional, las retropredicciones son fiables sobre casi todo el dominio para todos los rangos de predicción en MAM. Se han obtenido también muy buenos resultados en JJA y, en menor medida, en SON.

- **En el análisis de PR a escala anual, se han encontrado los mejores resultados para la habilidad predictora del WRF-DPLE, en comparación con el CESM-DPLE, para el rango de predicción de 6–9 años.** Se observan resultados positivos para MSSS_{G} en áreas extensas como consecuencia de los valores positivos generalizados obtenidos para $\Delta\text{ACC}_{\text{G}}$ en este rango de predicción y de la mejora encontrada también en términos de $\Delta\text{CB}_{\text{G}}$ en algunas regiones. Sin embargo, la significación estadística está restringida a áreas muy pequeñas en la región este central del dominio. Esta falta de significación estadística se observa también en otros rangos de predicción, pero con resultados positivos abarcando superficies menores. Las regiones del norte, del noroeste y del este central del dominio han obtenido generalmente los mejores resultados para los rangos de predicción de 2–5, 6–9 y 2–9 años. En el primer año, éstos se observan fundamentalmente a lo largo de la costa mediterránea y parte de la Submeseta Norte.

- **Los mejores resultados para la habilidad predictora del WRF-DPLE en relación a PR a escala estacional, en comparación con el CESM-DPLE, se han**

obtenido en JJA. Se han obtenido resultados positivos generalizados para $\mathrm{MSSS_G}$ en los rangos de predicción de 2–5, 6–9 y, especialmente, 2–9 años. La significación estadística, sin embargo, se limita a pequeñas regiones. Estos resultados son consecuencia de la acción conjunta de los resultados positivos obtenidos para $\Delta\mathrm{ACC_G}$ y los resultados significativos y positivos encontrados para $\Delta\mathrm{CB_G}$.

- **La fiabilidad de las retropredicciones del WRF-DPLE para PR es mayor que para las del CESM-DPLE.** Aunque estos resultados carecen de significación estadística, se han encontrado valores de $\Delta\mathrm{CRPSS_G}$ positivos sobre todo el dominio en general para los rangos de predicción de 1, 2–5 y 6–9 años en escala anual, siendo predominantes también en la costa y en algunas regiones interiores en el rango de predicción de 2–9 años. Estos resultados han sido producidos por una gran mejora en la representación de la dispersión del conjunto del WRF-DPLE en comparación con el CESM-DPLE, tal y como muestran los resultados obtenidos para $\mathrm{LESSS_G}$. Por las mismas razones, se han encontrado las mayores mejoras en términos de fiabilidad a escala anual en DJF y SON.

- **En comparación con los experimentos no inicializados del WRF-LE, el valor añadido a la habilidad predictora proporcionado por el WRF-DPLE en términos de $\mathrm{MSSS_U}$ se encuentra principalmente para el rango de predicción de 1 año en la parte oeste del dominio.** Los valores positivos obtenidos para $\Delta\mathrm{ACC_U}$ y $\Delta\mathrm{CB_U}$ en esas regiones conducen a tales resultados. Por otro lado, este valor añadido se limita fundamentalmente a regiones del norte y del sur para el rango de predicción de 2–5 años. A escala estacional, los resultados para $\mathrm{MSSS_U}$ son más prometedores en JJA y en SON, con un valor añadido del WRF-DPLE sobre grandes áreas del dominio, especialmente en el rango de predicción de 1 año, aunque los resultados estadísticamente significativos solo aparecen en algunas localizaciones. Como esta evaluación se ha limitado al periodo 1990–2005 por la falta de disponibilidad de datos del CESM-LE, estos resultados deberían ser tomados con cautela.

- **En el análisis de la retropredicciones espacialmente promediadas del WRF-DPLE para PR, algunos de los mejores resultados han sido obtenidos en la región NW y, en menor medida, también en la región CN.** En esos lugares, las series temporales de las retropredicciones son capaces de reproducir parte de los máximos y mínimos relativos presentes en las series observacionales. En general, sin embargo, no hay una buena representación de PR en ningún rango de predicción en términos de precisión. La magnitud de la señal del promedio del conjunto es muy baja en comparación con la de las series observacionales y con la amplitud de los intervalos de confianza, como consecuencia de la paradoja señal-ruido. Por otro lado, hay algunas regiones donde las retropredicciones han obtenido buenos resultados en términos de fiabilidad, particularmente en el rango de predicción de 2–9 años.

- **Los resultados obtenidos para el WRF-DPLE pueden haber estado parcialmente influenciados por la capacidad limitada del subconjunto de 4 miembros del CESM-DPLE para representar la variabilidad espacio-temporal de la SLP.** Aunque las retropredicciones del CESM-DPLE pueden reproducir parcialmente la variabilidad de la SLP, no son capaces de capturar la mayor parte de los modos principales de variabilidad espacio-temporal extraídos del PCA realizado para la SLP de ERA5. Ni el subconjunto de 4 miembros ni el de 10 miembros son capaces de simular de manera consistente la NAO. Todas las mejoras que puedan implementarse en esta línea podrían ayudar a incrementar la habilidad predictora del producto final del WRF-DPLE.

La evaluación de T_{max}, T_{min} y T_{mean} se ha presentado en el Capítulo 5. El procedimiento aplicado ha sido similar al seguido en el análisis de PR. Los resultados más notables de esta evaluación son los siguientes:

- **La paradoja señal-ruido también está presente en las retropredicciones de WRF-DPLE para NSAT.** Sin embargo, los resultados obtenidos para el RPC indican que su presencia es más débil que para PR. Aunque la adición de nuevos miembros al conjunto contribuiría a incrementar su habilidad predictora, se espera que la mejora sea menor de lo que sería para el mismo número de miembros en el caso de PR, tal y como se muestra en estudios previos.
- **Las retropredicciones del WRF-DPLE para las variables de NSAT muestran valores positivos generalizados de ACC en todo el domino a escala anual.** La significación estadística de estos resultados depende de la variable y el rango de predicción. La variable mejor representada es T_{min}, con resultados significativos y positivos predominantes para ACC en todos los rangos de predicción. Los valores más elevados de RMSE y los resultados con menor significación estadística en términos de ACC han sido obtenidos para T_{max}, aunque para esta variable también se han encontrado áreas extensas con resultados positivos y significativos de ACC en los rangos de predicción de 6–9 y 2–9 años. Los

resultados obtenidos para T_{mean} se encuentran a medio camino entre los de las otras dos variables de NSAT.

- **A escala estacional, los valores más elevados de ACC obtenidos por las retropredicciones del WRF-DPLE para las variables de NSAT han sido obtenidos en MAM y JJA**. Para T_{min} y T_{mean}, los resultados positivos se extienden por regiones amplias en los rangos de predicción de 2–5, 6–9 y 2–9 años en ambas estaciones. Se han obtenido resultados similares para T_{max}, pero con mejor significación estadística en general. En el rango de predicción de 1 año, los resultados son predominantemente positivos en MAM y, en menor medida, en JJA, aunque carecen de significación estadística.

- **La mayor habilidad predictora de las retropredicciones del WRF-DPLE, con la climatología como referencia, ha sido encontrada para T_{min}, seguida de T_{mean} y T_{max}, en ese orden.** Las áreas cubiertas por resultados positivos de $\mathrm{MSSS_C}$ son predominantes en todos los rangos de predicción para T_{min} y T_{mean}, mientras que algunas regiones con resultados negativos se extienden por parte del dominio en el caso de T_{max}. Estos resultados son consecuencia de los obtenidos para las distribuciones espaciales de ACC y CB. Mientras que para CB se han encontrado valores cercanos a cero e incluso no significativos sobre regiones amplias en el caso de T_{min} y, en menor medida, T_{mean}, los valores de CB han sido generalmente significativos y negativos sobre casi todo el dominio en todos los rangos de predicción para T_{max}.

- **A escala estacional, la evaluación de la habilidad predictora para NSAT, con la climatología como referencia, ha dado los mejores resultados en JJA y MAM.** Las distribuciones espaciales de $\mathrm{MSSS_C}$ obtenidas para las tres variables de NSAT recuerdan a aquellas obtenidas para ACC también a escala estacional. Las distribuciones espaciales de CB muestran grandes superficies con resultados sin significación estadística, principalmente en MAM, para las tres variables de NSAT en los rangos de predicción de 2–5 y 2–9 años, y también para el rango de 6–9 años en el caso de T_{min}.

- **A escala anual, las retropredicciones del WRF-DPLE para T_{max} y T_{min} son fiables sobre casi todo el dominio en los rangos de predicción de 2–5, 6–9 y 2–9 años.** Sin embargo, en el caso de T_{min}, las retropredicciones son fiables principalmente para el rango de 1 año en las regiones del norte de la IP. Los resultados obtenidos para CRPSS han sido determinados por los obtenidos para LESS. En el caso de T_{min}, se ha observado una subdispersión significativa del conjunto en todos los rangos de predicción, mientras que se han encontrado resultados no significantivos de LESS en los rangos de 6–9 y 2–9 años para T_{mean} y, sobre todo, para T_{max}. En escala estacional, los resultados varían dependiendo de la estación y la variable.

- **El valor añadido más robusto de WRF-DPLE a la habilidad predictora a escala anual, tomando al CESM-DPLE como referencia, ha sido encontrado para el rango de predicción de 6–9 años en el caso de T_{max} y T_{mean}, mientras que se muestra en el primer año para T_{min}.** En esos casos, los valores positivos de $\mathrm{MSSS_G}$ cubren la mayor parte del dominio, con áreas muy extensas que muestran significación estadística para los resultados de T_{max} y T_{min}. Se han encontrado grandes regiones con resultados positivos generalizados en el rango de 2–9 años para las tres variables. Como los resultados obtenidos para $\Delta\mathrm{ACC_G}$ son principalmente no significativos, los elevados resultados de $\mathrm{MSSS_G}$ han sido motivados por la mejora observada en términos de $\Delta\mathrm{CB_G}$.

- **En escala estacional, la habilidad predictora del WRF-DPLE, con el CESM-DPLE como referencia, depende de la estación y tiende a crecer cuando el desempeño del producto global es limitado.** Por ejemplo, los valores más altos de $\mathrm{MSSS_G}$ para T_{max} han sido observados para los rangos de predicción de 6–9 y 2–9 años en SON, siendo generalizados los resultados positivos y significativos a lo largo de la IP. Se presenta una situación similar para T_{mean}. En cuanto a T_{min}, los resultados positivos de $\mathrm{MSSS_G}$ son predominantes en todos los rangos de predicción en DJF, aunque no siempre con significación estadística. Como en escala anual, el papel de la mejora en términos de $\Delta\mathrm{CB_G}$ ha sido más determinante que los resultados de $\Delta\mathrm{ACC_G}$ en la obtención de los resultados positivos de $\mathrm{MSSS_G}$.

- **No se han encontrado diferencias significativas entre la fiabilidad de las retropredicciones del WRF-DPLE y del CESM-DPLE para las variables de NSAT.** Los resultados de $\Delta\mathrm{CRPSS_G}$ no son lo suficientemente grandes en valor absoluto para que exista esa significación. Las regiones que muestran una mejora o deterioro en la representación de la dispersión del conjunto dependen de la variable, el rango de predicción y la escala temporal del análisis (anual o estacional).

- **Los resultados que muestran un valor añadido de WRF-DPLE a la habilidad predictora a escala anual, tomando los experimentos no inicializados del WRF-LE como referencia, han sido encontrados principalmente en el rango**

de predicción de 1 año para las tres variables de NSAT. Los valores positivos de $\mathrm{MSSS_U}$ cubren la mayor parte del dominio para T_{max} y T_{mean} en este rango, mientras que éstos aparecen solo en las regiones del norte y de la mitad sur del dominio en el caso de T_{min}. Sin embargo, hay significación estadística en los resultados únicamente en unas pocas localizaciones. A escala estacional, se han encontrado los resultados más prometedores de $\mathrm{MSSS_U}$ en MAM, con valores positivos y significativos que cubren superficies extensas del dominio en los rangos de 1 y 2–5 años para las tres variables. La contribución de $\Delta\mathrm{ACC_U}$ a los resultados de $\mathrm{MSSS_U}$ es a menudo comparable a la de $\Delta\mathrm{CB_U}$ tanto en escala anual como estacional.

- **Se ha encontrado una sobreestimación generalizada de las anomalías observacionales de NSAT al comienzo del periodo de control en las series temporales de los promedios espaciales de las variables.** Esta sobreestimación ha contribuido a incrementar las diferencias entre las tendencias de las series del WRF-DPLE y las series observacionales. Estos errores en la representación de las tendencias fueron transferidos al WRF-DPLE por el CESM-DPLE durante las simulaciones DD, ya que estos también están presentes en el producto global.

El Capítulo 6 se ha dedicado al estudio de la sensibilidad de las simulaciones con WRF a condiciones iniciales extremas de humedad del suelo, centrando el análisis en el tiempo de spin-up requerido por algunas variables para alcanzar el equilibrio dinámico. Las simulaciones fueron inicializadas en dos fechas diferentes, 1990-01-01 y 1990-07-01, con ERA-Interim proporcionando las ICs y LBCs para todas las variables con excepción de la humedad del suelo. Las ICs de la humedad del suelo fueron calculadas combinando el SMI con algunas propiedades físicas del suelo determinadas por su textura. Se consideraron tres ICs diferentes para representar un suelo húmedo, uno seco y uno muy seco. A continuación, se resumen los resultados más importantes derivados de este análisis:

- **El tiempo de spin-up requerido por la humedad del suelo para garantizar el equilibrio dinámico en todas las capas de suelo tras partir de unas ICs extremas en la IP es de 8 años.** Esta es la duración máxima del periodo de spin-up obtenida para algunas localizaciones situadas principalmente en la mitad sur del dominio y en el valle del Ebro, pero se han encontrado periodos más cortos en otras regiones dependiendo de las ICs de humedad del suelo, la profundidad de la capa del suelo, la fecha de inicialización, las condiciones atmosféricas, la textura y el uso de suelo.

- **Los periodos de spin-up más largos de la humedad del suelo se han encontrado para la capa más profunda.** Las capas superiores estás sujetas a una mayor variabilidad porque interactúan con la atmósfera de manera más inmediata que las más profundas. El agua que proviene de la precipitación necesita más tiempo para llegar a las estas últimas, haciendo más duradero el estado del suelo definido por las ICs y, por tanto, conduciendo a tiempos de spin-up más largos.

- **Las condiciones secas contribuyen al incremento del tiempo de spin-up de la humedad del suelo.** La capacidad de almacenar agua es mayor en los suelos secos, de modo que la ausencia de humedad ralentiza el transporte de agua proveniente de la precipitación a través de las capas del suelo. Por tanto, la humedad del suelo tiende a alcanzar el equilibrio dinámico más tarde en los experimentos con ICs secas y muy secas.

- **La fecha de inicialización también influye en la longitud del periodo de spin-up de la humedad del suelo, aunque la forma en la que afecta depende de las ICs.** Las ICs húmedas en enero se encuentran más cerca del estado de control que en julio. En cambio, la desviación respecto a este estado es menor en julio que en enero para los experimentos seco y muy seco. El impacto producido por la inicialización de los experimentos desde estados más alejados del establecido por la simulación de control conduce a menudo a tiempos más largos para alcanzar el equilibrio dinámico, resultando en periodos de spin-up más largos. Estas diferencias entre la humedad del suelo de control en enero y julio son parcialmente causadas por las diferencias entre las condiciones meteorológicas de los meses precedentes (temperaturas menores y tasas de PR mayores en diciembre y la situación opuesta en junio).

- **La resistencia estomatal, cuyos valores dependen del uso de suelo (entre otros factores), afecta al contenido de humedad y, por tanto, tiene influencia en el tiempo de spin-up de la humedad del suelo.** Valores bajos de la resistencia estomatal, típicos de las áreas de cultivo, favorecen tasas altas de evapotranspiración desde la vegetación que tienen un impacto en el contenido de humedad del suelo. En regiones habituadas a exhibir una menor resistencia a este tipo de evapotranspiración, el estado del suelo en los escenarios seco y muy seco generalmente tiende a estar más cerca del estado de control que en el caso de las regiones donde la resistencia estomatal es elevada, conduciendo a periodos de spin-up más cortos. Este fenómeno se observa únicamente en las tres capas superiores, que comprenden la zona de raíz de este uso de suelo.

- **La inicialización con ICs extremas de humedad del suelo tiene un impacto en el tiempo de spin-up necesario para las variables atmosféricas. PR, con un tiempo de spin-up generalmente menor a 10 meses, ha sido la menos afectada**. No ha habido grandes diferencias entre los tres tipos de escenarios, lo que sugiere que el papel de la variabilidad interna del modelo prevalece sobre las ICs de humedad del suelo impuestas en la determinación de la evolución de PR. Sin embargo, existen diferencias entre los experimentos con diferentes fechas de inicialización, especialmente entre las simulaciones húmedas, influenciadas por la magnitud de la desviación de las ICs de humedad del suelo del estado de control.

- **El tiempo de spin-up para T_{max}, la variable más afectada por las ICs de humedad del suelo, alcanza valores superiores a los 36 meses (3 años) en algunas localizaciones.** Existen similitudes entre las distribuciones espaciales del periodo de spin-up requerido por T_{max} y por la humedad del suelo en las tres capas superiores, evidenciando la existencia de procesos de acoplamiento tierra-atmósfera que involucran a estas dos variables, de acuerdo con estudios previos. Esta relación, aunque se mantiene en cierto grado, no es tan marcada para T_{min} y T_{mean}. Tras PR, T_{min} es la variable menos afectada por las ICs de humedad del suelo. En general, los periodos de spin-up de T_{min} muestran duraciones menores a los 12 meses. Por otro lado, se requieren periodos mayores para T_{mean}, con duraciones máximas que van desde 24 hasta 32 meses en algunas localizaciones.

Los resultados obtenidos en el Capítulo 6 llevaron a considerar la inicialización del suelo en las simulaciones DD presentadas en el Capítulo 7, dedicado a analizar los experimentos del WRF-DPLE para la década 2015-2025. Como no se tuvo en cuenta ningún tiempo de spin-up en los análisis realizados en los capítulos Capítulos 4 y 5 (lo que habría implicado la pérdida de los primeros años simulados), la habilidad predictora puede haber experimentado algún deterioro en presencia de sesgos debidos al spin-up, al menos durante los primeros años de las simulaciones. Aunque el tiempo de spin-up requerido para esas variables puede ser más corto que el indicado en el Capítulo 6 bajo ICs normales de humedad de suelo, un estado de suelo dinámicamente equilibrado, tomado de la simulación de control con WRF, se utilizó para inicializar las simulaciones examinadas en el Capítulo 7 con el propósito de mejorar tanto como fuese posible la habilidad predictora de las predicciones de alta resolución.

Los resultados presentados en el Capítulo 7 corresponden a las predicciones de las mismas variables analizadas en el periodo de control: PR, T_{max}, T_{min} y T_{mean}. Como la información observacional está disponible hasta 2022, las variables pronosticadas han sido comparadas con los valores observacionales en los rangos de predicción de 1 y 2–5 años. Los resultados más relevantes son resumidos a continuación:

- **En las regiones con predicciones fiables para PR, las anomalías pronosticadas por el WRF-DPLE$_4$ para esta variable en escala anual son generalmente positivas al comienzo de la década y se vuelven negativas durante la segunda mitad de la misma.** La predicción para el rango de 2–9 años muestra anomalías negativas generalizadas sobre casi todo el dominio. En los rangos de 6–9 y de 2–9 años, se han encontrado las anomalías negativas más intensas en las regiones del noroeste, en el Sistema Central y en los Pirineos, con valores por debajo de -12 mm/mes en algunas localizaciones. Sin embargo, en todos los rangos de predicción, la amplitud de los intervalos de confianza es mucho mayor que la magnitud de las anomalías y, por tanto, éstos usualmente contienen valores de anomalías con signos opuestos. En escala estacional, las superficies más extensas con pronósticos fiables de anomalías negativas en el rango de 2–9 años se muestran en SON. En MAM, hay muchas regiones en las que se han encontrado valores observacionales fuera de los intervalos de confianza, a pesar de que las predicciones son fiables allí. Aunque hay una probabilidad del 10 % asociada a este tipo de casos, también podrían deberse parcialmente al tamaño del espacio vacío que existe entre el fin del periodo de control y la década 2015–2025 y a que la duración del periodo de control, en comparación con el tamaño de este espacio, no es lo suficientemente larga.
- **Las predicciones del WRF-DPLE$_{10}$ para PR son cualitativamente similares a las del WRF-DPLE$_4$. No obstante, con algunas excepciones, las predicciones del WRF-DPLE$_{10}$ generalmente obtienen errores más pequeños que las del WRF-DPLE$_4$ en escala anual para los rangos de predicción de 1 y 2–5 años.** Adicionalmente, las predicciones del WRF-DPLE$_{10}$ muestran unas anomalías ligeramente más moderadas que las del WRF-DPLE$_4$ en escala anual. En escala estacional, las áreas cubiertas por errores más bajos de WRF-DPLE$_{10}$ que de WRF-DPLE$_4$ son generalmente más grandes que aquellas que indican una mayor precisión para WRF-DPLE$_4$, con las claras excepciones de MAM y JJA en el rango de predicción de 1 año, cuando el desempeño de WRF-DPLE$_4$ es

mejor sobre la mayor parte del dominio.

- **Los resultados obtenidos para los promedios espaciales de las anomalías de PR reproducen aquello que se ha observado desde la perspectiva de punto de rejilla.** En los lugares con predicciones fiables, las anomalías más intensas son negativas y han sido encontradas en las regiones CN y NW en los rangos de predicción de 6–9 y 2–9 años para WRF-DPLE$_4$. Los resultados obtenidos para WRF-DPLE$_{10}$ son cualitativamente similares, pero obteniendo errores menores que WRF-DPLE$_4$ en los rangos de 1 y 2–5 años en general.

- **Las predicciones del WRF-DPLE$_4$ para las variables de NSAT muestran anomalías positivas en escala anual para todos los rangos de predicción sobre todo el dominio.** En regiones con predicciones fiables, las anomalías más elevadas han sido encontradas en el rango de 2–5 años para las tres variables de NSAT. Estas anomalías son generalmente superiores a 1 K en ese rango de predicción, alcanzando los valores máximos entre 1.5 K y 1.75 K para T_{max} en regiones del Sistema Ibérico. En escala estacional, las anomalías llegan a ser más exacerbadas. Los pronósticos de anomalías más altas se han encontrado en JJA para las tres variables, con máximos de hasta 2 K en extensas superficies del dominio con predicciones fiables para T_{max} en el rango de 6–9 años en algunas regiones del sureste y noreste. En escala anual, los intervalos de confianza no contienen en general valores de distinto signo en los rangos de 2–5, 6–9 y 2–9 años para ninguna variable de NSAT. Igual que en el caso de PR, hay valores observacionales que caen fuera de los intervalos de confianza, siendo más frecuentes en DJF para T_{max} y T_{mean}. Además del tamaño del periodo de control y del espacio vacío entre el fin de este periodo y la década 2015-2025, esto podría estar en cierto grado relacionado con el hecho de estas predicciones fueron inicializadas desde un estado de suelo dinámicamente equilibrado, que favorece la reducción del tiempo de spin-up necesario para las mismas, al contrario que las retropredicciones que se usaron para calcular esos intervalos.

- **En cuanto al WRF-DPLE$_{10}$, se han obtenido anomalías de NSAT cualitativamente similares a las del WRF-DPLE$_4$ en todos los rangos de predicción, encontrándose en general valores menores en la magnitud de las anomalías para la escala anual.** Las áreas donde la precisión del WRF-DPLE$_{10}$ es mayor que la del WRF-DPLE$_4$ en los rangos de 1 y 2–5 años para las tres variables de NSAT son generalmente más grandes que las áreas en las que el desempeño de WRF-DPLE$_4$ es mejor. En escala estacional, las diferencias entre WRF-DPLE$_{10}$

y WRF-DPLE$_4$ son ligeramente mayores que en la anual, con algunas localizaciones obteniendo anomalías de distinto signo dependiendo del tamaño del conjunto.

- **Los resultados obtenidos por los promedios espaciales de las predicciones para las variables de NSAT en escala anual resumen lo que se ha obtenido desde la perspectiva de punto de rejilla.** Todas las regiones muestran anomalías positivas para todas las variables de NSAT en todos los rangos de predicción para los dos tamaños de conjunto. Las anomalías son comúnmente superiores a 0.5 K y a menudo alcanzan valores mayores que 1 K. Con la excepción del rango de 1 año, los intervalos de confianza no contienen anomalías de signos diferentes para ninguna variable en ninguna región. En general, el error cometido por WRF-DPLE$_{10}$ es menor que el de WRF-DPLE$_4$ en los rangos de 1 y 2–5 años, con algunas excepciones.

Finalmente, una serie de métodos de corrección de deriva han sido examinados en el Capítulo 8. El método MDC contribuye a reducir el sesgo promedio dependiente del rango de predicción en las ICs y LBCs proporcionadas por el CESM-DPLE, pero no tiene en cuenta otros sesgos de orden superior, como los observados en la representación de las tendencias. Por tanto, varias técnicas de corrección adicionales han sido evaluadas para explorar alternativas al método MDC y así mejorar tanto como sea posible la habilidad predictora de los datos de entrada en las simulaciones DD. El mejor método de corrección sobre diferentes dominios de especial interés en un contexto de DD (los dominios EUR, SA y NA) ha sido utilizado para corregir los datos del CESM-DPLE y seleccionar un subconjunto de 3 miembros para posibles experimentos DD futuros. Las principales conclusiones que se extraen de este análisis son:

- **El método TrDC$_{\mathrm{kNN}}$ ha sido escogido como uno de los más adecuados para la corrección de la deriva en los dominios EUR y NA, mientras que el método ICDC$_{\mathrm{kNN}}$ puede ser apropiado para el dominio SA.** Los métodos de la familia TrDC muestran resultados prometedores en la predicción de la NAO, obteniendo correlaciones positivas significativas para el índice NAO en el rango de predicción de 2–9 años, con valores más elevados que los obtenidos por otros estudios que consideraron conjuntos de mayor tamaño sin corregir. Los métodos de la familia ICDC han obtenido generalmente mejores resultados en el rango de 1 año para todas las variables analizadas en los tres dominios. Además, el método ICDC$_{\mathrm{kNN}}$, en particular, mejora ligeramente la representación de la SST en el Pacífico subtropical en términos de ACC en el rango de 2–9 años, aportando además un pequeño valor añadido a la representación del TNI frente a ICDC y ICDC$_{\mathrm{FIT}}$ en la segunda mitad de la década. Si un método más directo y menos demandante computacionalmente fuese requerido, el método ICDC podría ser también una buena opción, ya que adicionalmente presenta un desempeño ligeramente mejor que ICDC$_{\mathrm{kNN}}$ en lo que respecta al TNI al comienzo de la década.

- **El modesto ENS3 podría ser una buena alternativa a conjuntos más grandes en un contexto de acceso limitado a recursos computacionales para algunas aplicaciones específicas.** El valor añadido del incremento del tamaño del conjunto a la habilidad predictora de la SST es muy pequeño en comparación con el incremento de los requerimientos computacionales de las simulaciones DD. Este comportamiento es compartido por otras variables de temperatura, tal y como muestran otros estudios previos y como se ha discutido en el Capítulo 5. Sin embargo, otros campos como PR se beneficiarían claramente de emplear conjuntos de gran tamaño para generar predicciones de alta resolución.
- **Las predicciones corregidas para SST carecen de fiabilidad, independientemente del tamaño del conjunto.** Los resultados obtenidos para el ⟨CRPSS⟩ de la SST muestran valores negativos y estadísticamente significativos, de modo que la dispersión promedio del conjunto no es apropiada para cuantificar la incertidumbre de las predicciones. Otras técnicas alternativas que consideren una corrección explícita tanto de los sesgos incondicionales como de condicionales, como el método DeFoReSt, pueden contribuir a mejorar las predicciones también en términos probabilísticos.

Trabajos futuros

La investigación presentada en esta Tesis evidencia el valor del papel que puede desempeñar WRF en la generación de DCPs de alta resolución en la IP. Se han encontrado mejoras significativas en los resultados de las predicciones en comparación con los experimentos globales del CESM-DPLE y los no inicializados del WRF-LE para las variables de NSAT, mientras que el valor añadido por el WRF-DPLE a la habilidad predictora para PR es más limitado. Sin embargo, como la paradoja señal-ruido es tan fuerte en el caso de PR, tal y como revelan los RPCs calculados en la Sección 4.1 y como encontraron Smith et al. (2019) para un gran conjunto compuesto por diferentes DPSs, estos resultados podrían ser ampliamente mejorados añadiendo nuevos miembros al conjunto de alta resolución. Reyers et al. (2019) mostraron el grado en el que la habilidad predictora para PR y, en menor medida, para las variables de NSAT, puede ser mejorada incrementando el tamaño del conjunto. Adicionalmente, en el mismo trabajo, así como en Sienz et al. (2016), se mostró también que el incremento del tamaño de la muestra (es decir, del número fechas de inicialización) tiene una influencia positiva en la solidez de las predicciones. Por tanto, el siguiente paso ideal consistiría en incrementar el tamaño del conjunto de predicciones de alta resolución hasta los 10 miembros (el máximo tamaño que se puede alcanzar con el CESM-DPLE proporcionando la información de entrada) y progresivamente ir añadiendo nuevas fechas de inicialización desde 1969 hacia atrás, para así aumentar tanto como se pueda la habilidad predictora del producto de alta resolución. Sin embargo, esto solo sería posible con acceso a recursos computacionales suficientes para realizar las simulaciones.

El análisis de las DCPs de alta resolución podría continuar con una evaluación de los extremos de temperatura y precipitación, ya que su frecuencia e intensidad podría incrementarse en condiciones de cambio climático en la IP durante las próximas décadas (Cardoso Pereira et al., 2020; Lorenzo y Alvarez, 2022, 2020). Además, como una de las pricipales ventajas del DD sobre otras técnicas de reducción de escala es que el RCM es capaz de producir un amplio abanico de variables para cada experimento, este análisis se podría extender también a otros fenómenos de especial interés en la IP, como la sequía o los incendios, que suelen involucrar a múltiples variables climáticas en sus evaluaciones. La IP no es solo vulnerable en la actualidad a estos eventos naturales por motivos humanos, medioambientales y económicos (Cammalleri et al., 2020; San-Miguel-Ayanz et al., 2018), sino que además se espera que estos fenómenos tengan una mayor presencia en el futuro (García-Valdecasas et al., 2021; Turco et al., 2018).

Otra línea de investigación interesante sería la de explorar en mayor profundidad el valor añadido que la inicialización del suelo puede proporcionar a la habilidad predictora del conjunto de alta resolución. Kothe et al. (2016) estudiaron diferentes estrategias de inicialización que son potentialmente aplicables en simulaciones DD decenales. La inicialización a partir de un estado de suelo dinámicamente equilibrado, reseñada en ese estudio, ha sido utilizada aquí. No obstante, Kothe et al. (2016) también examinaron métodos más complejos que podrían conducir a resultados mejores, como la generación del estado inicial del suelo con la ejecución del LSM del RCM de forma independiente o con la implementación de técnicas de asimilación de datos.

Las múltiples aplicaciones del DD en la rama de la DCP y sus potenciales ramificaciones abren un amplio campo de investigación que podría ser explorado en trabajos futuros tomando el estudio presentado en esta Tesis como un sólido punto de partida.

A

Supplementary tables

A.1. Köppen-Geiger climate classification

Table A.1: Definition of the Köppen-Geiger climate classes. Adapted from Beck et al. (2023). A description of the acronyms and symbols is available in the next page.

Letter symbol				
1st	**2nd**	**3rd**	**Description**	**Criterion**
A			Tropical	Not (B) & $T_{\text{cold}} \geq 18$
	f		- Rainforest	$P_{\text{dry}} \geq 60$
	m		- Monsoon	Not (Af) & $P_{\text{dry}} \geq 100 - \text{MAP}/25$
	w		- Savannah	Not (Af) & $P_{\text{dry}} < 100 - \text{MAP}/25$
B			Arid	MAP $< 10 \times P_{\text{threshold}}$
	W		- Desert	MAP $< 5 \times P_{\text{threshold}}$
	S		- Steppe	MAP $\geq 5 \times P_{\text{threshold}}$
		h	- Hot	MAT ≥ 18
		k	- Cold	MAT < 18
C			Temperate	Not (B) & $T_{\text{hot}} > 10$ & $0 < T_{\text{cold}} < 18$
	s		- Dry summer	$P_{\text{sdry}} < 40$ & $P_{\text{sdry}} < P_{\text{wwet}}/3$
	w		- Dry winter	$P_{\text{wdry}} < P_{\text{swet}}/10$
	f		- Without dry season	Not (Cs) or (Cw)
		a	- Hot summer	$T_{\text{hot}} \geq 22$
		b	- Warm summer	Not (a) & $T_{\text{mon10}} \geq 4$
		c	- Cold summer	Not (a or b) & $1 \leq T_{\text{mon10}} < 4$
D			Cold	Not (B) & $T_{\text{hot}} > 10$ & $< T_{\text{cold}} \leq 10$
	s		- Dry summer	$P_{\text{sdry}} < 40$ & $P_{\text{sdry}} < P_{\text{wwet}}/3$
	w		- Dry winter	$P_{\text{wdry}} < P_{\text{swet}}/10$
	f		- Without dry season	Not (Ds) or (Dw)
		a	- Hot summer	$T_{\text{hot}} \geq 22$
		b	- Warm summer	Not (a) & $T_{\text{mon10}} \geq 4$
		c	- Cold summer	Not (a, b, or d)
		d	- Very cold winter	Not (a or b) & $T_{\text{hot}} \leq -38$
E			Polar	Not (B) & $T_{\text{hot}} \leq 10$
	T		- Tundra	$T_{\text{hot}} > 0$
	F		- Frost	$T_{\text{hot}} \leq 0$

In Table A.1, MAT is the mean annual air temperature (°C); T_{cold} is the air temperature of the coldest month (°C); T_{hot} is the air temperature of the warmest month (°C); T_{mon10} is the number of months with air temperature > 10 °C (unitless); MAP is the mean annual precipitation (mm y^{-1}); P_{dry} is precipitation in the driest month (mm month^{-1}); P_{sdry} is precipitation in the driest month in summer (mm month^{-1}); P_{wdry} is precipitation in the driest month in winter (mm month^{-1}); P_{swet} is precipitation in the wettest month in summer (mm month^{-1}); P_{wwet} is precipitation in the wettest month in winter (mm month^{-1}); $P_{\mathrm{threshold}} = 2 \times \mathrm{MAT}$ if > 70 % of precipitation falls in winter, $P_{\mathrm{threshold}} = 2 \times \mathrm{MAT} + 28$ if > 70 % of precipitation falls in summer, otherwise $P_{\mathrm{threshold}} = 2 \times \mathrm{MAT} + 14$. Summer (winter) is the six-month period that is warmer (colder) between April–September and October–March.

A.2. Trends of precipitation and near-surface air temperature fields

Table A.2: Trends of spatially averaged WRF-DPLE multiannual mean anomalies of PR, T_{max}, T_{min} and T_{mean} in lead years 1, 2-5, 6-9 and 2-9 at annual scale for each region defined in Section 3.6. While β^{Y} denotes the trend of the WRF-DPLE lead time series, β^{Y-X} identifies the trend of the difference between the WRF-DPLE and AEMET series. The bold formatting indicates that the results are different from zero at the 90 % confidence level.

Lead years	Region (PR)	PR (mm/decade)		Region (NSAT)	T_{max} (K/decade)		T_{min} (K/decade)		T_{mean} (K/decade)	
		β^{Y}	β^{Y-X}		β^{Y}	β^{Y-X}	β^{Y}	β^{Y-X}	β^{Y}	β^{Y-X}
1	EI	-13.38	17.31	SW	**0.24**	-0.04	0.44	-0.06	**0.26**	-0.12
2-5		**-6.35**	-0.19		**0.20**	-0.02	**0.22**	**-0.26**	**0.25**	**-0.12**
6-9		-6.07	-1.41		**0.19**	0.07	**0.26**	**-0.07**	**0.22**	-0.03
2-9		**-5.67**	-0.12		**0.19**	0.03	**0.24**	**-0.17**	**0.23**	**-0.08**
1	WI	3.29	0.44	NO	**0.25**	-0.23	**0.35**	-0.11	**0.29**	-0.17
2-5		-0.44	-21.89		**0.32**	-0.10	**0.23**	**-0.24**	**0.27**	**-0.19**
6-9		-6.76	0.57		**0.26**	0.04	**0.27**	-0.05	**0.26**	0.00
2-9		-4.71	-1.83		**0.28**	-0.04	**0.25**	-0.13	**0.26**	-0.07
1	NE	-18.97	-2.01	CI	**0.22**	**-0.24**	**0.50**	-0.12	**0.27**	**-0.27**
2-5		-4.97	-7.17		**0.29**	-0.02	**0.21**	**-0.32**	**0.26**	-0.15
6-9		**-16.67**	-9.90		**0.28**	**0.19**	**0.27**	-0.07	**0.27**	0.04
2-9		**-12.75**	-3.45		**0.27**	0.07	**0.24**	**-0.20**	**0.25**	-0.05
1	CS	-4.93	24.26	NE	**0.25**	**-0.29**	**0.41**	-0.12	**0.32**	-0.19
2-5		-3.88	4.18		**0.31**	-0.18	**0.22**	**-0.30**	**0.27**	**-0.23**
6-9		-4.14	20.41		**0.32**	-0.01	**0.29**	-0.09	**0.32**	-0.05
2-9		-5.12	13.67		**0.31**	-0.09	**0.25**	**-0.18**	**0.30**	-0.13
1	NW	14.34	14.66	CS	**0.50**	-0.16	**0.53**	-0.12	**0.47**	-0.19
2-5		-7.56	-37.30		**0.30**	**-0.33**	**0.27**	**-0.42**	**0.28**	**-0.41**
6-9		-11.44	39.17		**0.32**	-0.03	**0.29**	**-0.25**	**0.30**	-0.15
2-9		-10.08	0.34		**0.29**	-0.21	**0.27**	**-0.34**	**0.28**	**-0.28**
1	EA	0.09	32.58	EA	**0.25**	**-0.24**	**0.32**	-0.12	**0.26**	**-0.18**
2-5		-6.41	3.86		**0.29**	-0.12	**0.23**	**-0.19**	**0.27**	**-0.15**
6-9		**-15.11**	**-34.41**		**0.27**	0.07	**0.27**	0.02	**0.28**	0.04
2-9		**-11.36**	-15.39		**0.28**	-0.02	**0.25**	-0.07	**0.27**	-0.05
1	SW	10.94	1.51	MT	**0.28**	-0.22	**0.40**	-0.18	**0.30**	**-0.23**
2-5		-1.10	**-45.89**		**0.29**	-0.10	**0.25**	**-0.30**	**0.26**	**-0.21**
6-9		-4.94	-5.45		**0.29**	-0.04	**0.28**	-0.07	**0.26**	-0.07
2-9		-4.14	**-27.96**		**0.28**	-0.06	**0.27**	-0.16	**0.25**	-0.14
1	CN	-20.97	**50.61**	WI	**0.21**	-0.17	**0.35**	-0.15	**0.24**	-0.22
2-5		**-28.65**	**45.19**		**0.29**	0.00	**0.21**	**-0.23**	**0.26**	-0.12
6-9		-10.25	**38.73**		**0.24**	0.12	**0.26**	0.01	**0.24**	0.06
2-9		**-21.58**	34.32		**0.25**	0.05	**0.23**	-0.10	**0.24**	-0.02

Table A.3: As Table A.2 but in DJF.

Lead years	**Region (PR)**	**PR (mm/decade)**		**Region (NSAT)**	**T_{max} (K/decade)**		**T_{min} (K/decade)**		**T_{mean} (K/decade)**	
		β^{Y}	β^{Y-X}		β^{Y}	β^{Y-X}	β^{Y}	β^{Y-X}	β^{Y}	β^{Y-X}
1	EI	16.75	60.82	SW	**0.19**	**-0.28**	**0.25**	-0.23	**0.23**	-0.29
2-5		-0.95	29.76		**0.14**	-0.08	**0.06**	**-0.24**	**0.20**	-0.04
6-9		-1.89	**64.76**		**0.22**	0.12	**0.14**	0.14	**0.17**	0.11
2-9		-2.05	**48.46**		**0.19**	0.02	**0.10**	-0.02	**0.17**	0.01
1	WI	46.09	101.30	NO	**0.20**	**-0.43**	**0.12**	-0.26	**0.14**	**-0.48**
2-5		-8.63	13.51		**0.19**	**-0.17**	**0.19**	-0.14	**0.21**	**-0.13**
6-9		5.29	**148.38**		**0.12**	0.02	**0.18**	0.09	**0.15**	0.02
2-9		-3.34	**95.94**		**0.16**	-0.16	**0.18**	-0.05	**0.17**	**-0.14**
1	NE	32.83	69.08	CI	**0.20**	**-0.41**	**0.23**	-0.21	**0.24**	-0.33
2-5		6.46	20.24		**0.17**	-0.10	**0.10**	-0.09	**0.14**	-0.08
6-9		-10.75	8.64		**0.22**	0.12	**0.08**	0.18	**0.08**	0.10
2-9		-2.41	18.88		**0.19**	-0.03	**0.10**	0.02	**0.11**	-0.07
1	CS	28.23	**93.10**	NE	**0.21**	**-0.43**	**-0.03**	**-0.41**	**0.12**	-0.37
2-5		-16.44	27.00		**0.29**	-0.09	**0.14**	-0.10	**0.25**	-0.06
6-9		1.86	**120.97**		**0.26**	0.01	**0.13**	0.07	**0.17**	0.00
2-9		-9.95	**84.89**		**0.27**	-0.13	**0.14**	-0.04	**0.21**	-0.08
1	NW	50.62	**271.10**	CS	**0.26**	**-0.48**	**0.24**	-0.29	**0.25**	**-0.38**
2-5		5.12	186.86		**0.24**	**-0.32**	**0.14**	**-0.20**	**0.23**	**-0.21**
6-9		26.84	**394.72**		**0.31**	-0.15	**0.12**	0.02	**0.20**	-0.12
2-9		9.70	**316.17**		**0.28**	**-0.21**	**0.13**	-0.10	**0.21**	**-0.13**
1	EA	**-5.50**	-29.50	EA	**0.17**	-0.25	**0.10**	-0.13	**0.13**	-0.19
2-5		-10.26	-12.49		**0.26**	0.06	**0.15**	0.04	**0.16**	0.03
6-9		**-21.36**	-39.22		**0.27**	0.20	**0.07**	0.20	**0.09**	0.15
2-9		**-14.47**	-33.46		**0.25**	0.14	**0.12**	0.12	**0.13**	0.09
1	SW	28.14	100.72	MT	**0.21**	**-0.53**	**0.10**	**-0.51**	**0.14**	**-0.57**
2-5		-35.56	-38.96		**0.19**	-0.21	**0.23**	**-0.17**	**0.22**	**-0.17**
6-9		-4.42	**124.72**		**0.24**	-0.07	**0.17**	-0.04	**0.23**	-0.02
2-9		-22.03	27.17		**0.23**	-0.15	**0.20**	-0.19	**0.22**	-0.17
1	CN	5.81	140.75	WI	**0.21**	**-0.42**	**0.19**	-0.36	**0.18**	**-0.41**
2-5		-15.69	107.65		**0.20**	**-0.16**	**0.08**	-0.20	**0.16**	**-0.12**
6-9		-16.07	76.63		**0.19**	0.06	**0.12**	0.17	**0.11**	0.08
2-9		**-25.08**	55.32		**0.20**	-0.11	**0.10**	-0.06	**0.14**	-0.10

Table A.4: As Table A.2 but in MAM.

Lead years	**Region (PR)**	**PR (mm/decade)**		**Region (NSAT)**	**T_{max} (K/decade)**		**T_{min} (K/decade)**		**T_{mean} (K/decade)**	
		β^{Y}	β^{Y-X}		β^{Y}	β^{Y-X}	β^{Y}	β^{Y-X}	β^{Y}	β^{Y-X}
1	EI	-18.99	36.69	SW	**0.28**	**-0.65**	**0.29**	**-0.44**	**0.28**	**-0.57**
2-5		**-11.91**	20.85		**0.30**	**-0.41**	**0.25**	**-0.50**	**0.28**	**-0.45**
6-9		-1.92	-0.98		**0.33**	-0.28	**0.28**	**-0.37**	**0.31**	**-0.33**
2-9		-5.82	15.59		**0.33**	**-0.36**	**0.27**	**-0.49**	**0.30**	**-0.40**
1	WI	2.31	15.63	NO	**0.34**	**-0.63**	**0.16**	**-0.40**	**0.27**	**-0.50**
2-5		-6.66	-29.32		**0.32**	**-0.49**	**0.20**	**-0.42**	**0.27**	**-0.45**
6-9		0.49	-2.48		**0.32**	**-0.41**	**0.24**	**-0.24**	**0.31**	**-0.31**
2-9		-7.79	**-21.09**		**0.34**	**-0.56**	**0.24**	**-0.32**	**0.29**	**-0.42**
1	NE	-2.09	92.35	CI	**0.31**	**-0.63**	**0.26**	**-0.40**	**0.27**	**-0.55**
2-5		-4.40	26.98		**0.32**	**-0.44**	**0.22**	**-0.46**	**0.28**	**-0.46**
6-9		-10.14	34.76		**0.36**	-0.30	**0.30**	**-0.26**	**0.34**	**-0.28**
2-9		-4.09	**35.51**		**0.37**	**-0.40**	**0.26**	**-0.38**	**0.33**	**-0.41**
1	CS	-1.47	74.38	NE	**0.27**	**-0.69**	**0.21**	**-0.43**	**0.27**	**-0.55**
2-5		-9.98	39.77		**0.35**	**-0.54**	**0.24**	**-0.46**	**0.30**	**-0.50**
6-9		-0.96	-3.52		**0.38**	**-0.47**	**0.31**	**-0.25**	**0.35**	**-0.35**
2-9		-8.33	14.29		**0.38**	**-0.52**	**0.28**	**-0.37**	**0.34**	**-0.43**
1	NW	-9.54	-22.47	CS	**0.20**	**-0.94**	**0.23**	**-0.54**	**0.22**	**-0.74**
2-5		-4.18	-68.28		**0.30**	**-0.67**	**0.25**	**-0.57**	**0.26**	**-0.64**
6-9		**-23.35**	33.09		**0.35**	**-0.37**	**0.27**	**-0.41**	**0.32**	**-0.38**
2-9		**-19.05**	-50.22		**0.33**	**-0.55**	**0.27**	**-0.52**	**0.30**	**-0.51**
1	EA	1.95	**126.42**	EA	**0.19**	**-0.67**	**0.21**	**-0.29**	**0.22**	**-0.48**
2-5		6.14	57.39		**0.28**	**-0.48**	**0.24**	**-0.33**	**0.26**	**-0.40**
6-9		-15.21	-38.21		**0.35**	-0.21	**0.29**	**-0.18**	**0.34**	-0.15
2-9		**-6.93**	2.54		**0.34**	**-0.33**	**0.27**	**-0.21**	**0.32**	**-0.27**
1	SW	5.85	75.75	MT	**0.34**	**-0.59**	**0.26**	**-0.51**	**0.33**	**-0.53**
2-5		-10.55	-2.04		**0.34**	**-0.57**	**0.21**	**-0.54**	**0.30**	**-0.53**
6-9		-1.37	**-42.10**		**0.37**	**-0.55**	**0.25**	**-0.37**	**0.35**	**-0.42**
2-9		-4.94	-34.33		**0.38**	**-0.57**	**0.25**	**-0.47**	**0.34**	**-0.49**
1	CN	-45.93	**143.67**	WI	**0.29**	**-0.56**	**0.23**	**-0.31**	**0.29**	**-0.40**
2-5		-26.01	**155.54**		**0.34**	**-0.39**	**0.16**	**-0.43**	**0.28**	**-0.36**
6-9		-14.66	**133.11**		**0.34**	-0.39	**0.21**	**-0.22**	**0.32**	-0.24
2-9		-13.58	**143.55**		**0.36**	**-0.39**	**0.20**	**-0.32**	**0.30**	**-0.32**

Table A.5: As Table A.2 but in JJA.

Lead years	**Region** (PR)	**PR** (mm/decade)		**Region** (NSAT)	T_{max} (K/decade)		T_{min} (K/decade)		T_{mean} (K/decade)	
		β^{Y}	β^{Y-X}		β^{Y}	β^{Y-X}	β^{Y}	β^{Y-X}	β^{Y}	β^{Y-X}
1	EI	**-40.70**	10.69	SW	**0.32**	0.12	**0.20**	-0.22	**0.25**	-0.09
2-5		**-15.33**	26.99		0.34	0.11	**0.28**	**-0.23**	**0.29**	-0.05
6-9		**-11.21**	14.55		**0.39**	-0.02	**0.38**	**-0.26**	**0.39**	-0.13
2-9		**-13.20**	19.89		**0.36**	0.02	**0.32**	**-0.26**	**0.34**	**-0.12**
1	WI	**-27.91**	19.16	NO	0.19	-0.03	**0.08**	**-0.32**	**0.18**	-0.12
2-5		-8.17	**25.62**		**0.32**	**0.11**	**0.23**	**-0.20**	**0.28**	-0.04
6-9		-6.15	12.88		**0.37**	0.04	**0.36**	-0.10	**0.37**	-0.03
2-9		**-7.33**	**27.69**		**0.33**	0.03	**0.29**	**-0.16**	**0.31**	-0.05
1	NE	-52.42	-1.38	CI	**0.30**	-0.12	0.16	**-0.42**	**0.24**	-0.24
2-5		**-21.53**	57.30		**0.36**	-0.00	**0.25**	**-0.32**	**0.30**	-0.16
6-9		-6.78	**42.98**		**0.42**	0.07	**0.42**	**-0.24**	**0.45**	-0.05
2-9		**-12.41**	**39.19**		**0.39**	0.00	**0.33**	**-0.30**	**0.36**	-0.13
1	CS	**-22.83**	10.12	NE	0.24	-0.12	**0.21**	-0.21	0.28	-0.10
2-5		**-6.73**	**32.68**		**0.32**	-0.10	**0.26**	**-0.26**	**0.30**	-0.13
6-9		**-7.32**	**32.63**		**0.42**	-0.01	**0.41**	-0.13	**0.42**	-0.08
2-9		**-6.06**	**32.42**		**0.38**	-0.07	**0.33**	**-0.23**	**0.36**	-0.18
1	NW	-1.08	21.47	CS	**0.31**	-0.30	**0.22**	**-0.48**	**0.30**	**-0.37**
2-5		**-20.68**	-8.06		**0.38**	**-0.31**	**0.30**	**-0.71**	**0.34**	**-0.53**
6-9		-8.17	4.28		**0.43**	**-0.33**	**0.41**	**-0.63**	**0.43**	**-0.50**
2-9		**-13.28**	-7.44		**0.41**	**-0.34**	**0.35**	**-0.62**	**0.38**	**-0.45**
1	EA	-21.51	18.87	EA	**0.23**	-0.26	**0.21**	-0.17	**0.21**	-0.24
2-5		-5.61	**36.64**		**0.38**	-0.18	**0.29**	**-0.16**	**0.33**	-0.15
6-9		-5.67	**15.48**		**0.39**	-0.15	**0.39**	-0.10	**0.42**	-0.05
2-9		**-6.99**	**29.81**		**0.39**	-0.17	**0.34**	**-0.17**	**0.38**	-0.17
1	SW	**-9.72**	**20.22**	MT	0.25	-0.08	0.13	**-0.17**	0.21	-0.16
2-5		-1.53	**28.19**		**0.34**	0.05	**0.24**	**-0.26**	**0.27**	-0.09
6-9		**-4.25**	**26.01**		**0.40**	0.03	**0.37**	-0.16	**0.38**	-0.08
2-9		-2.35	**23.36**		**0.36**	-0.02	**0.30**	**-0.23**	**0.32**	-0.14
1	CN	-13.56	28.77	WI	0.24	-0.02	0.08	-0.20	0.18	-0.11
2-5		**-36.51**	7.77		**0.36**	0.12	**0.22**	**-0.18**	**0.29**	-0.03
6-9		-20.58	42.73		0.41	0.12	**0.33**	-0.09	**0.36**	-0.01
2-9		**-24.18**	**30.39**		**0.35**	0.06	**0.27**	-0.15	**0.30**	-0.05

Table A.6: As Table A.2 but in SON.

Lead years	Region (PR)	PR (mm/decade)		Region (NSAT)	T_{max} (K/decade)		T_{min} (K/decade)		T_{mean} (K/decade)	
		β^Y	β^{Y-X}		β^Y	β^{Y-X}	β^Y	β^{Y-X}	β^Y	β^{Y-X}
1	EI	0.07	**-65.98**	SW	**-0.12**	0.23	**0.19**	-0.07	**0.24**	0.18
2-5		-2.22	**-79.22**		**0.14**	**0.58**	**0.18**	-0.22	**0.14**	0.16
6-9		**-15.79**	**-77.09**		**-0.04**	**0.50**	**0.23**	0.12	**0.15**	0.38
2-9		**-7.67**	**-77.95**		**0.06**	**0.43**	**0.21**	0.01	**0.15**	**0.23**
1	WI	-39.60	**-129.01**	NO	**0.29**	0.27	**0.22**	-0.19	**0.24**	0.06
2-5		-8.70	**-114.29**		**0.23**	0.29	**0.17**	-0.15	**0.19**	0.04
6-9		-11.54	**-120.80**		**0.13**	0.43	**0.25**	0.08	**0.21**	0.29
2-9		**-12.44**	**-113.62**		**0.19**	**0.32**	**0.22**	0.01	**0.21**	0.19
1	NE	5.64	-105.52	CI	**-0.20**	-0.01	**0.18**	-0.26	**0.23**	0.13
2-5		**-9.36**	**-159.03**		**0.13**	0.43	**0.12**	-0.35	**0.14**	0.08
6-9		**-45.18**	**-107.32**		0.02	0.54	**0.20**	0.07	**0.12**	0.30
2-9		**-28.23**	**-116.43**		**0.07**	**0.39**	**0.18**	-0.06	**0.13**	0.17
1	CS	**-6.89**	-71.87	NE	**0.22**	0.12	**0.19**	-0.44	**0.20**	-0.15
2-5		-3.45	**-89.54**		**0.19**	0.10	**0.12**	-0.31	**0.14**	-0.08
6-9		**-24.94**	**-64.43**		**0.20**	0.38	**0.24**	0.06	**0.26**	0.23
2-9		**-14.40**	**-75.70**		**0.20**	0.17	**0.19**	-0.09	**0.21**	0.08
1	NW	-44.38	**-341.90**	CS	**0.14**	0.02	**0.23**	-0.24	**0.28**	-0.01
2-5		-33.88	**-255.18**		**0.17**	0.05	**0.14**	**-0.43**	**0.15**	-0.17
6-9		5.50	**-263.38**		**0.16**	0.26	**0.22**	0.01	**0.21**	0.16
2-9		-14.66	**-278.92**		**0.15**	0.07	**0.19**	-0.19	**0.18**	-0.02
1	EA	**-16.18**	-29.51	EA	**0.08**	-0.10	**0.22**	-0.19	**0.26**	0.01
2-5		0.82	**-60.94**		**0.10**	0.09	**0.11**	-0.27	**0.12**	-0.05
6-9		**-25.06**	-61.31		**0.11**	0.33	**0.23**	0.09	**0.17**	0.22
2-9		-9.44	-54.38		**0.10**	0.18	**0.17**	-0.05	**0.14**	0.05
1	SW	**-24.13**	-109.21	MT	**0.04**	0.19	**0.20**	-0.20	**0.23**	0.05
2-5		9.01	**-136.35**		**0.14**	0.38	**0.18**	-0.13	**0.18**	0.14
6-9		**-31.61**	**-93.38**		**0.07**	0.45	**0.25**	0.18	**0.17**	0.30
2-9		-12.47	**-114.17**		**0.12**	**0.30**	**0.22**	0.05	**0.19**	0.19
1	CN	**-54.09**	-112.47	WI	0.03	0.24	**0.18**	-0.18	**0.23**	0.16
2-5		**-25.29**	-47.92		**0.15**	0.60	**0.14**	-0.22	**0.15**	0.16
6-9		3.78	-75.08		**0.04**	**0.62**	**0.22**	0.18	**0.10**	0.36
2-9		-11.00	**-66.22**		0.09	**0.49**	**0.18**	0.04	0.13	0.24

B

Supplementary figures

B.1. Retrospective decadal climate predictions for precipitation

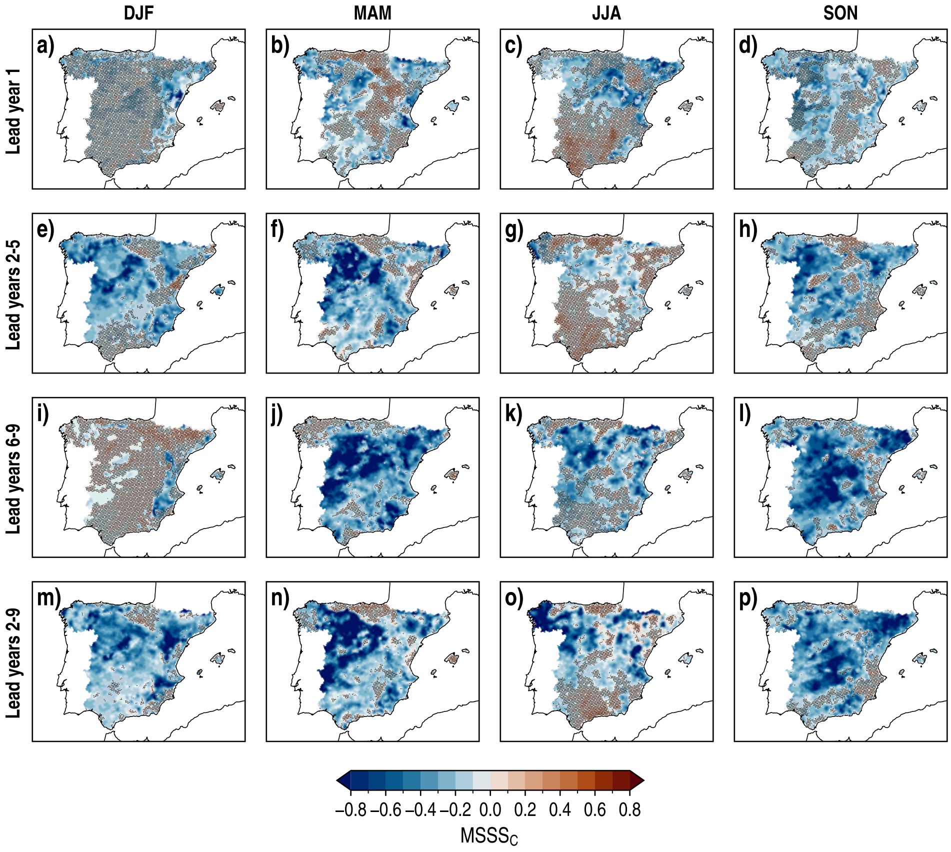

Figure B.1: Spatial distributions of MSSS$_C$, with climatology as reference, for the WRF-DPLE multiannual mean anomalies of PR for lead years 1, 2–5, 6–9 and 2–9 (rows) in DJF, MAM, JJA and SON (columns). The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90% confidence level.

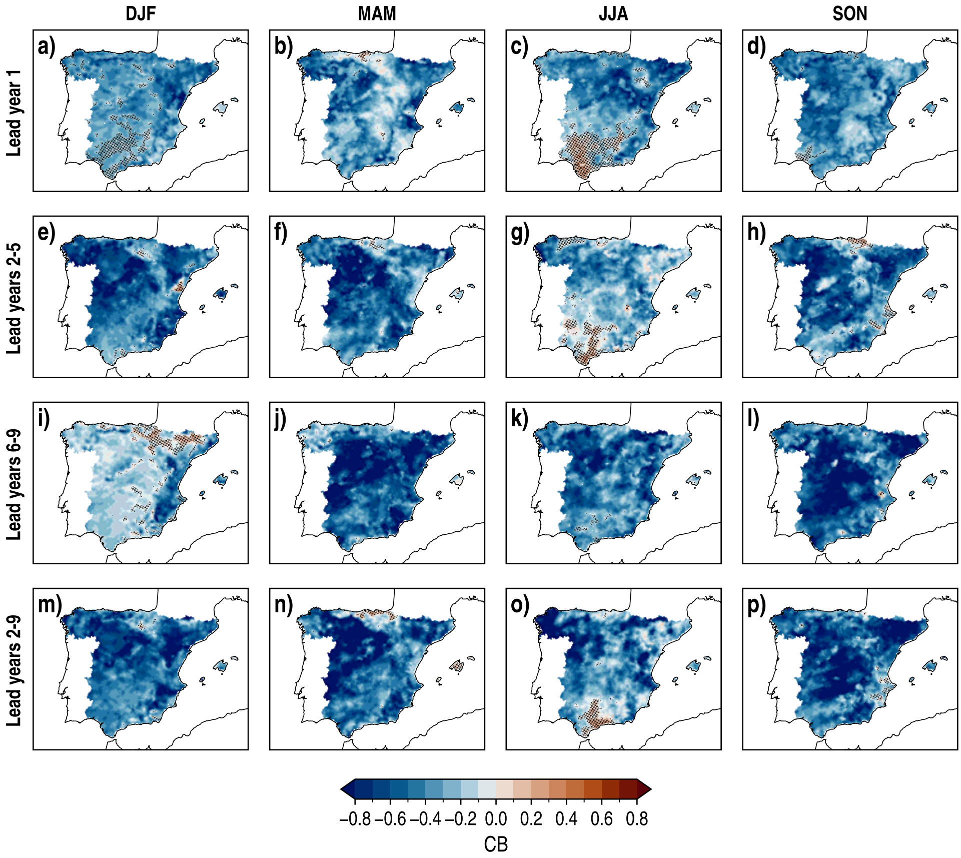

Figure B.2: As Figure B.1 but for CB.

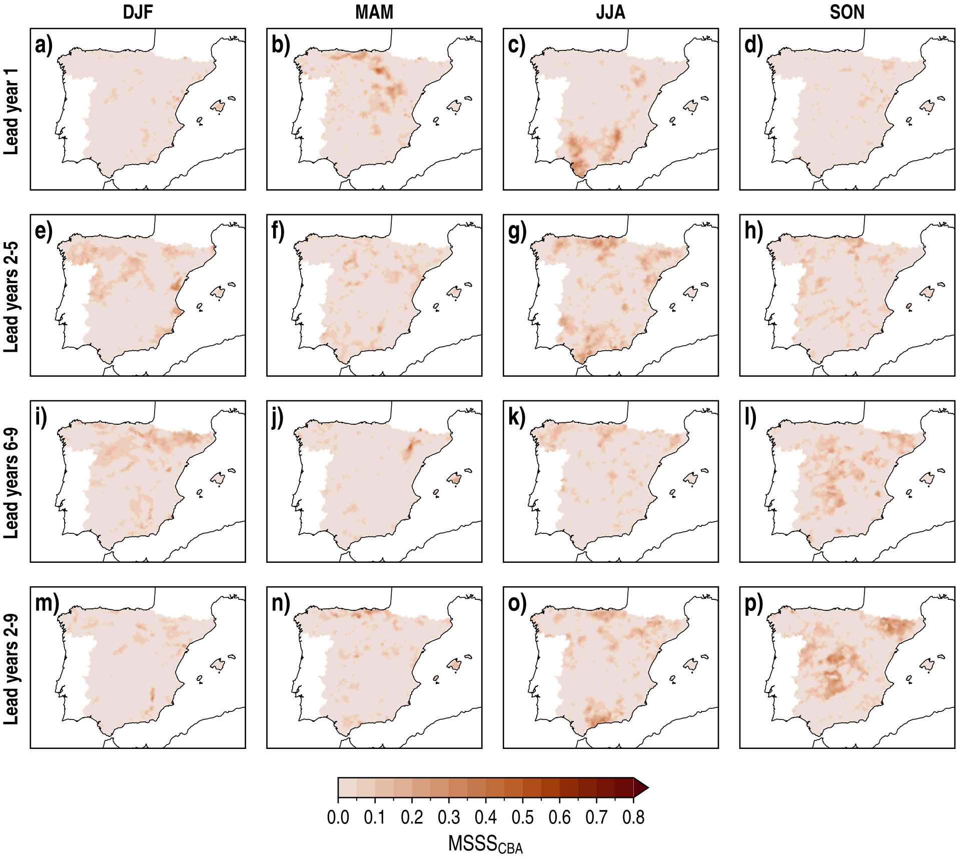

Figure B.3: As Figure B.1 but for MSSS$_{CBA}$ (MSSS$_{C}$ calculated for lead time series with an adjusted CB, i.e., equal to zero).

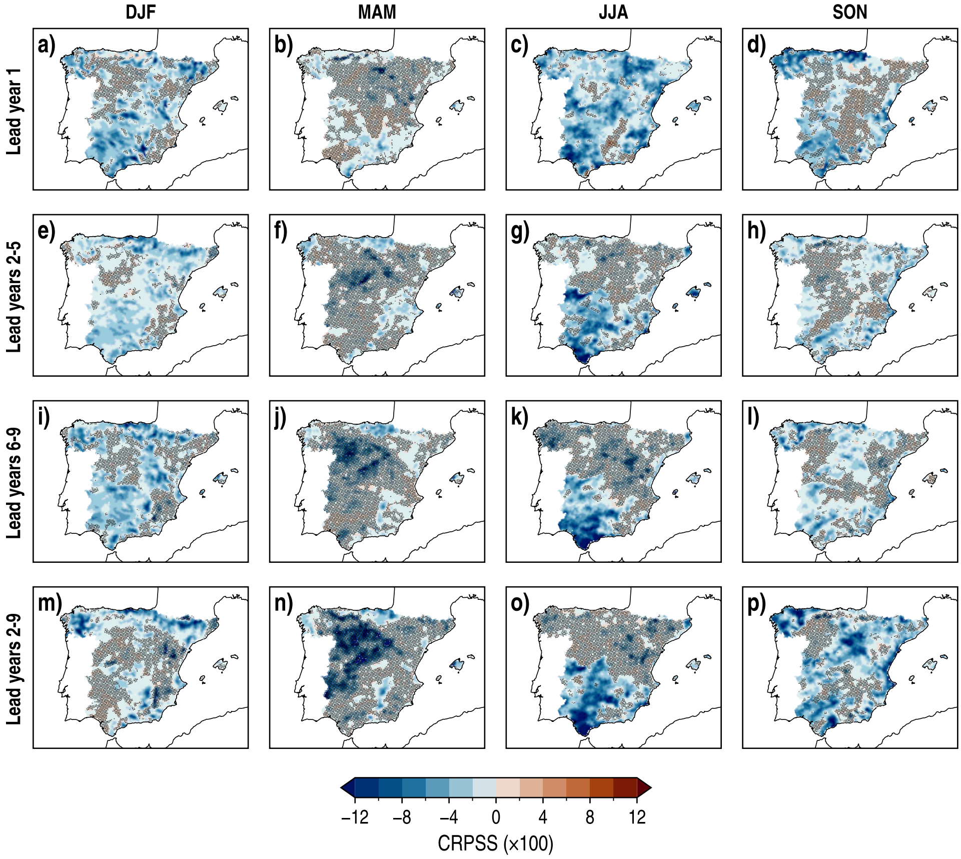

Figure B.4: As Figure B.1 but for CRPSS.

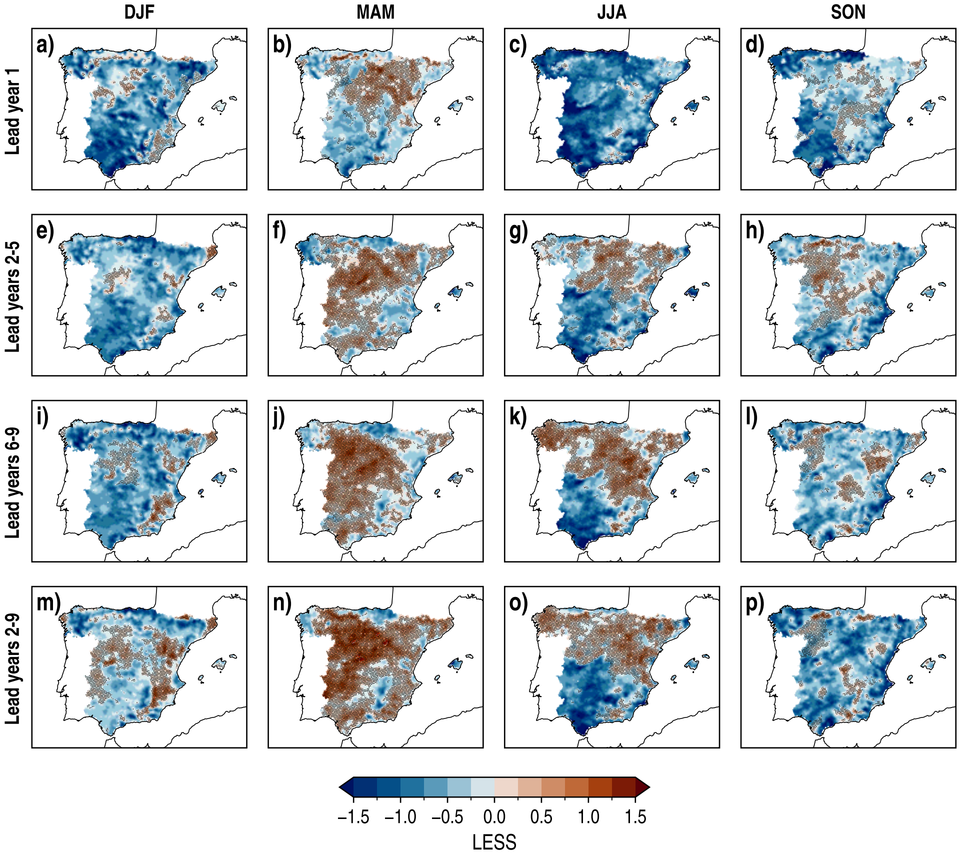

Figure B.5: As Figure B.1 but for LESS.

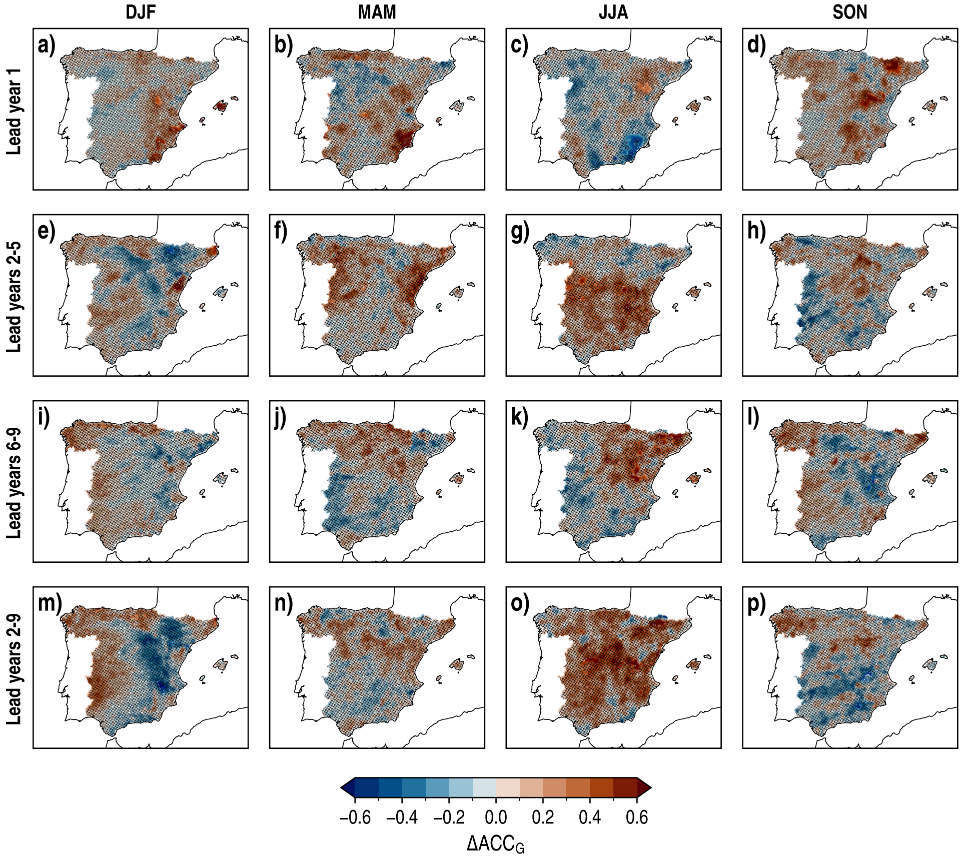

Figure B.6: As Figure B.1 but for $\Delta\text{ACC}_\text{G}$, with CESM-DPLE as reference.

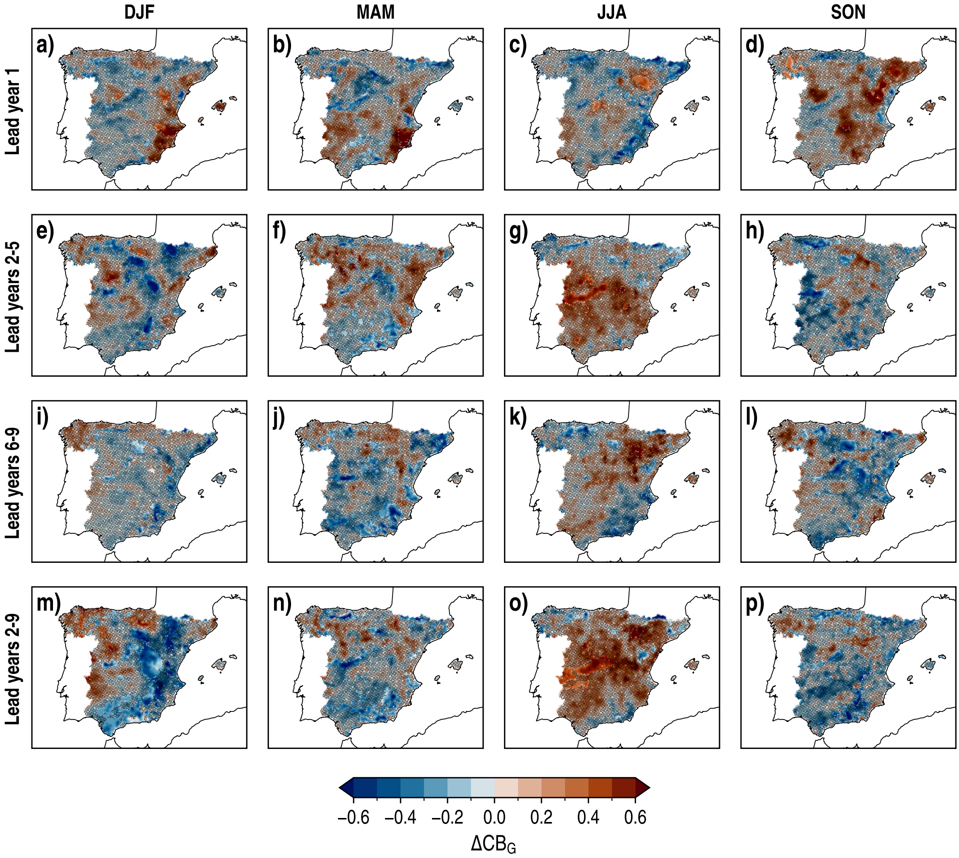

Figure B.7: As Figure B.1 but for ΔCB_G, with CESM-DPLE as reference.

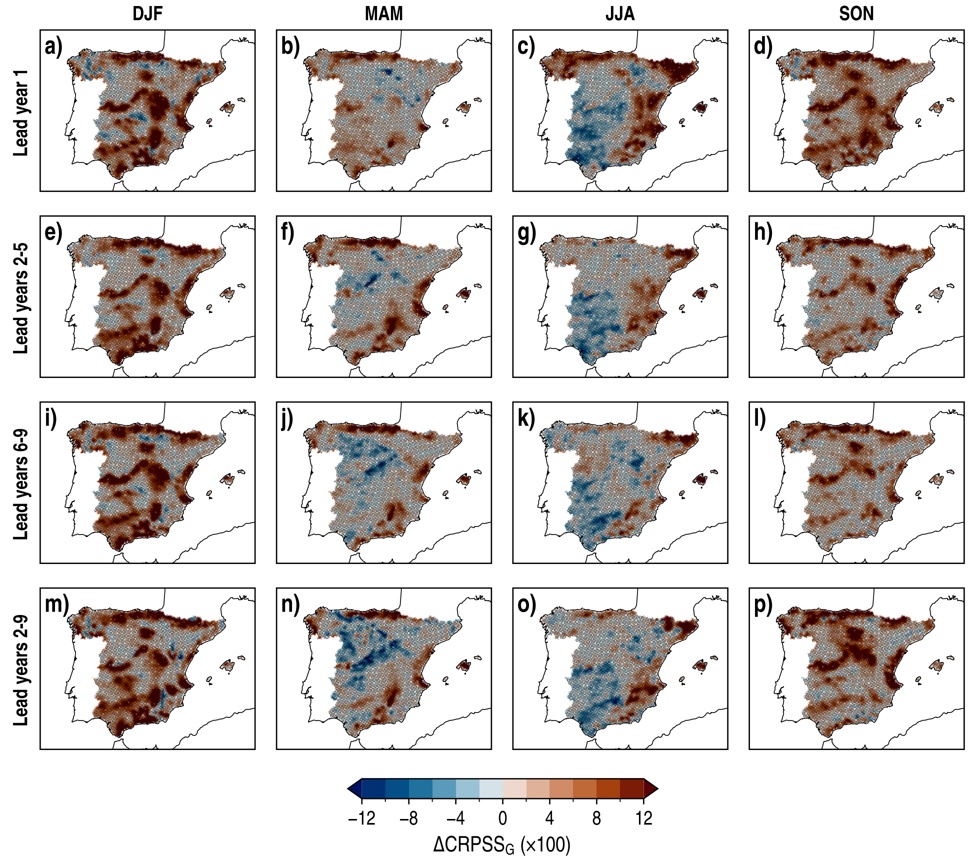

Figure B.8: As Figure B.1 but for ΔCRPSS$_G$, with CESM-DPLE as reference.

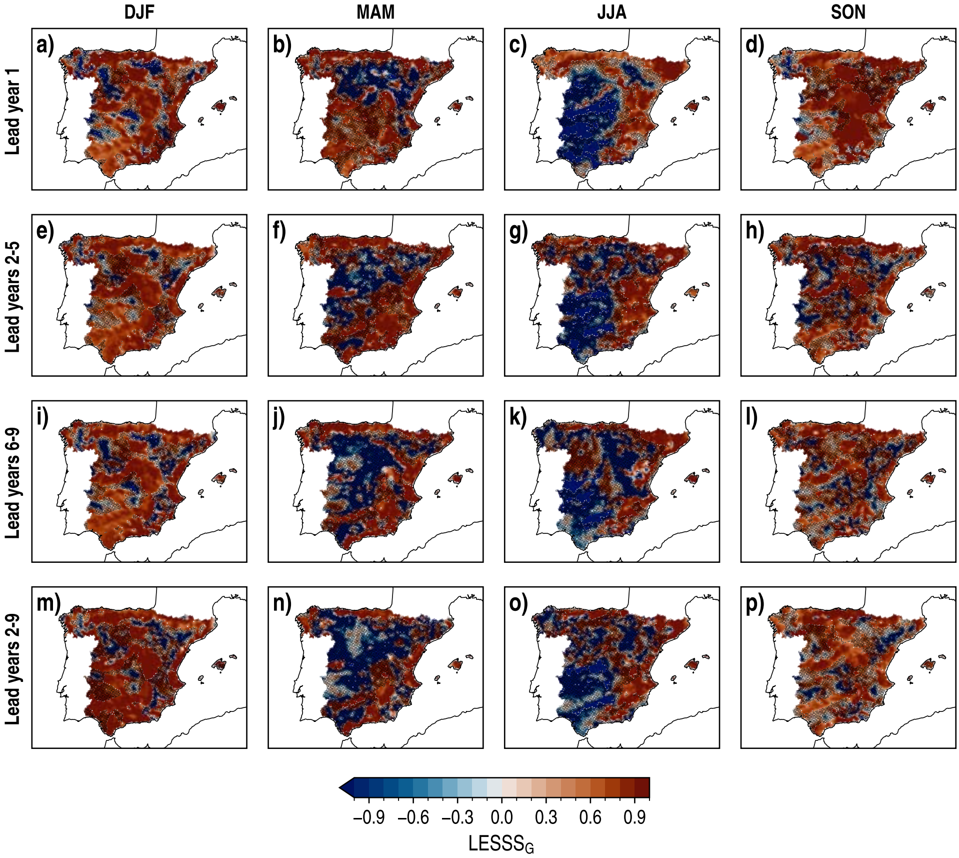

Figure B.9: As Figure B.1 but for LESSS$_G$, with CESM-DPLE as reference.

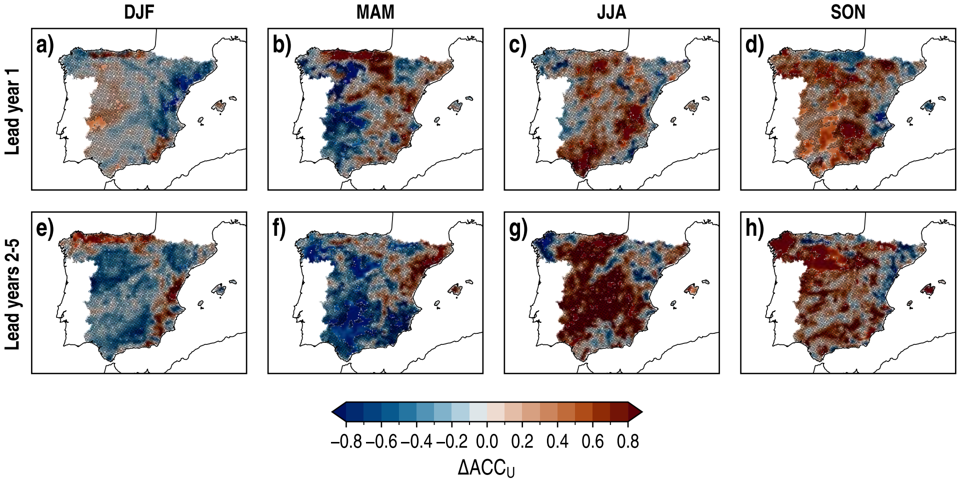

Figure B.10: As Figure B.1 but for $\Delta\mathrm{ACC}_{\mathrm{U}}$, with WRF-LE as reference, only for lead years 1 and 2–5 (rows).

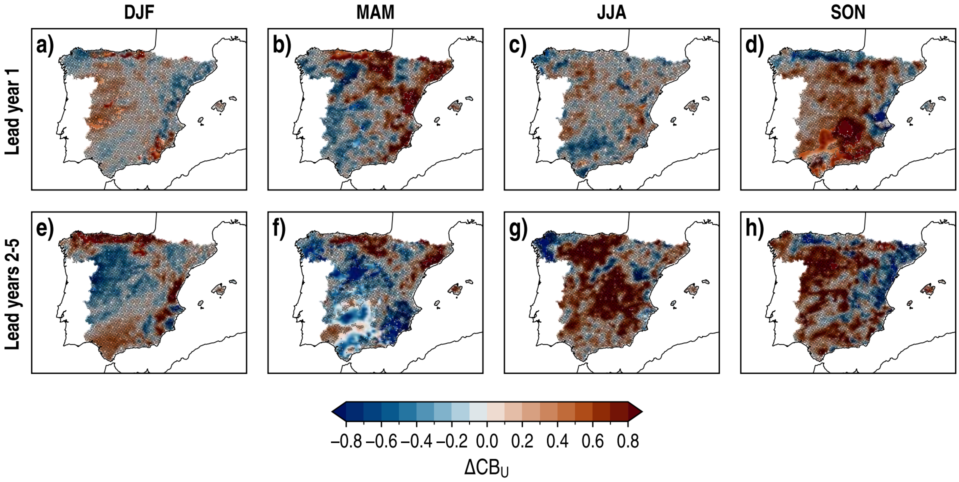

Figure B.11: As Figure B.1 but for $\Delta\mathrm{CB}_{\mathrm{U}}$, with WRF-LE as reference, only for lead years 1 and 2–5 (rows).

B.2. Retrospective decadal climate predictions for near-surface air temperature

B.2.1. *Daily maximum near-surface air temperature*

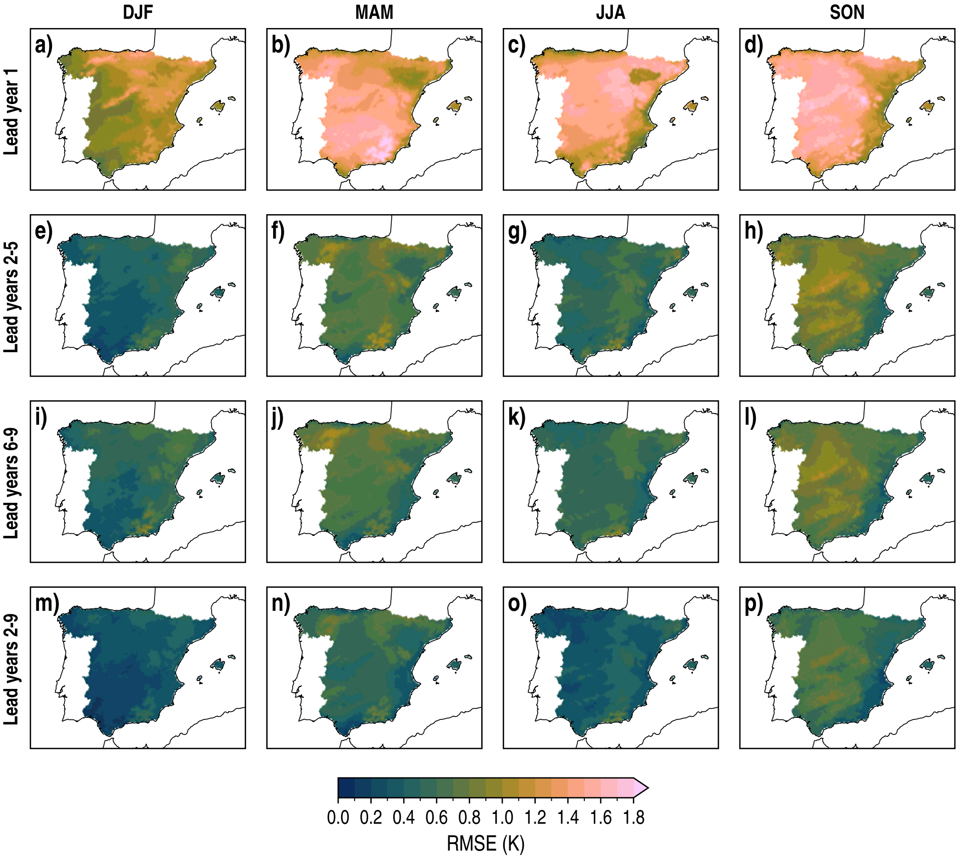

Figure B.12: Spatial distributions of RMSE for the WRF-DPLE multiannual mean anomalies of $T_{\max}$ for lead years 1, 2–5, 6–9 and 2–9 (rows) in DJF, MAM, JJA and SON (columns).

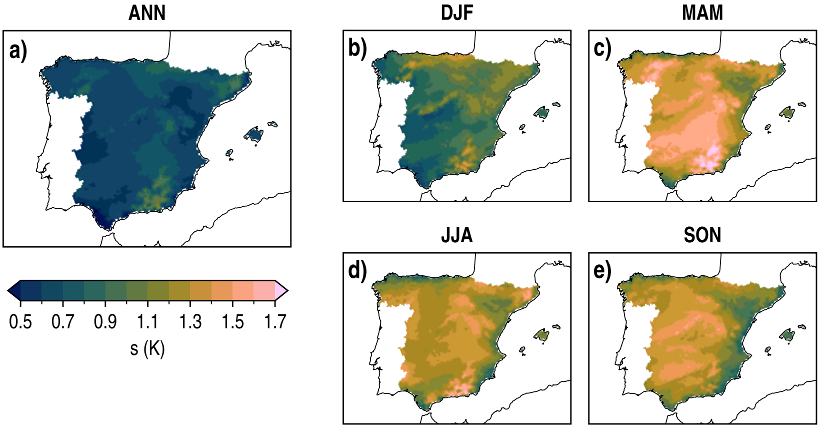

Figure B.13: Spatial distributions of the standard deviation (s) of the AEMET $T_{\max}$ at annual and seasonal scales for the period 1970-2009. While the annual series covers the period from 1970-11 to 2009-10, the seasonal series span the period from 1970-12 to 2009-11.

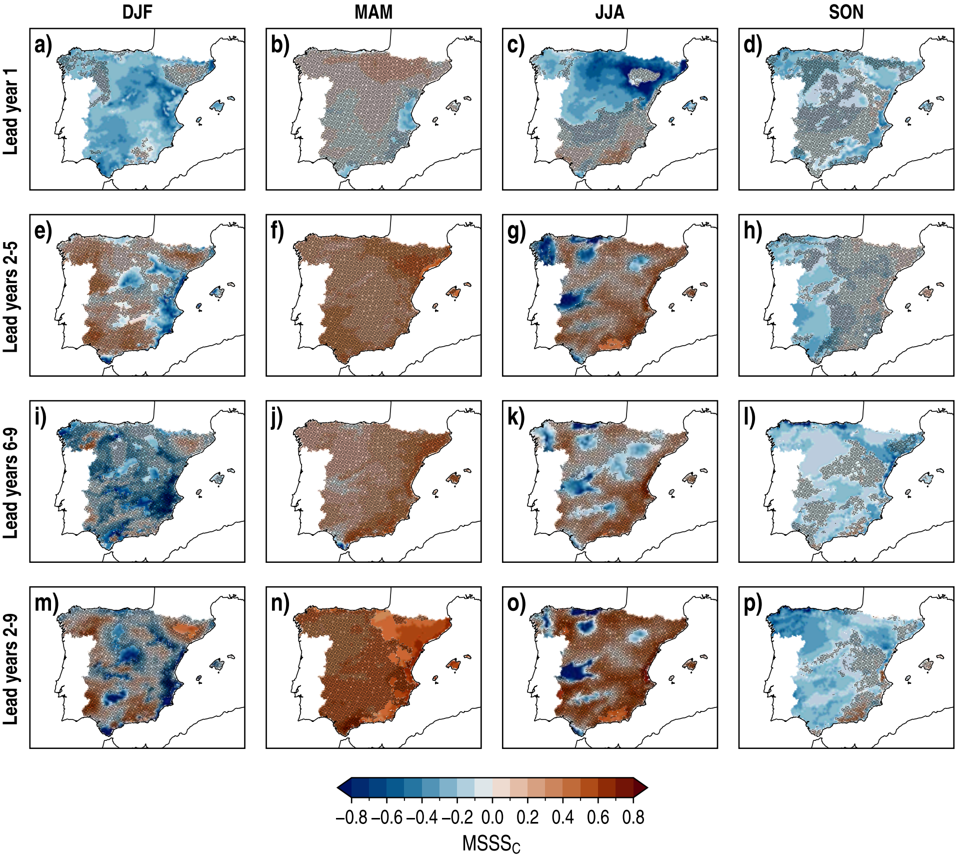

Figure B.14: Spatial distributions of MSSS$_C$, with climatology as reference, for the WRF-DPLE multiannual mean anomalies of T_{max} for lead years 1, 2–5, 6–9 and 2–9 (rows) in DJF, MAM, JJA and SON (columns). The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90% confidence level.

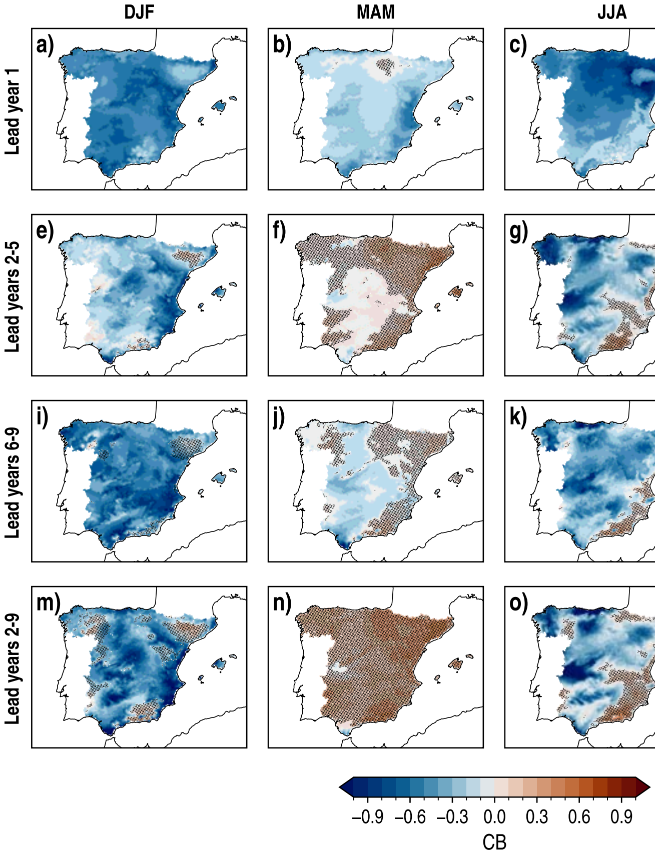

Figure B.15: As Figure B.14 but for CB.

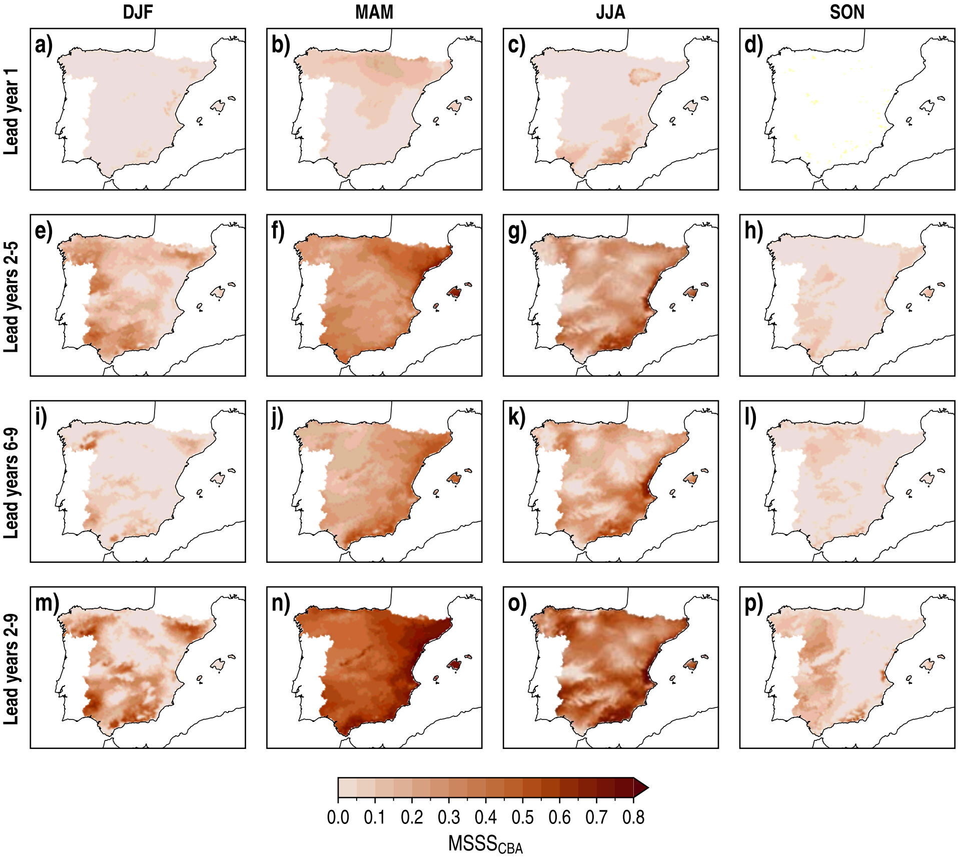

Figure B.16: As Figure B.14 but for MSSS$_{CBA}$ (MSSS$_C$ calculated for lead time series with an adjusted CB, i.e., equal to zero).

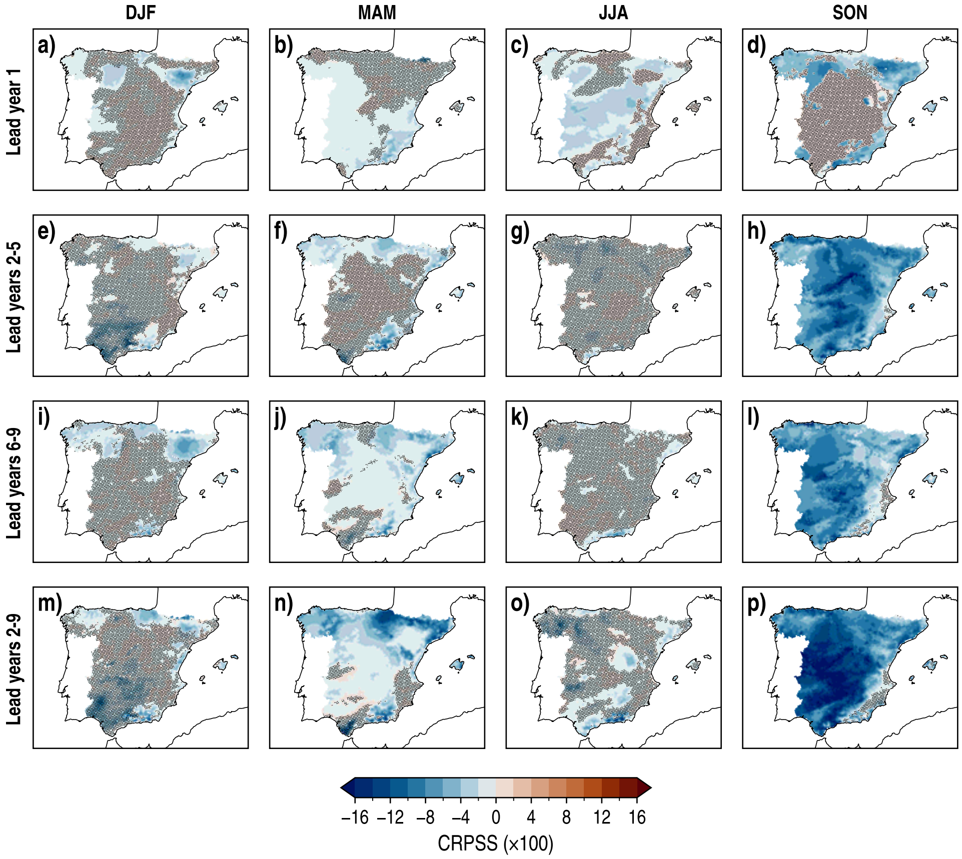

Figure B.17: As Figure B.14 but for CRPSS.

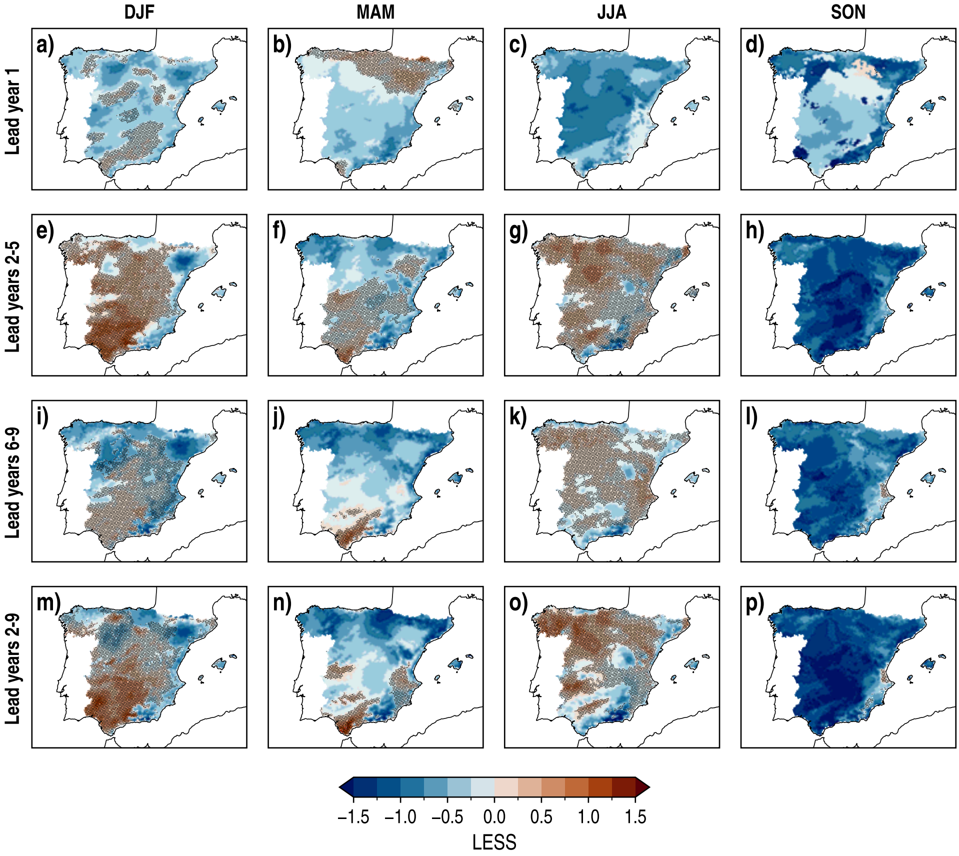

Figure B.18: As Figure B.14 but for LESS.

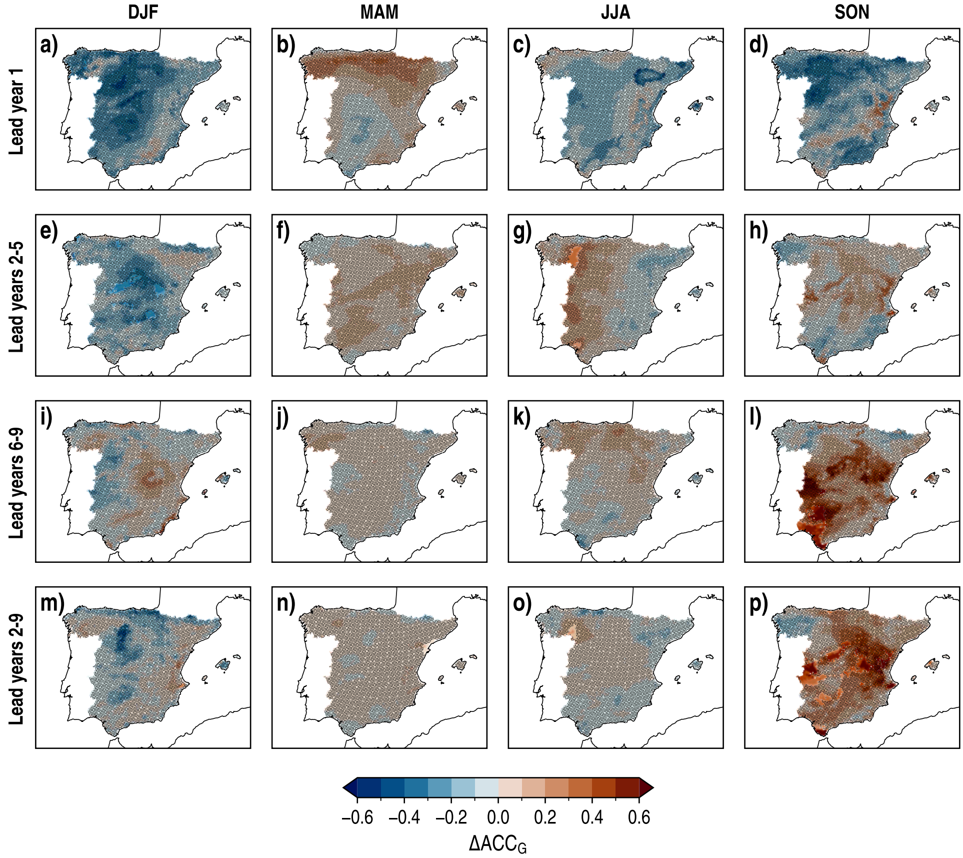

Figure B.19: As Figure B.14 but for $\Delta\mathrm{ACC_G}$, with CESM-DPLE as reference.

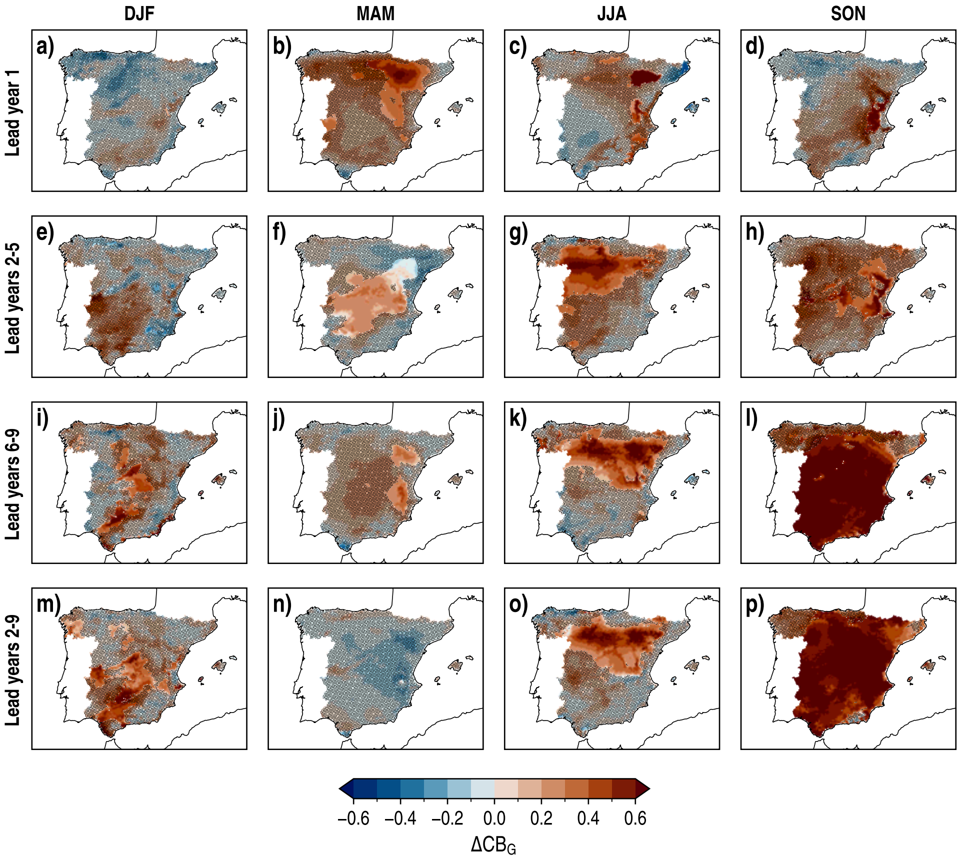

Figure B.20: As Figure B.14 but for ΔCB_G, with CESM-DPLE as reference.

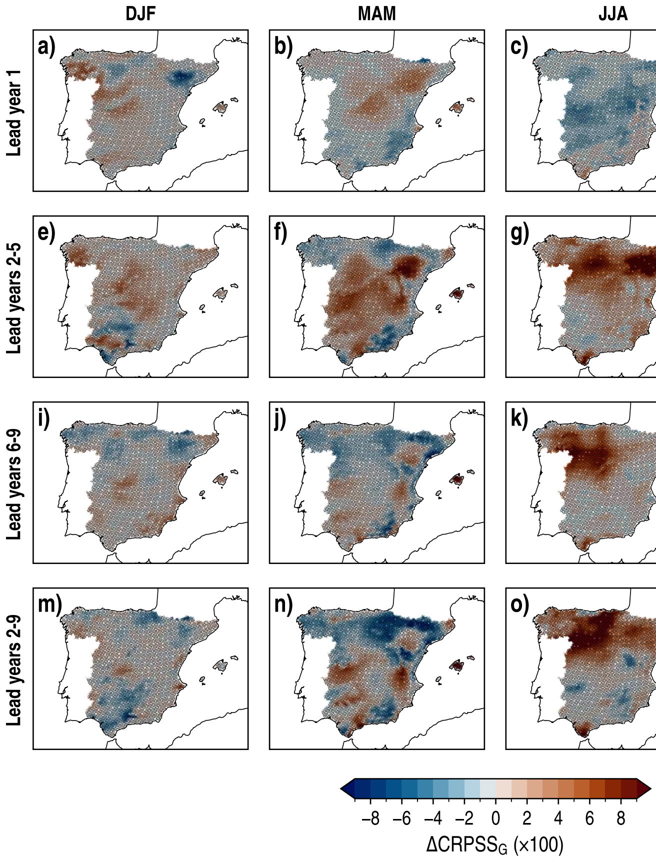

Figure B.21: As Figure B.14 but for ΔCRPSS$_G$, with CESM-DPLE as reference.

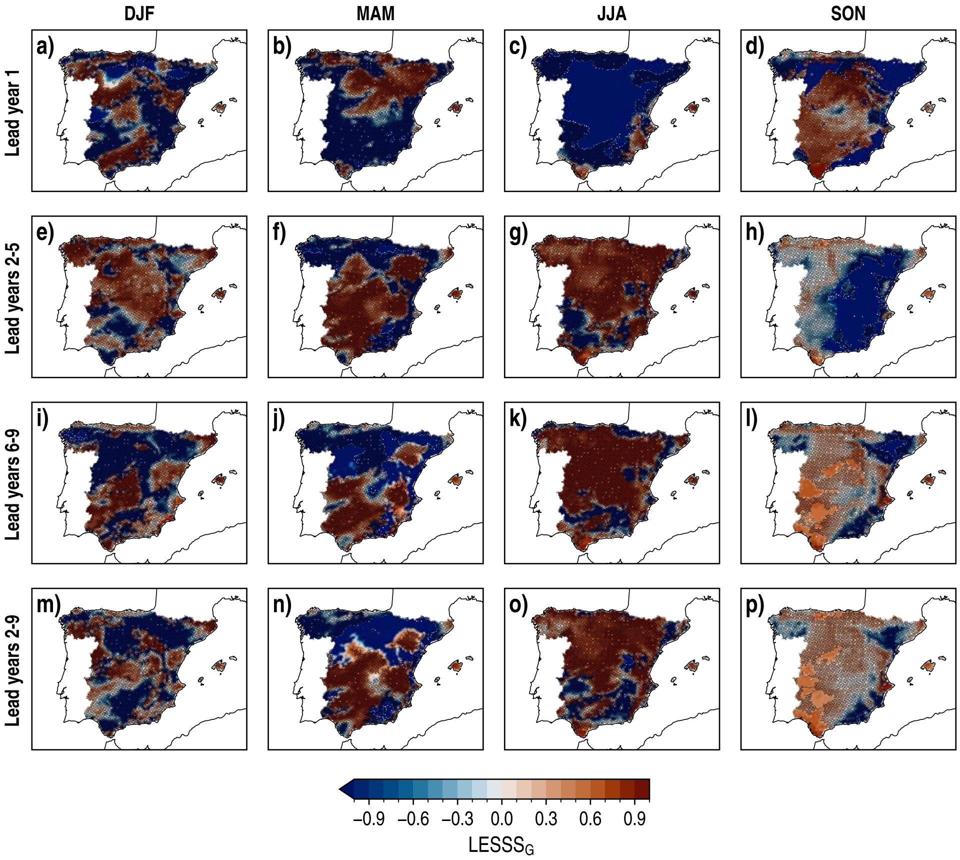

Figure B.22: As Figure B.14 but for $LESSS_G$, with CESM-DPLE as reference.

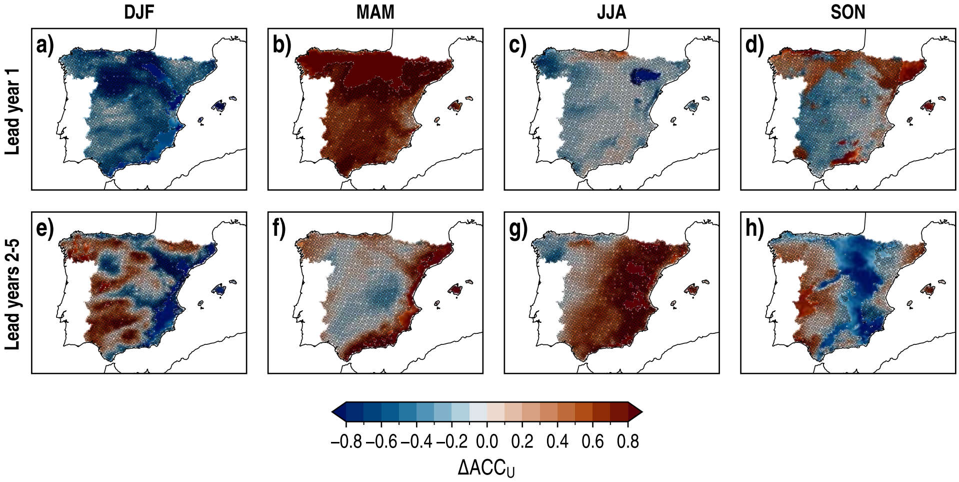

Figure B.23: As Figure B.14 but for ΔACC_U, with WRF-LE as reference, only for lead years 1 and 2–5 (rows).

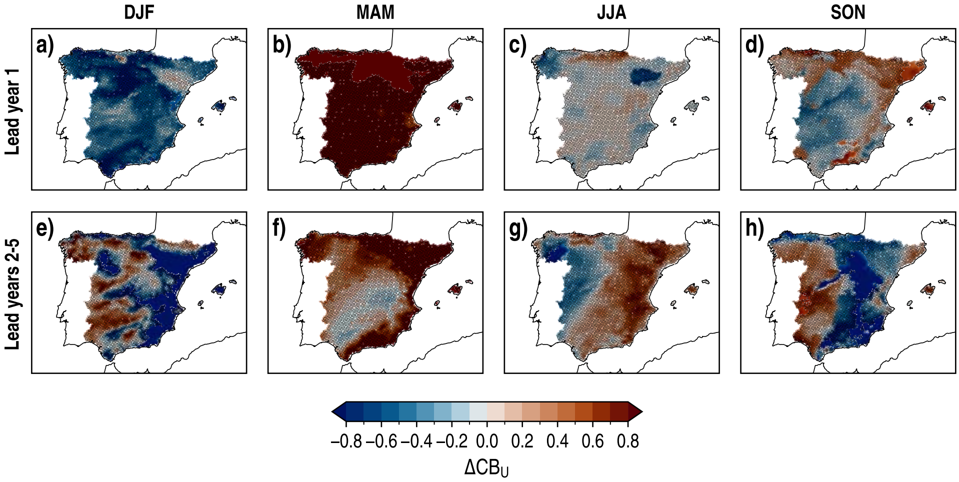

Figure B.24: As Figure B.14 but for ΔCB_U, with WRF-LE as reference, only for lead years 1 and 2–5 (rows).

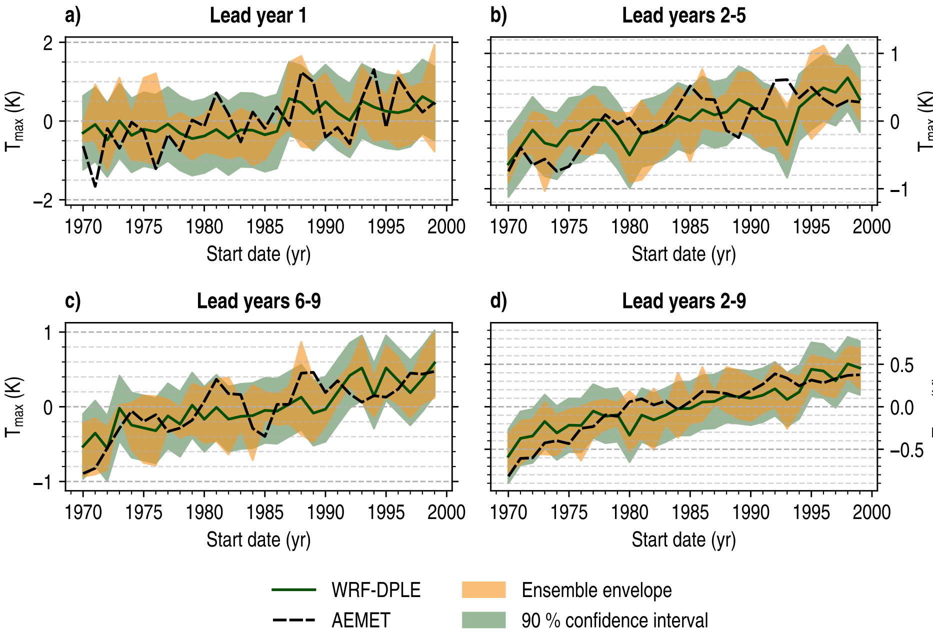

Figure B.25: Time series of the spatially averaged multiannual mean anomalies of $T_{\max}$ in the MT region for lead years 1, 2–5, 6–9 and 2–9 at annual scale. Solid green lines identify the WRF-DPLE ensemble mean, whereas dashed black lines correspond to AEMET. Shaded green surfaces indicate the 90 % confidence interval for a WRF-DPLE single member, calculated from the average ensemble spread (Eq. [3.32]). Shaded yellow surfaces show the ensemble envelope which encloses the trajectories followed by the members composing the WRF-DPLE ensemble.

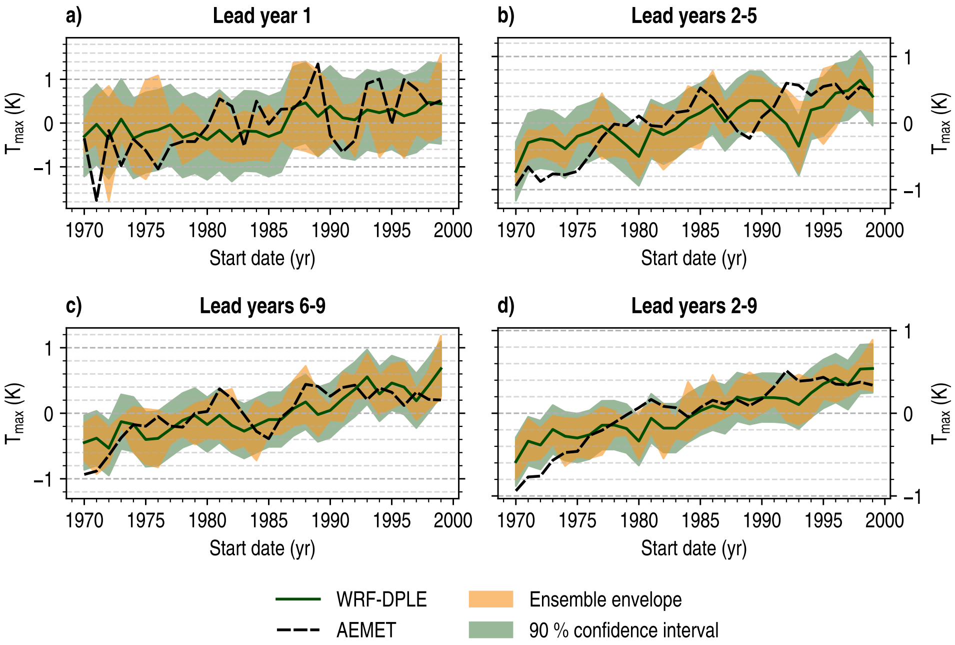

Figure B.26: As Figure B.25 but for the NE region.

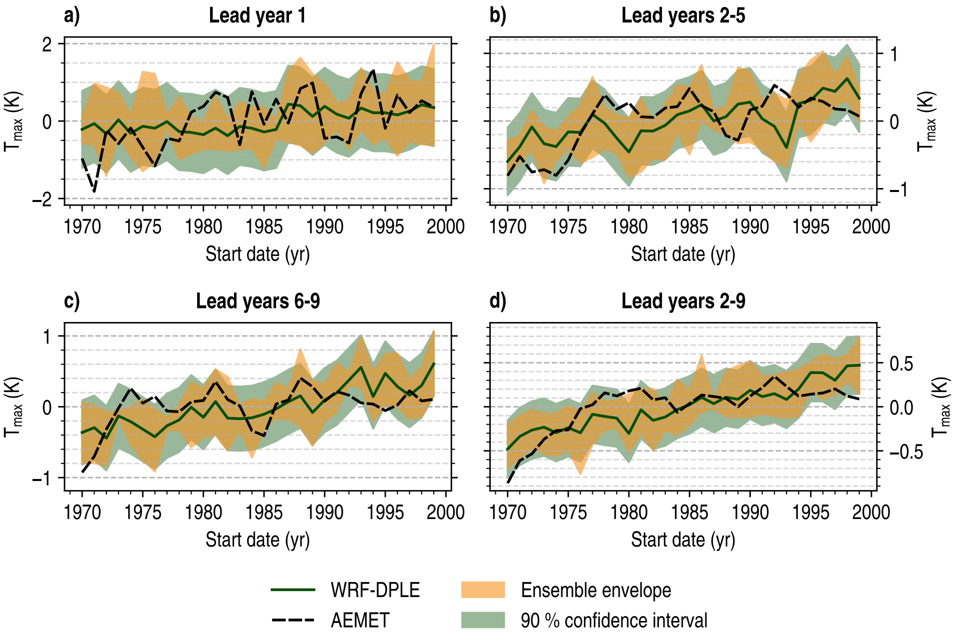

Figure B.27: As Figure B.25 but for the CI region.

B.2.2. *Daily minimum near-surface air temperature*

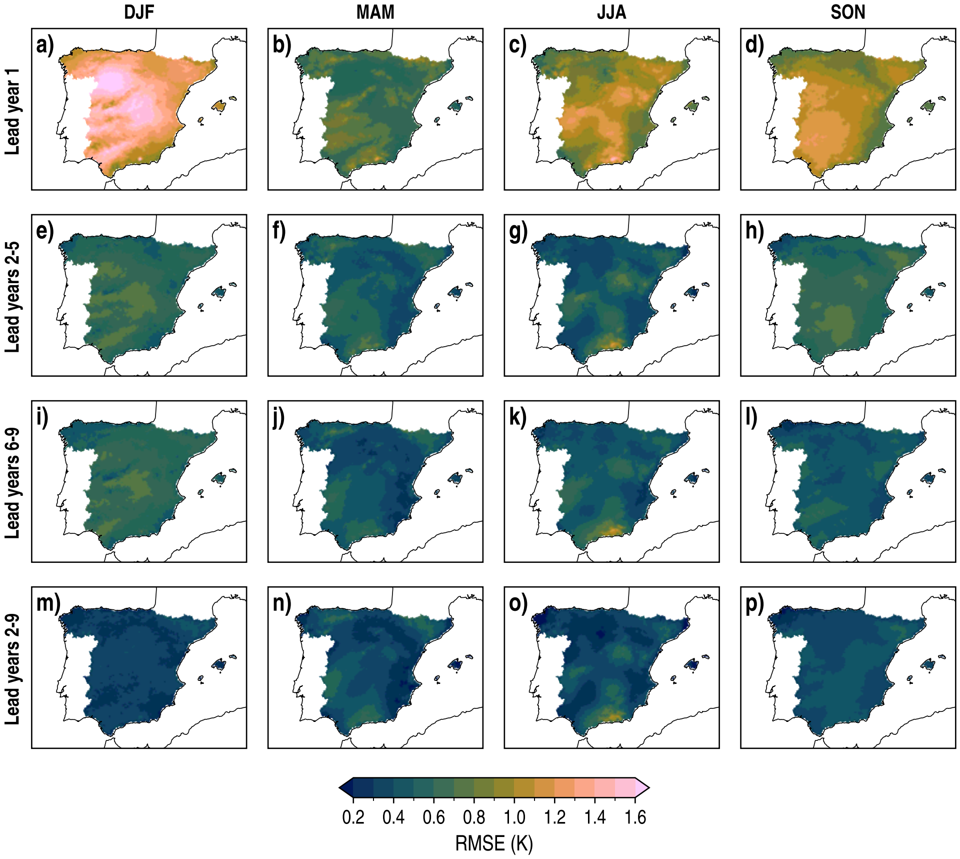

Figure B.28: Spatial distributions of RMSE for the WRF-DPLE multiannual mean anomalies of T_{min} for lead years 1, 2–5, 6–9 and 2–9 (rows) in DJF, MAM, JJA and SON (columns).

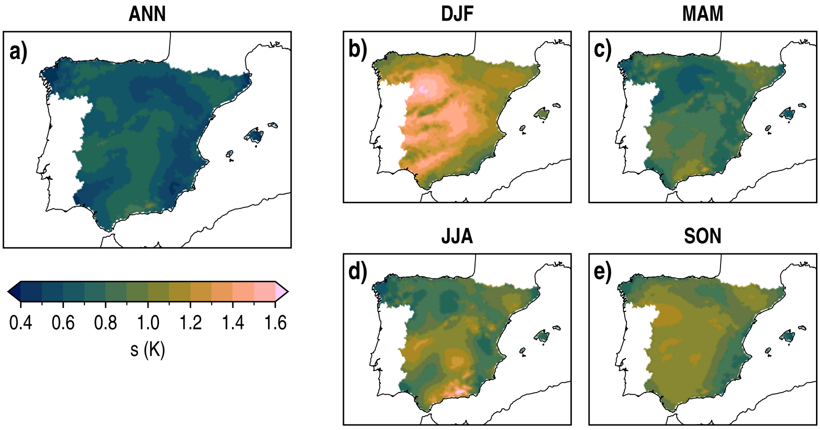

Figure B.29: Spatial distributions of the standard deviation (s) of the AEMET T_{min} at annual and seasonal scales for the period 1970-2009. While the annual series covers the period from 1970-11 to 2009-10, the seasonal series span the period from 1970-12 to 2009-11.

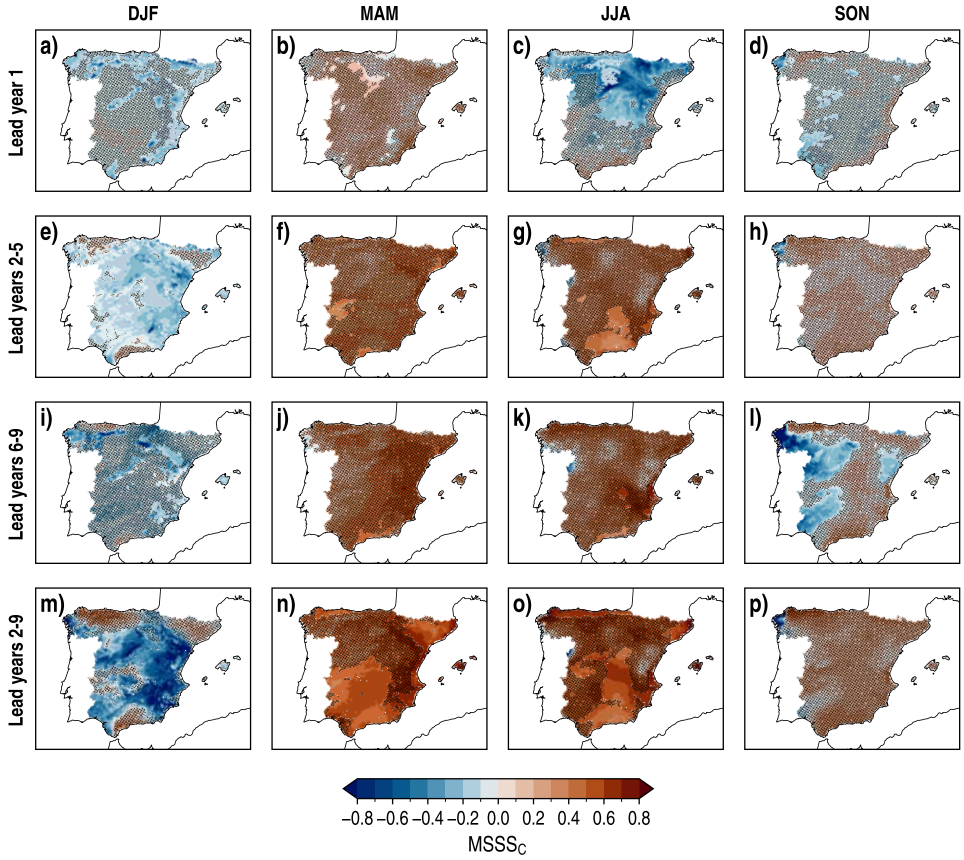

Figure B.30: Spatial distributions of MSSS$_C$, with climatology as reference, for the WRF-DPLE multiannual mean anomalies of $T_{\min}$ for lead years 1, 2–5, 6–9 and 2–9 (rows) in DJF, MAM, JJA and SON (columns). The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90% confidence level.

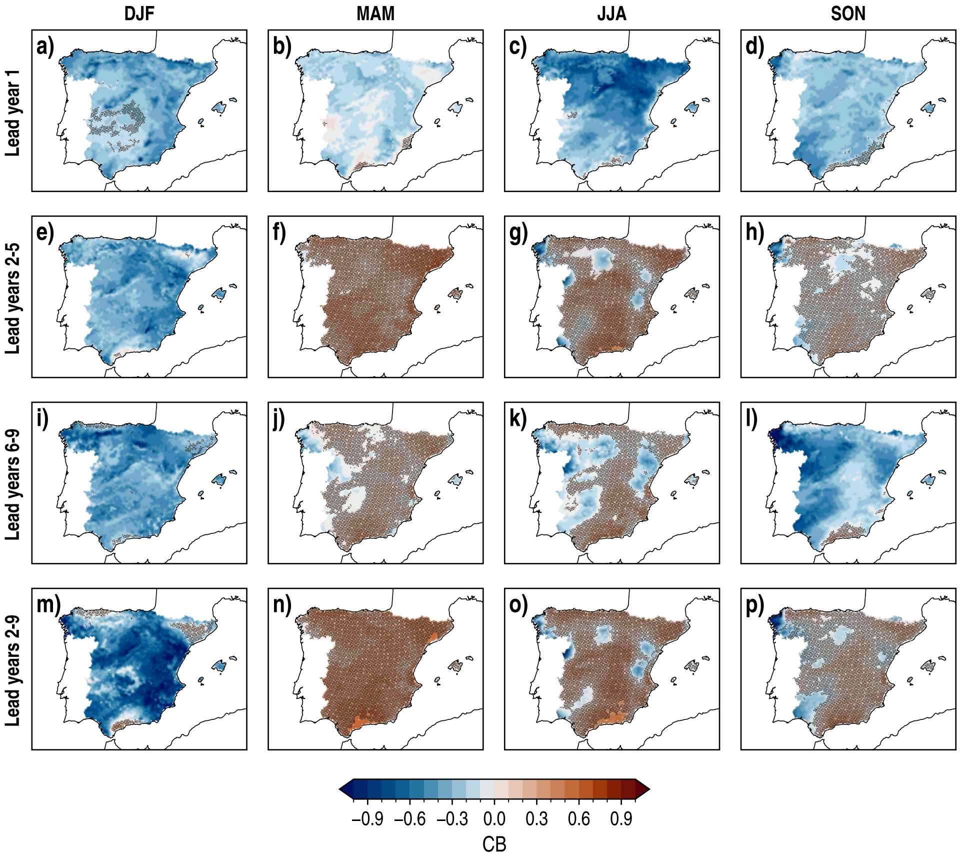

Figure B.31: As Figure B.30 but for CB.

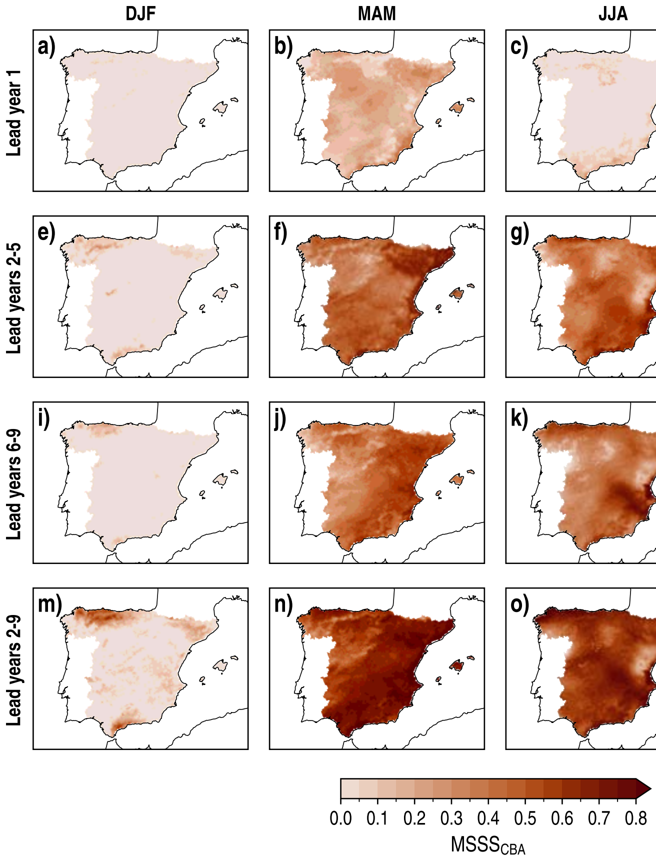

Figure B.32: As Figure B.30 but for MSSS$_{CBA}$ (MSSS$_{C}$ calculated for lead time series with an adjusted CB, i.e., equal to zero).

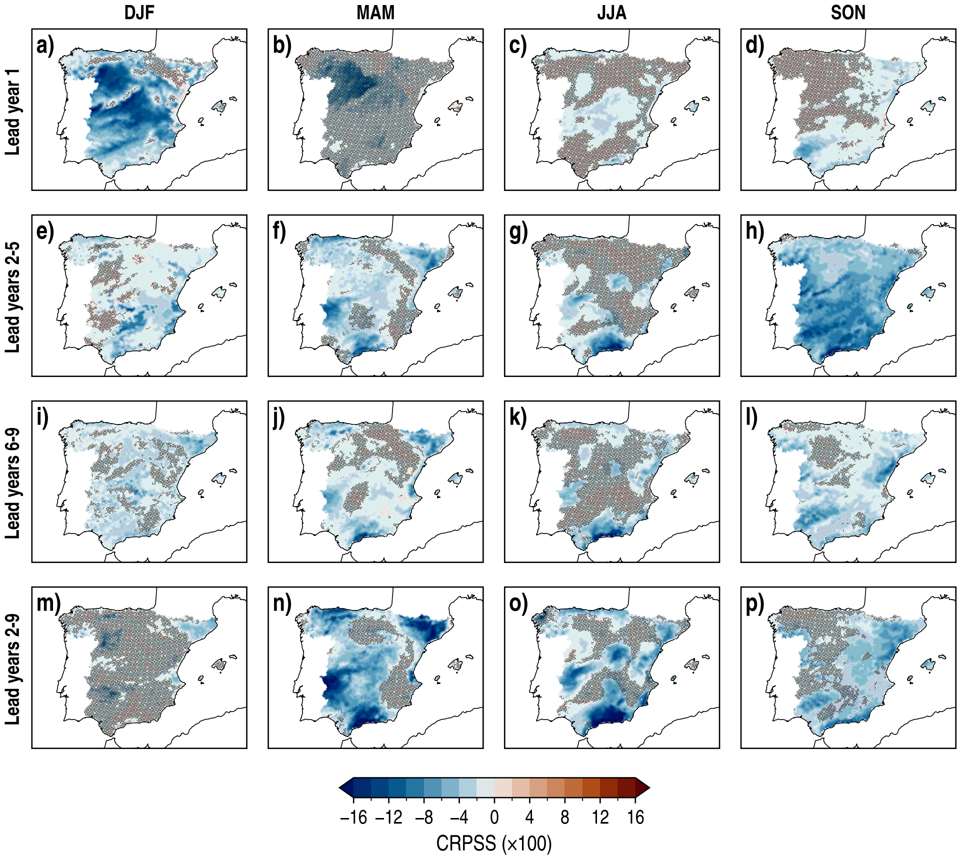

Figure B.33: As Figure B.30 but for CRPSS.

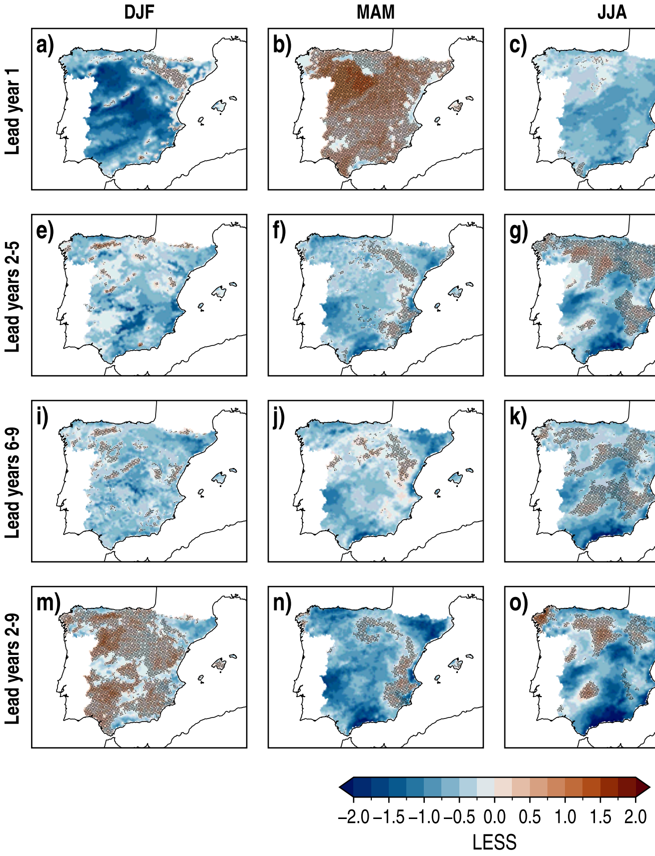

Figure B.34: As Figure B.30 but for LESS.

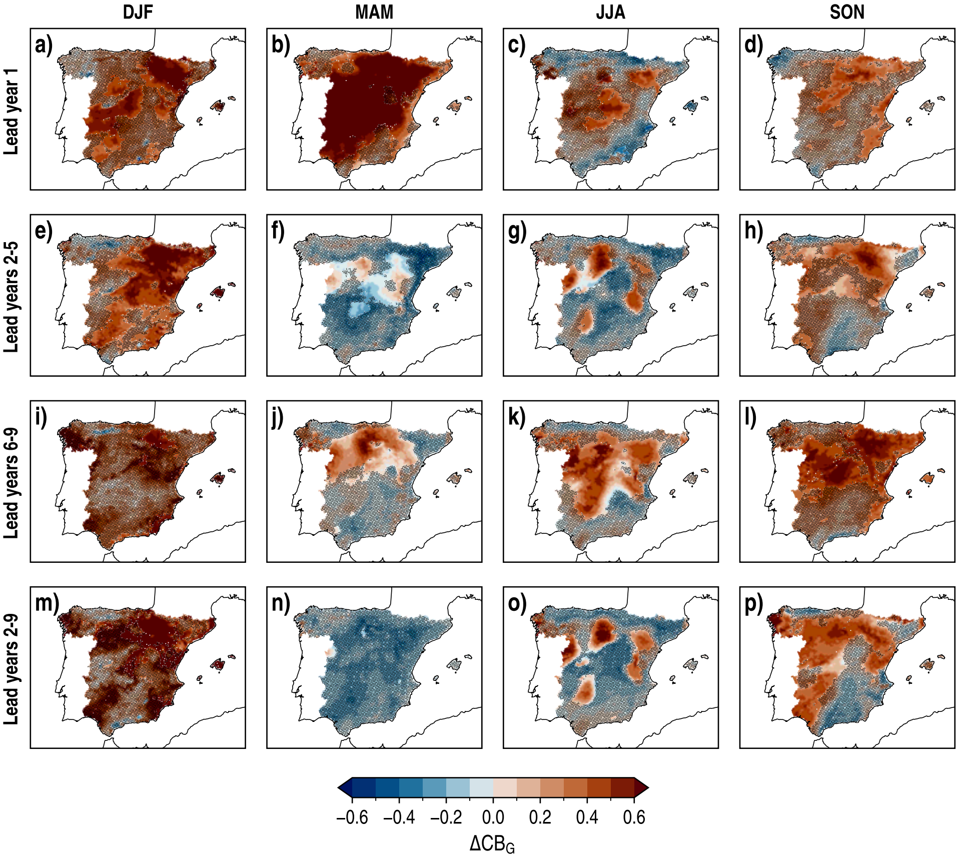

Figure B.35: As Figure B.30 but for ΔCB_G, with CESM-DPLE as reference.

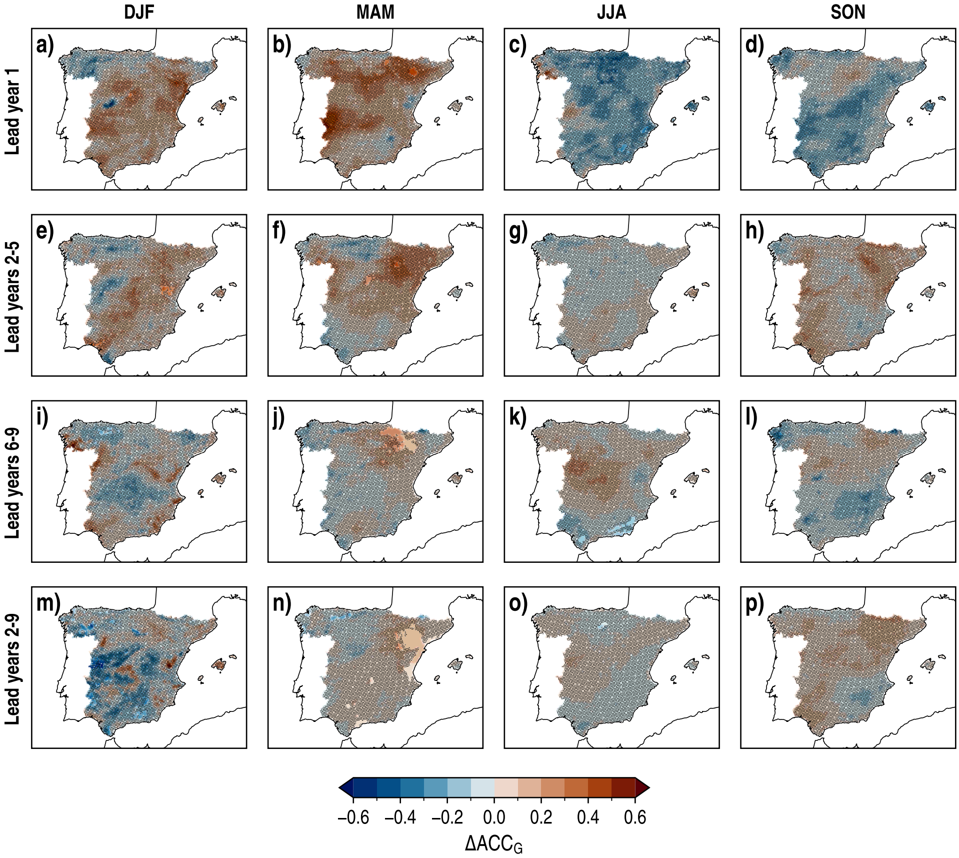

Figure B.36: As Figure B.30 but for ΔACC_G, with CESM-DPLE as reference.

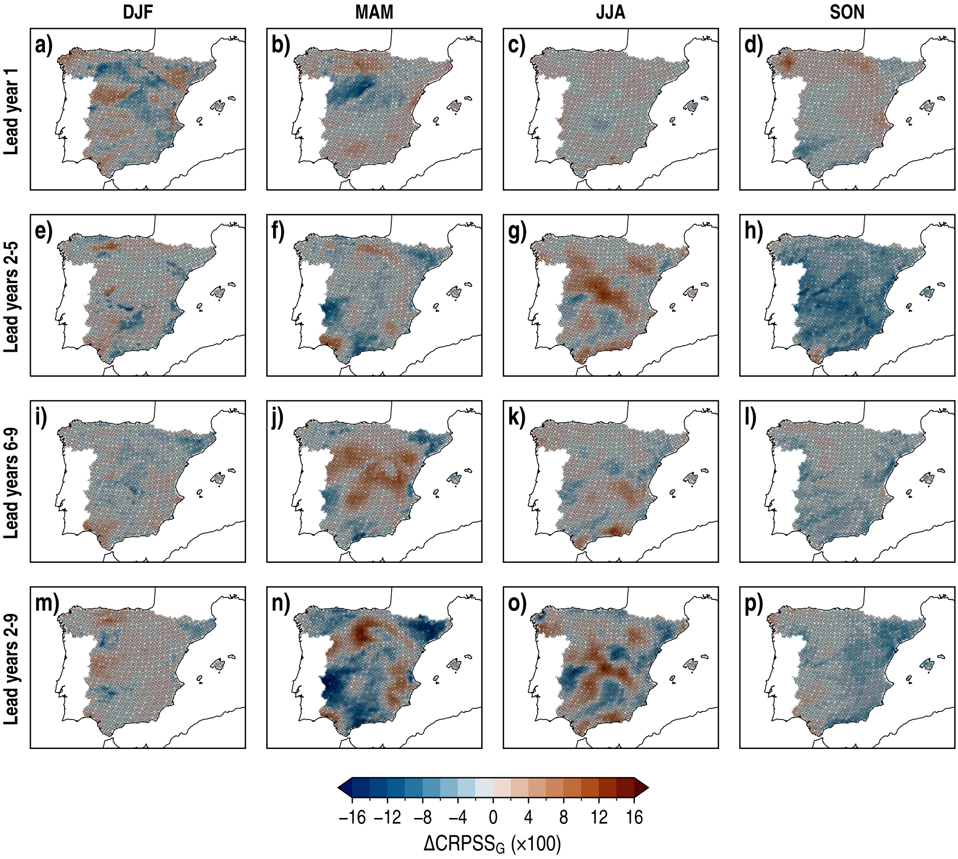

Figure B.37: As Figure B.30 but for ΔCRPSS$_G$, with CESM-DPLE as reference.

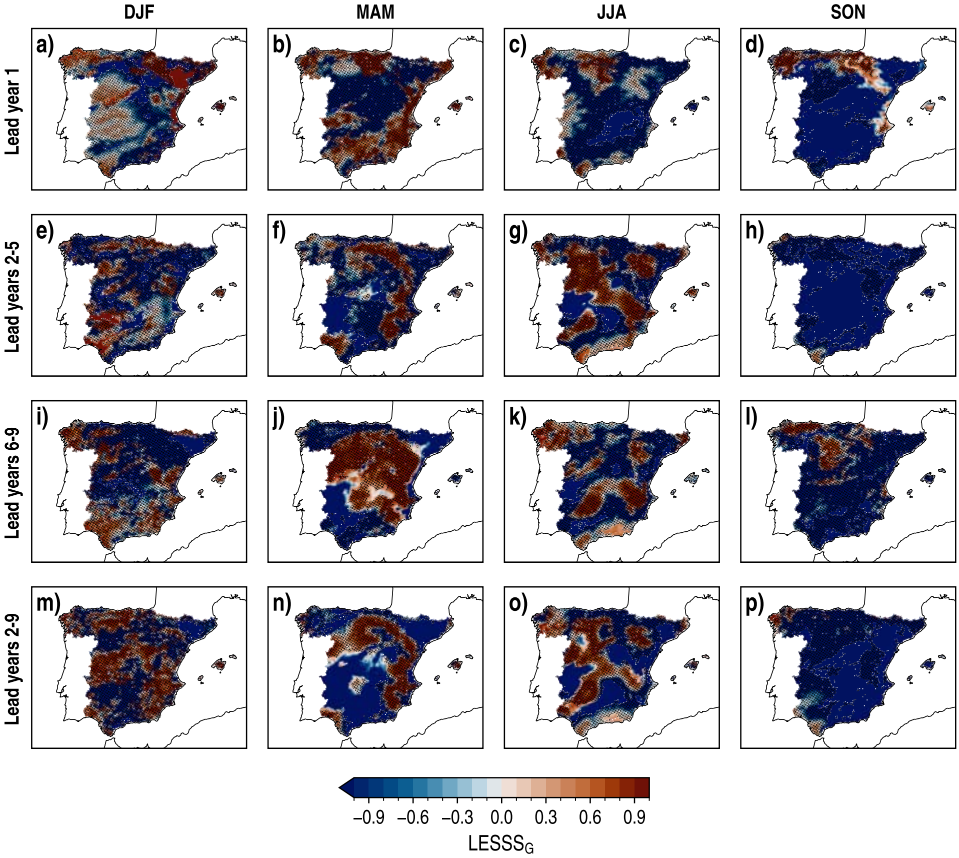

Figure B.38: As Figure B.30 but for LESSS$_G$, with CESM-DPLE as reference.

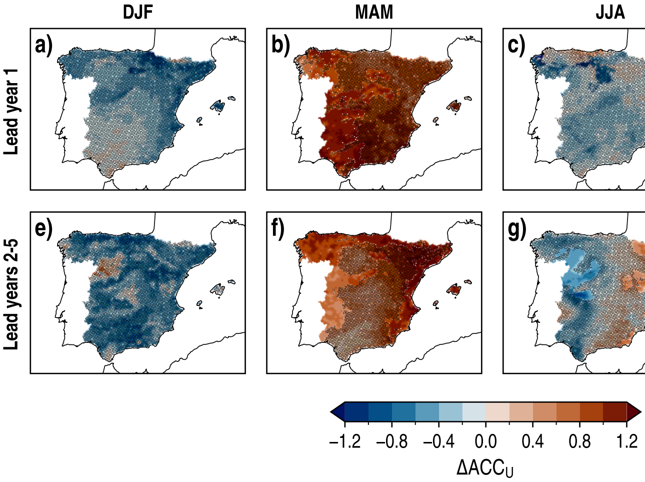

Figure B.39: As Figure B.30 but for ΔACC_U, with WRF-LE as reference, only for lead years 1 and 2–5 (rows).

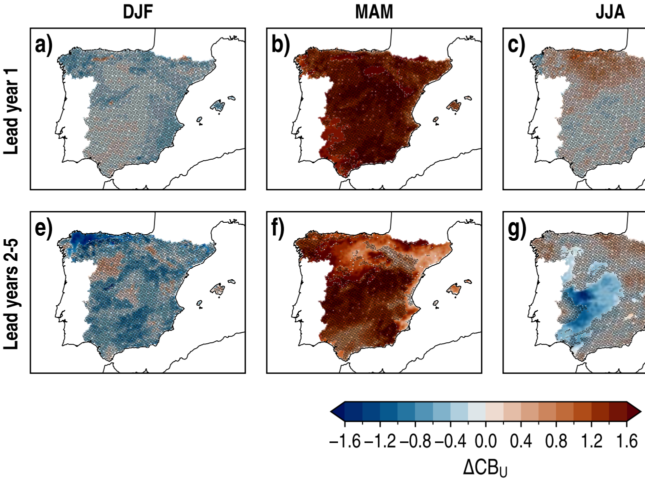

Figure B.40: As Figure B.30 but for ΔCB_U, with WRF-LE as reference, only for lead years 1 and 2–5 (rows).

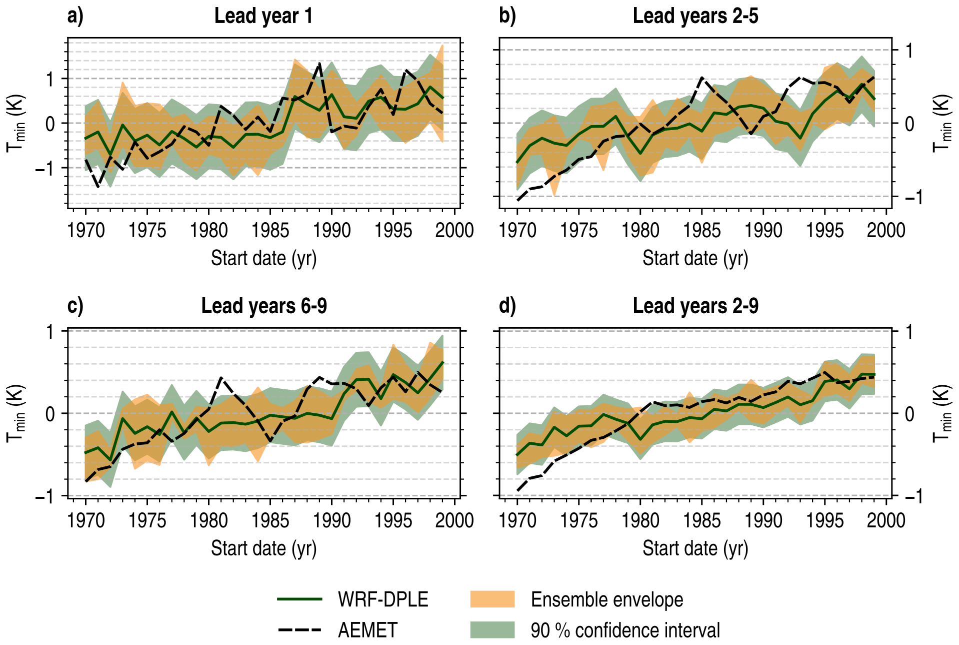

Figure B.41: Time series of the spatially averaged multiannual mean anomalies of T_{min} in the MT region for lead years 1, 2–5, 6–9 and 2–9 at annual scale. Solid green lines identify the WRF-DPLE ensemble mean, whereas dashed black lines correspond to AEMET. Shaded green surfaces indicate the 90 % confidence interval for a WRF-DPLE single member, calculated from the average ensemble spread (Eq. [3.32]). Shaded yellow surfaces show the ensemble envelope which encloses the trajectories followed by the members composing the WRF-DPLE ensemble.

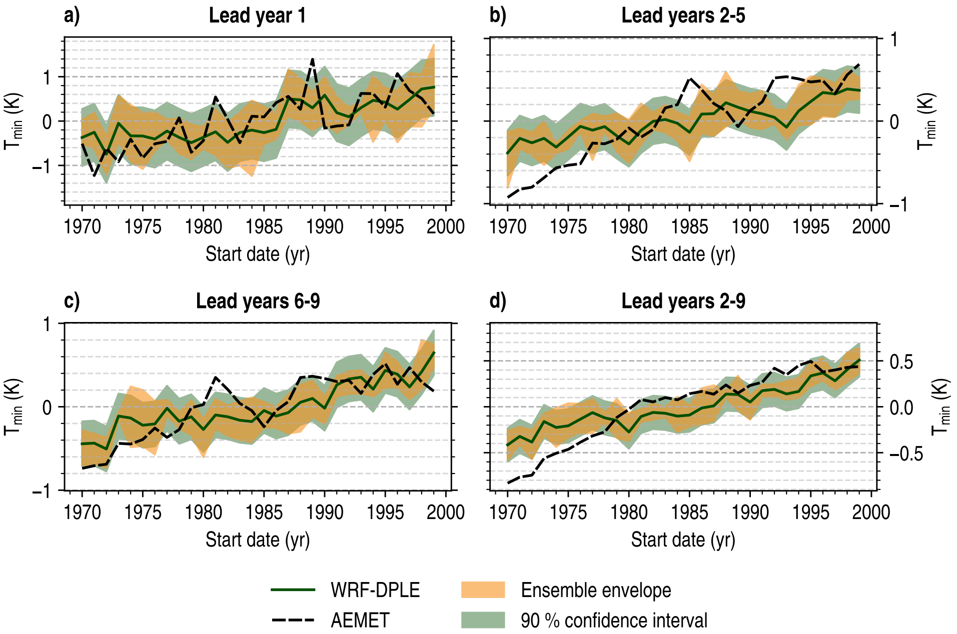

Figure B.42: As Figure B.41 but for the NE region.

B.2.3. *Daily mean near-surface air temperature*

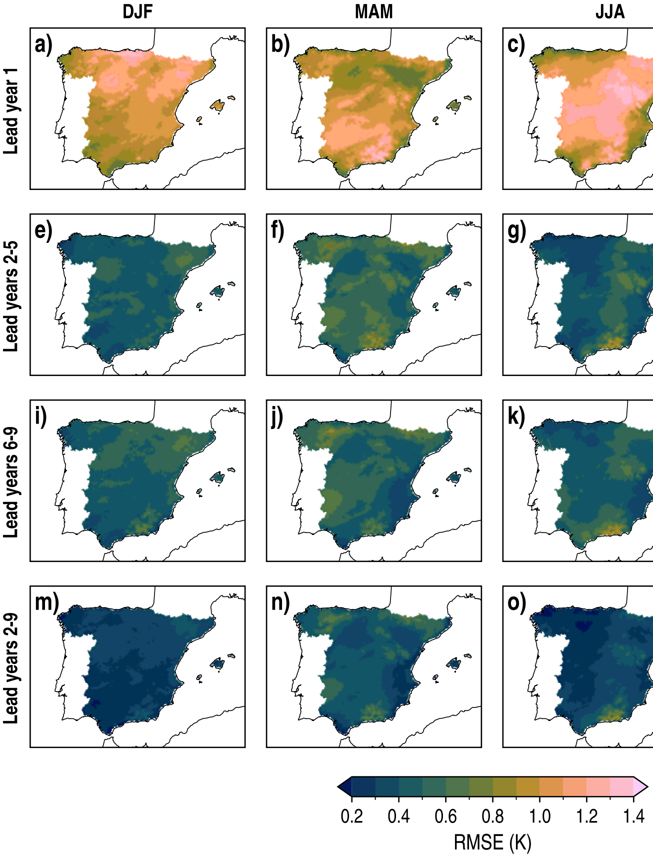

Figure B.43: Spatial distributions of RMSE for the WRF-DPLE multiannual mean anomalies of T_{mean} for lead years 1, 2–5, 6–9 and 2–9 (rows) in DJF, MAM, JJA and SON (columns).

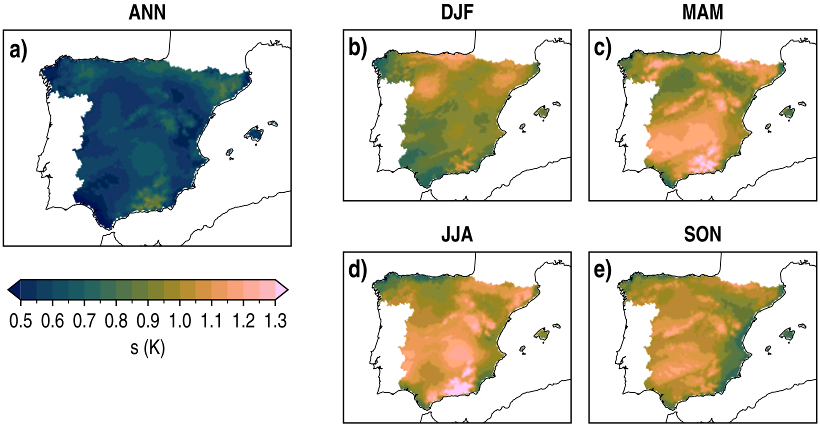

Figure B.44: Spatial distributions of the standard deviation (s) of the AEMET T_{mean} at annual and seasonal scales for the period 1970-2009. While the annual series covers the period from 1970-11 to 2009-10, the seasonal series span the period from 1970-12 to 2009-11.

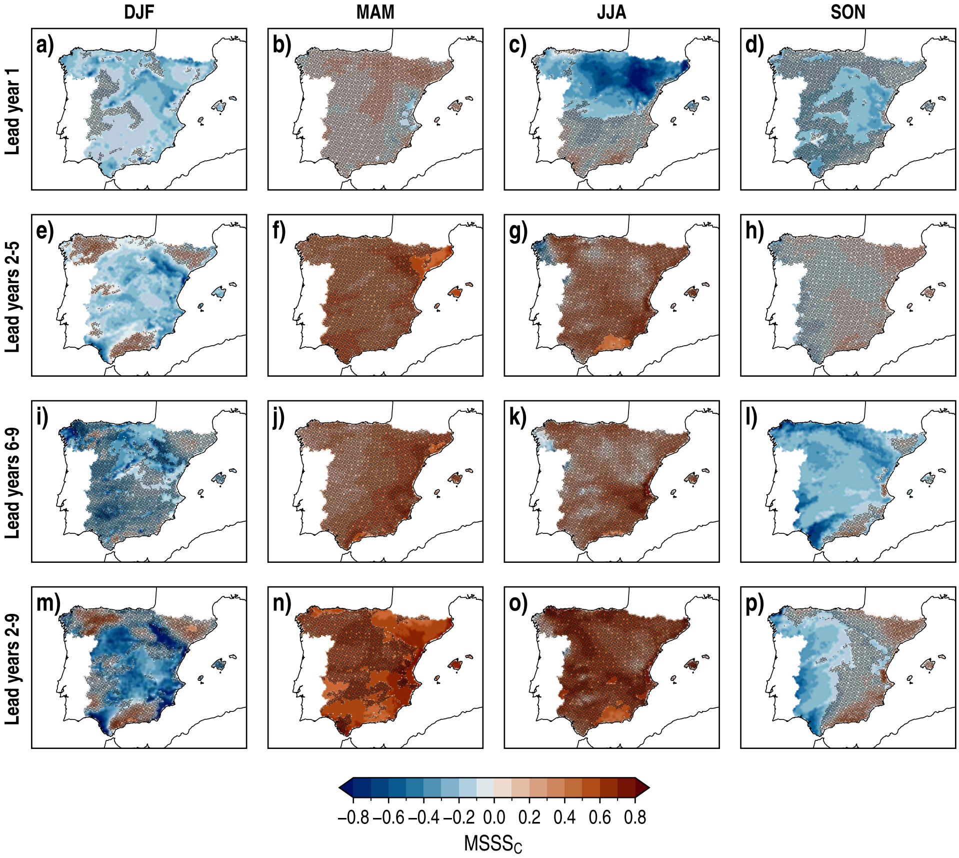

Figure B.45: Spatial distributions of $MSSS_C$, with climatology as reference, for the WRF-DPLE multiannual mean anomalies of T_{mean} for lead years 1, 2–5, 6–9 and 2–9 (rows) in DJF, MAM, JJA and SON (columns). The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 90% confidence level.

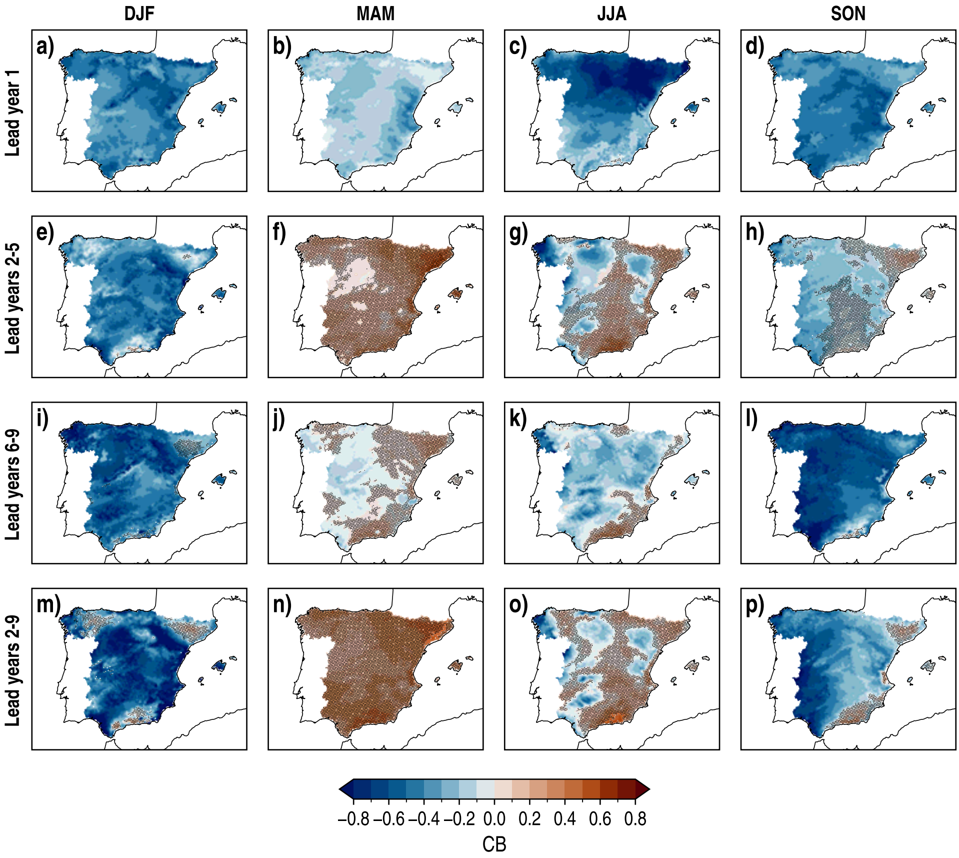

Figure B.46: As Figure B.45 but for CB.

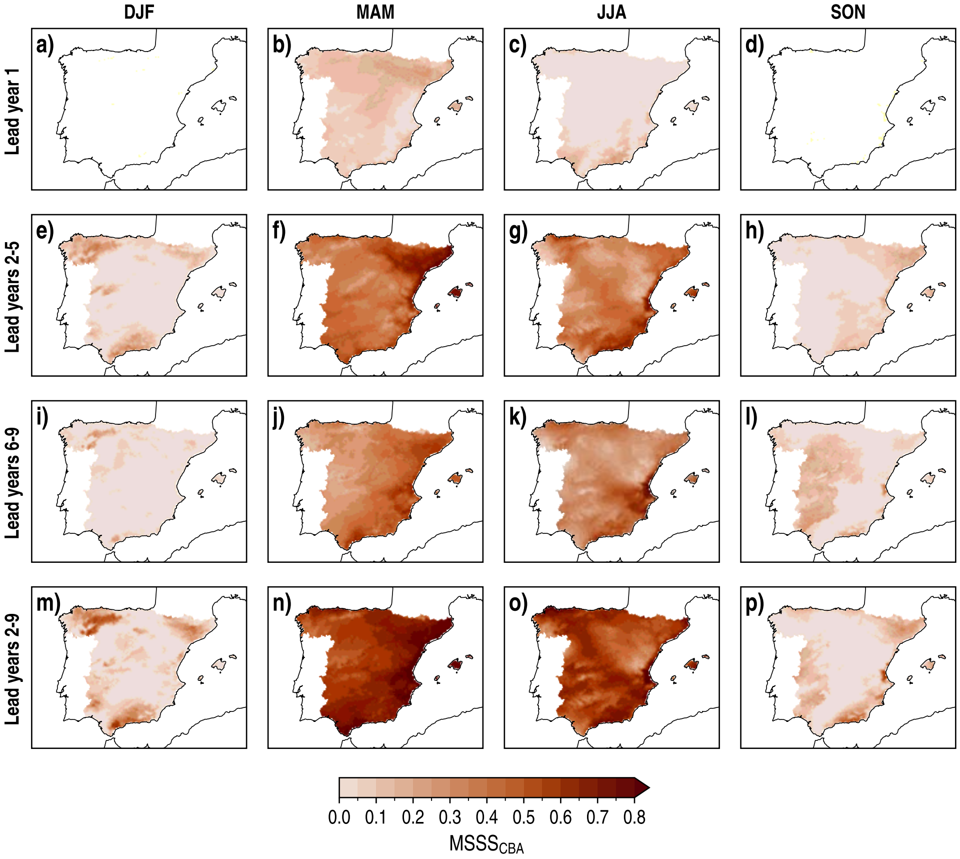

Figure B.47: As Figure B.45 but for $MSSS_{CBA}$ ($MSSS_C$ calculated for lead time series with an adjusted CB, i.e., equal to zero).

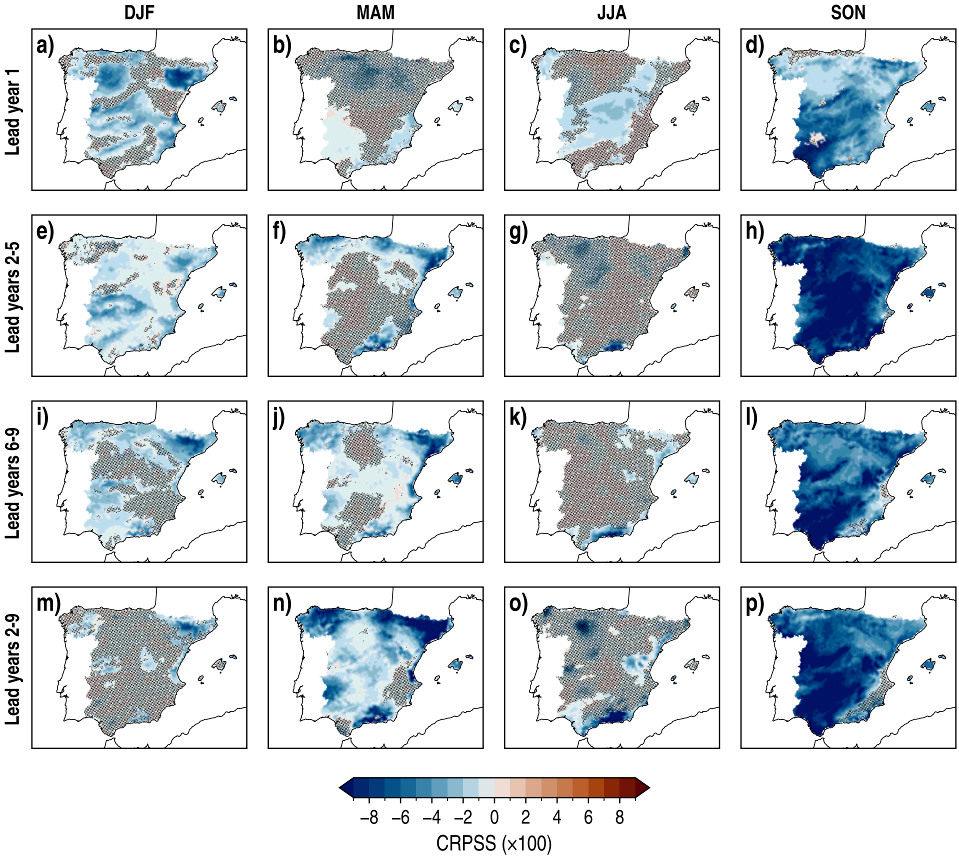

Figure B.48: As Figure B.45 but for CRPSS.

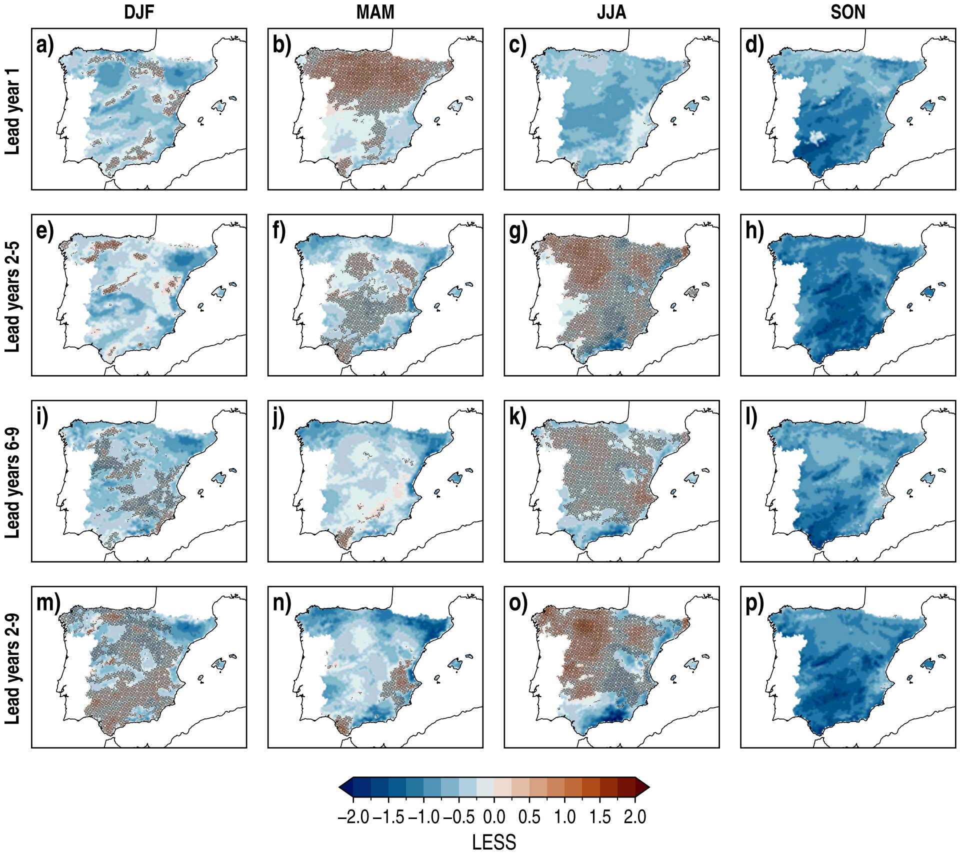

Figure B.49: As Figure B.45 but for LESS.

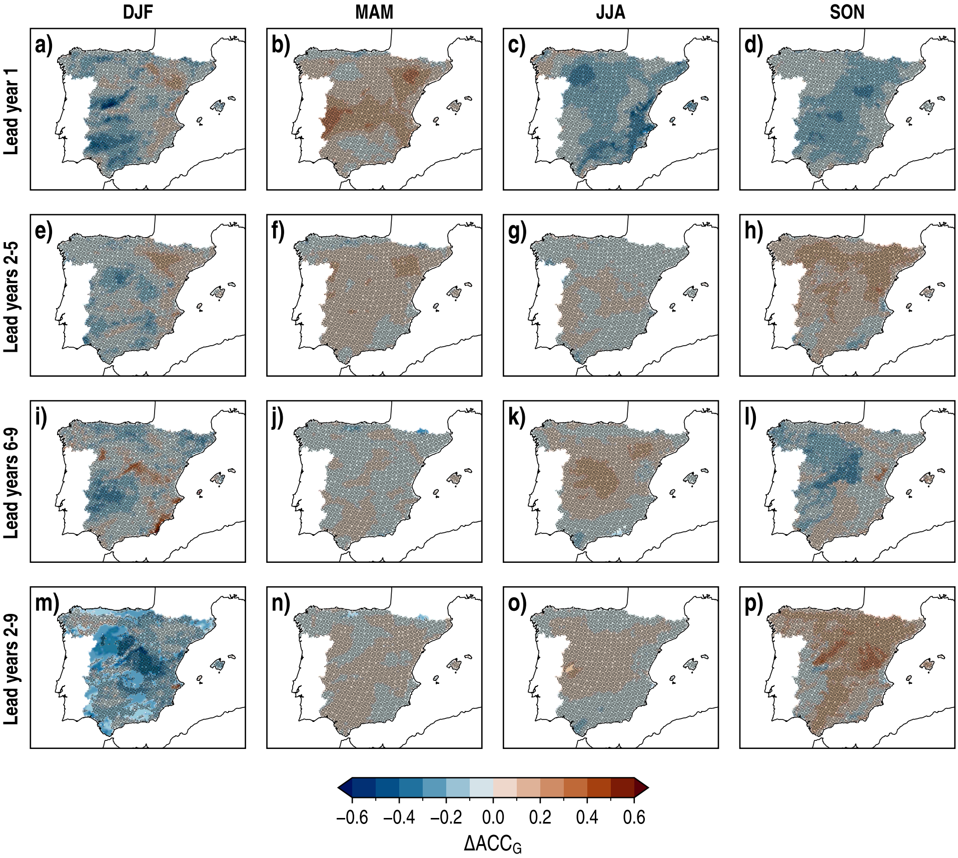

Figure B.50: As Figure B.45 but for ΔACC_G, with CESM-DPLE as reference.

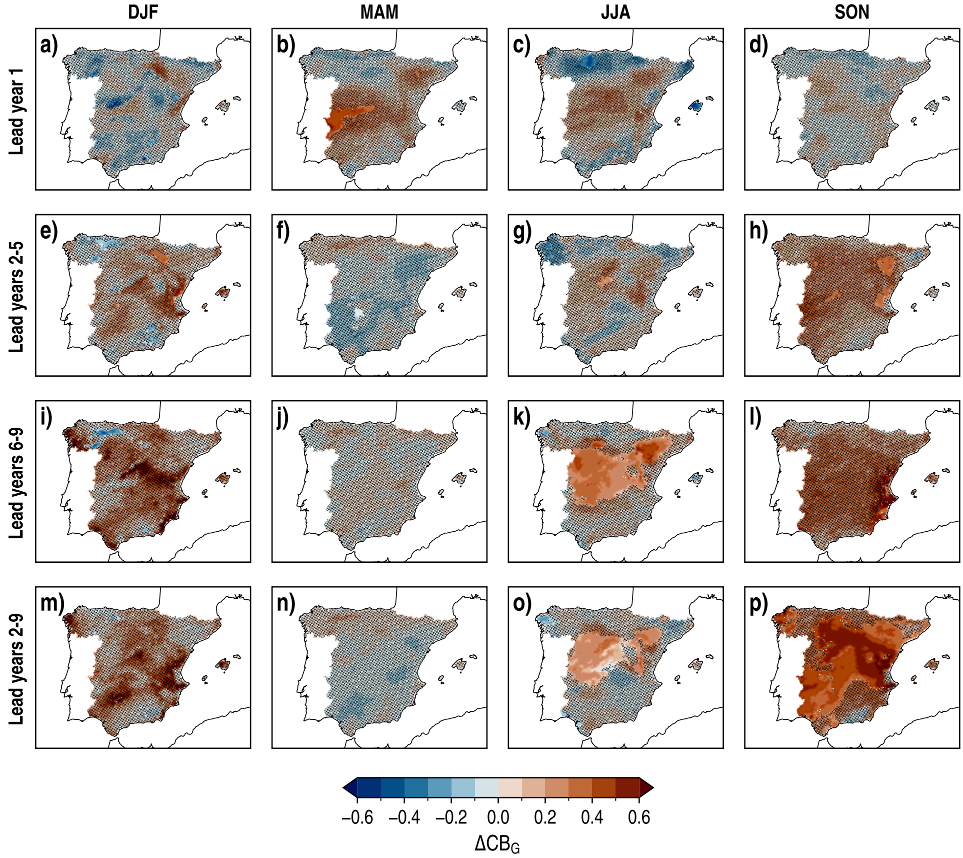

Figure B.51: As Figure B.45 but for ΔCB_G, with CESM-DPLE as reference.

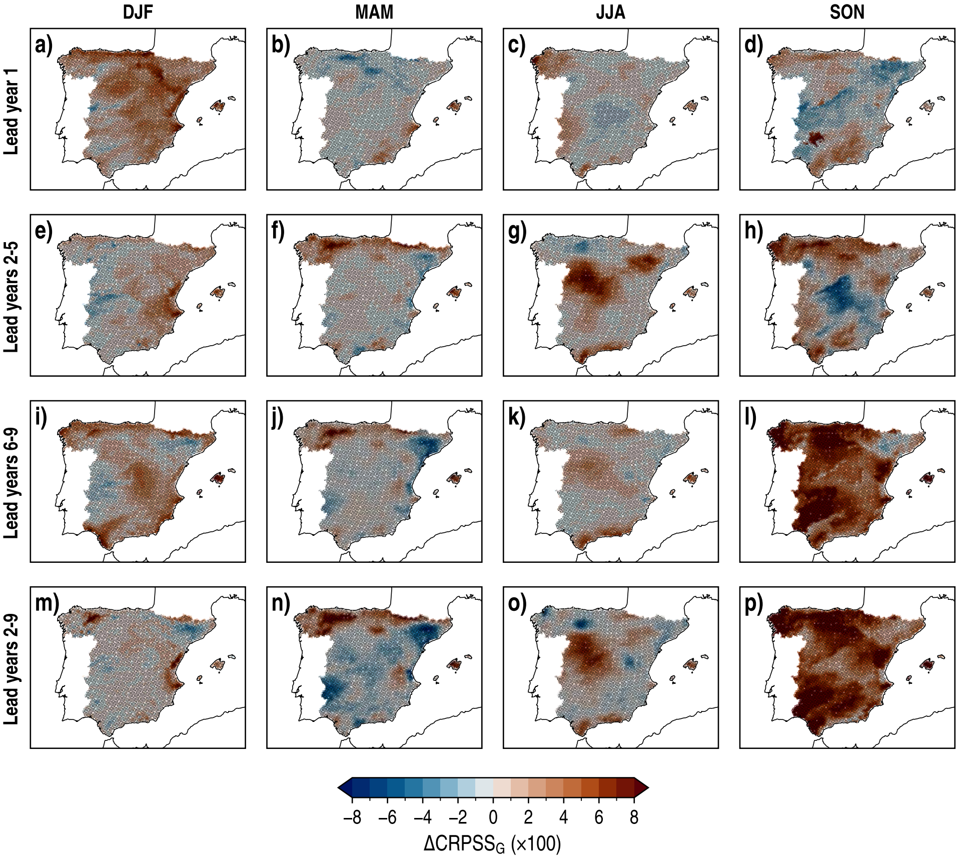

Figure B.52: As Figure B.45 but for $\Delta\text{CRPSS}_\text{G}$, with CESM-DPLE as reference.

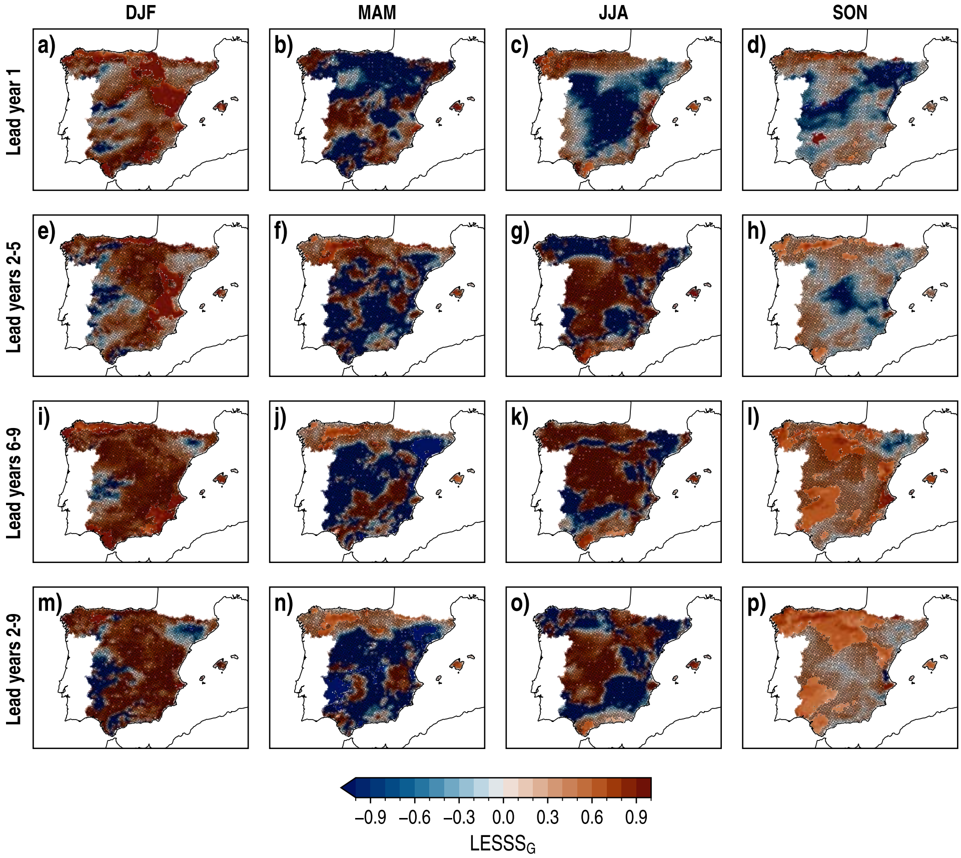

Figure B.53: As Figure B.45 but for LESSS$_G$, with CESM-DPLE as reference.

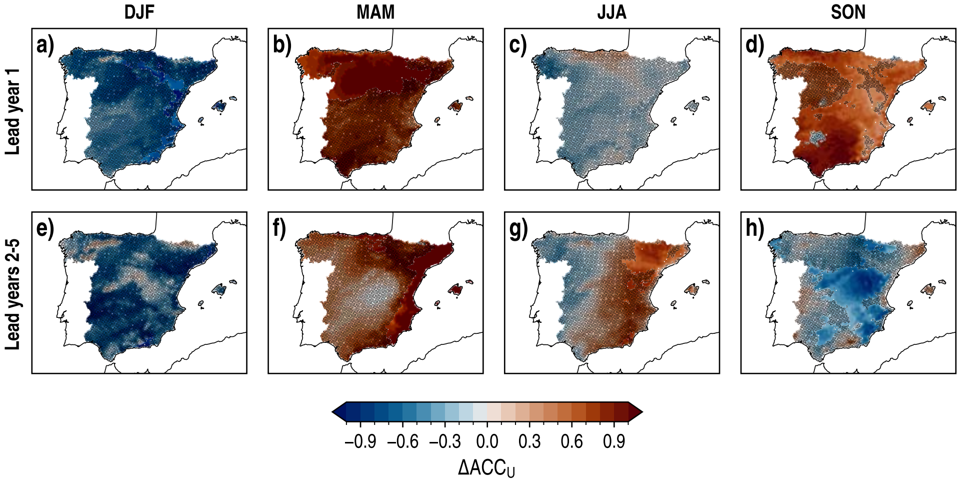

Figure B.54: As Figure B.45 but for ΔACC$_U$, with WRF-LE as reference, only for lead years 1 and 2–5 (rows).

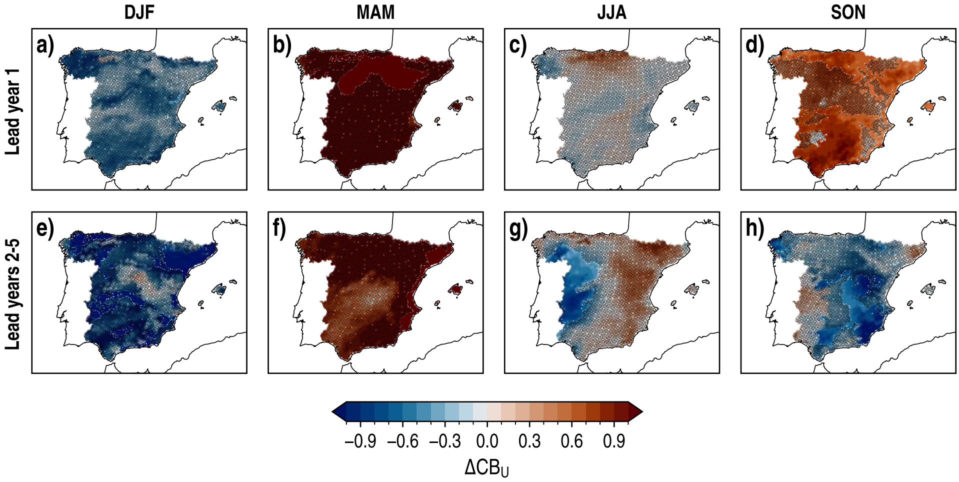

Figure B.55: As Figure B.45 but for ΔCB$_U$, with WRF-LE as reference, only for lead years 1 and 2–5 (rows).

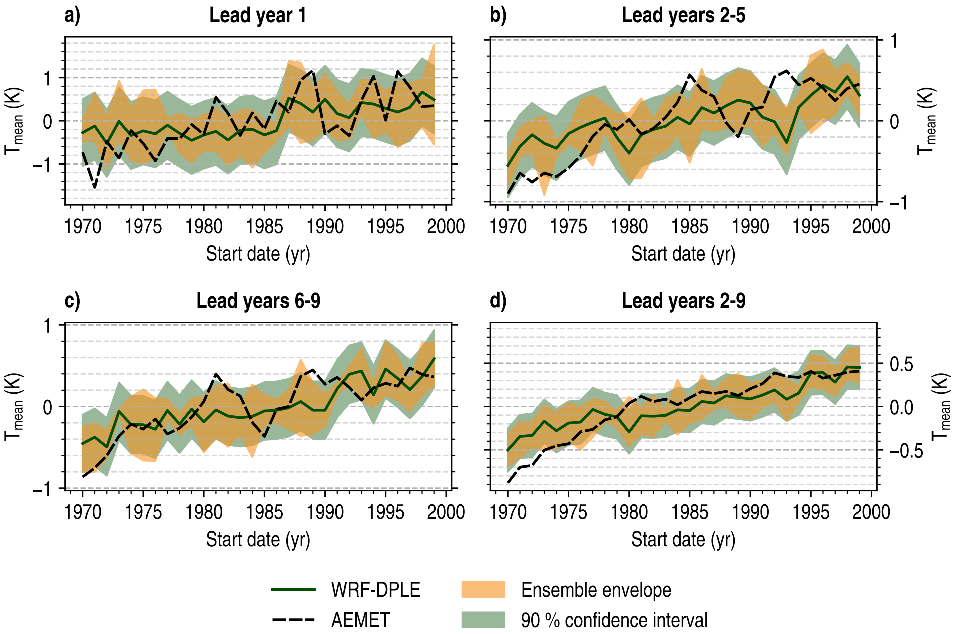

Figure B.56: Time series of the spatially averaged multiannual mean anomalies of T_{mean} in the MT region for lead years 1, 2–5, 6–9 and 2–9 at annual scale. Solid green lines identify the WRF-DPLE ensemble mean, whereas dashed black lines correspond to AEMET. Shaded green surfaces indicate the 90 % confidence interval for a WRF-DPLE single member, calculated from the average ensemble spread (Eq. [3.32]). Shaded yellow surfaces show the ensemble envelope which encloses the trajectories followed by the members composing the WRF-DPLE ensemble.

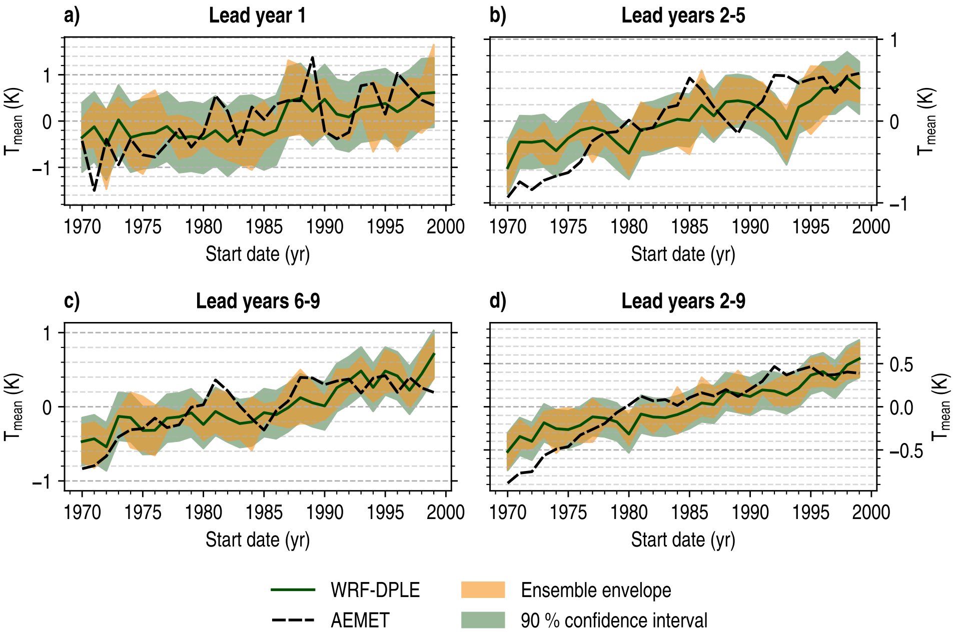

Figure B.57: As Figure B.56 the NE region.

B.3. Decadal climate predictions for the period 2015–2025

B.3.1. *Precipitation*

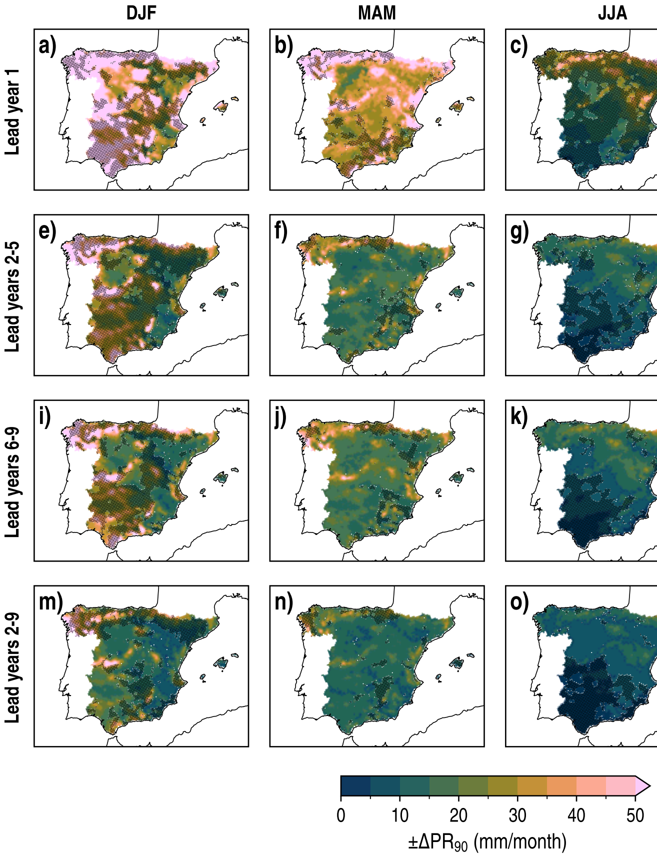

Figure B.58: Spatial distributions of the confidence intervals of PR at the 90 % level for a single WRF-DPLE4 member (±ΔPR_{90}) for lead years 1, 2–5, 6–9 and 2–9 (rows) in DJF, MAM, JJA and SON (columns). The absence (presence) of black dots indicates the locations where the confidence intervals represent the forecast uncertainty at the 90 % confidence level.

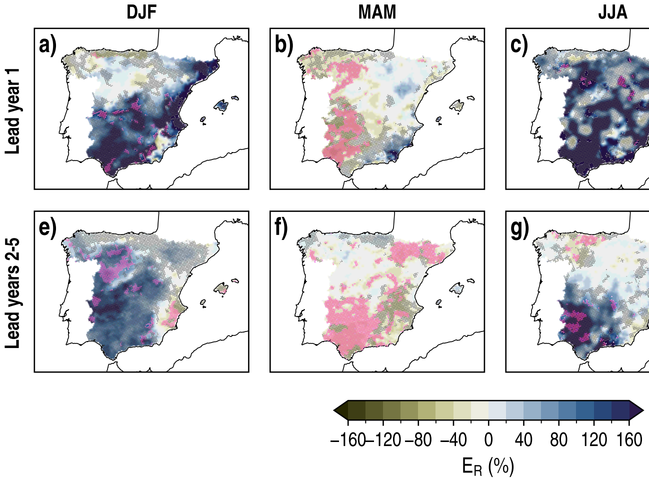

Figure B.59: As Figure B.58 but for the WRF-DPLE$_4$ relative anomaly error of PR (E_R) in lead years 1 and 2–5 (rows). Pink triangles indicate the locations where the forecast uncertainty is represented by the confidence intervals but the observational anomalies fall outside them.

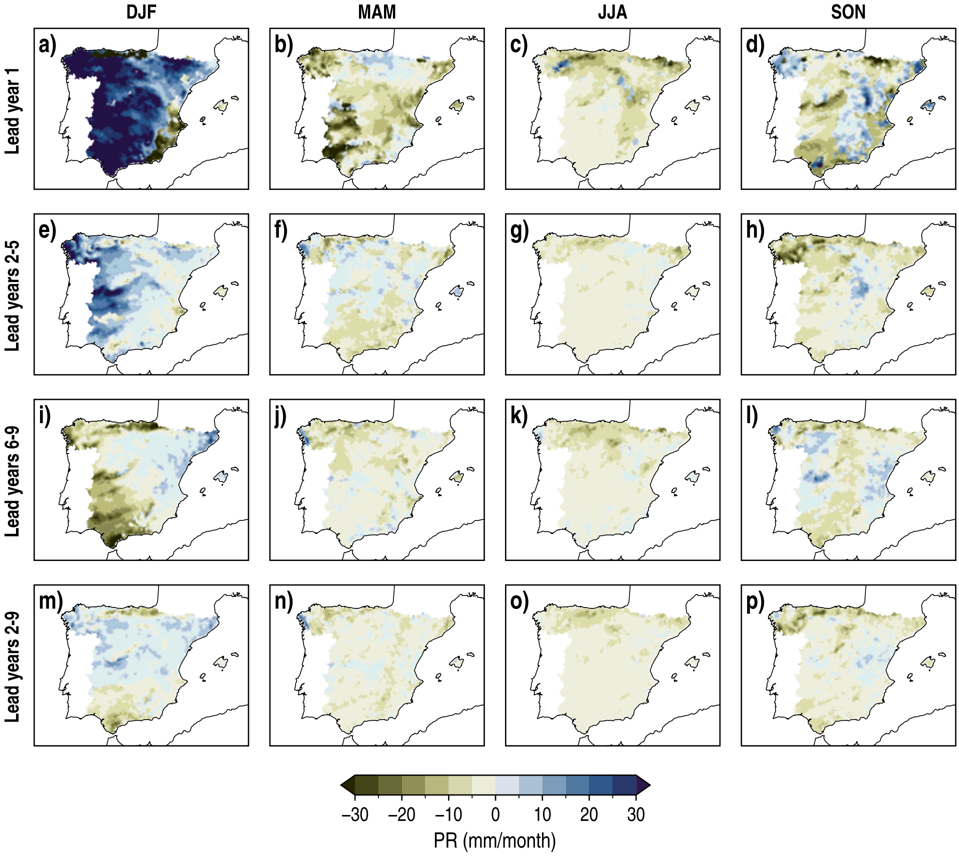

Figure B.60: Spatial distributions of the WRF-DPLE$_{10}$ multiannual mean anomalies of PR for lead years 1, 2–5, 6–9 and 2–9 (rows) in DJF, MAM, JJA and SON (columns).

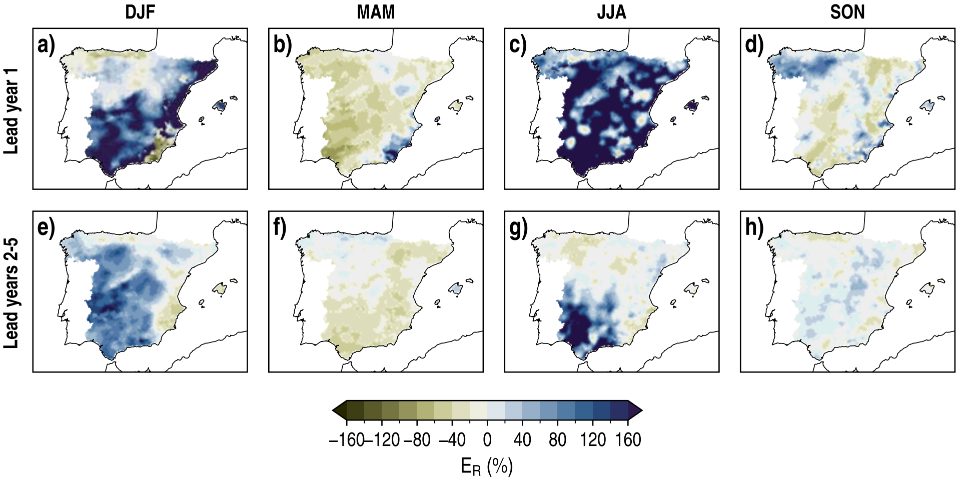

Figure B.61: Spatial distributions of the WRF-DPLE$_{10}$ relative anomaly error of PR (E_{R}) for lead years 1 and 2–5 (rows) in DJF, MAM, JJA and SON (columns).

B.3.2. *Daily maximum near-surface air temperature*

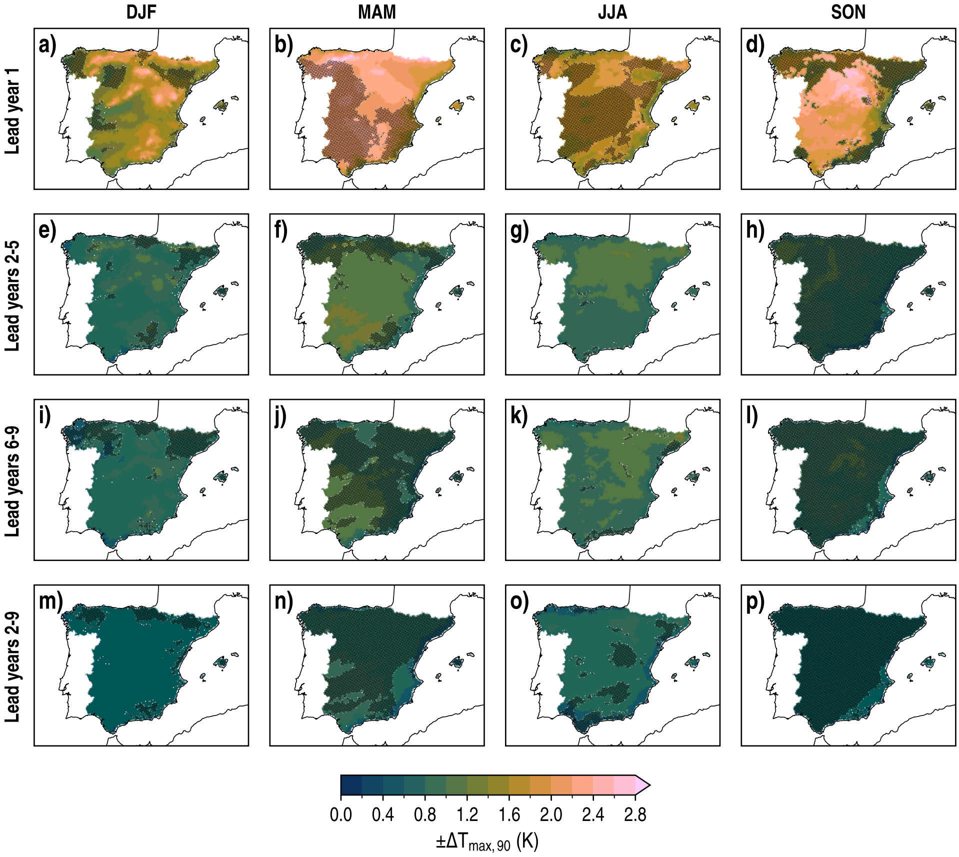

Figure B.62: Spatial distributions of the confidence intervals of T_{max} at the 90 % level for a single WRF-DPLE$_4$ member ($\pm\Delta T_{\text{max,90}}$) for lead years 1, 2–5, 6–9 and 2–9 (rows) in DJF, MAM, JJA and SON (columns). The absence (presence) of black dots indicates the locations where the confidence intervals represent the forecast uncertainty at the 90 % confidence level.

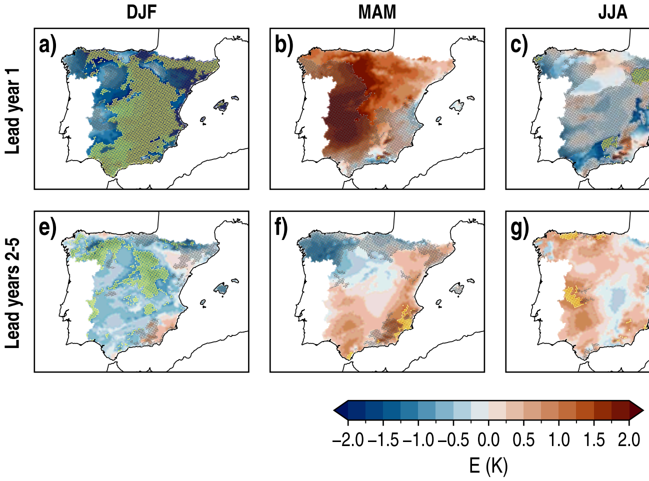

Figure B.63: As Figure B.62 but for the WRF-DPLE$_4$ anomaly error of $T_{\max}$ (E) in lead years 1 and 2–5 (rows). Yellow triangles indicate the locations where the forecast uncertainty is represented by the confidence intervals but the observational anomalies fall outside them.

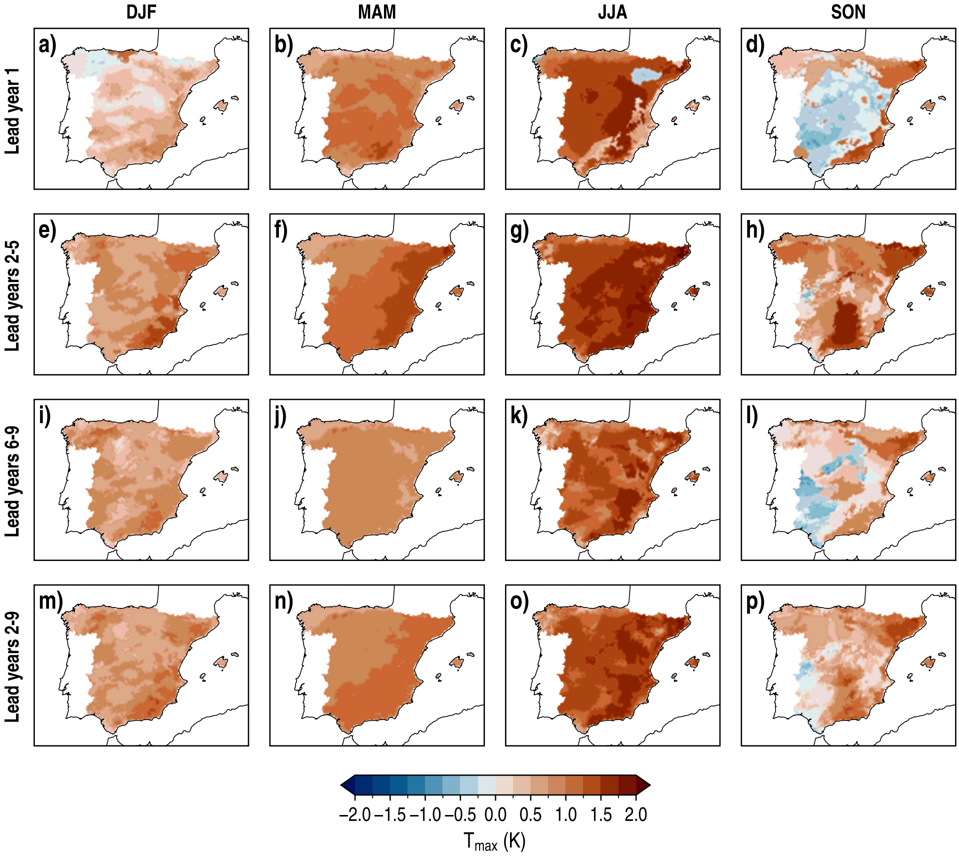

Figure B.64: Spatial distributions of the WRF-DPLE$_{10}$ multiannual mean anomalies of $T_{\max}$ for lead years 1, 2–5, 6–9 and 2–9 (rows) in DJF, MAM, JJA and SON (columns).

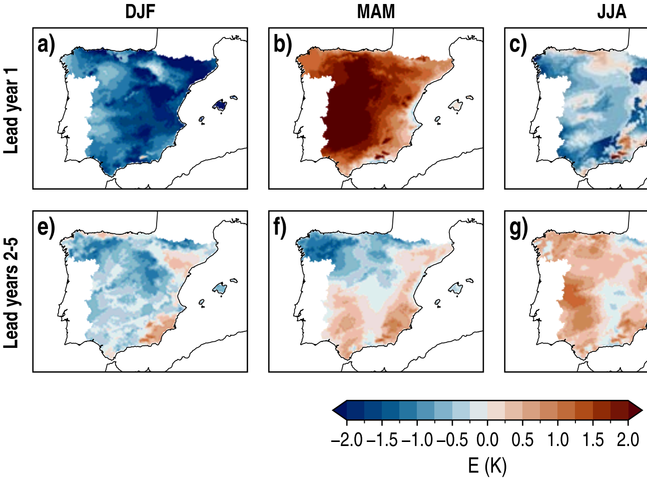

Figure B.65: Spatial distributions of the WRF-DPLE$_{10}$ anomaly error of $T_{\max}$ (E) for lead years 1 and 2–5 (rows) in DJF, MAM, JJA and SON (columns).

B.3.3. *Daily minimum near-surface air temperature*

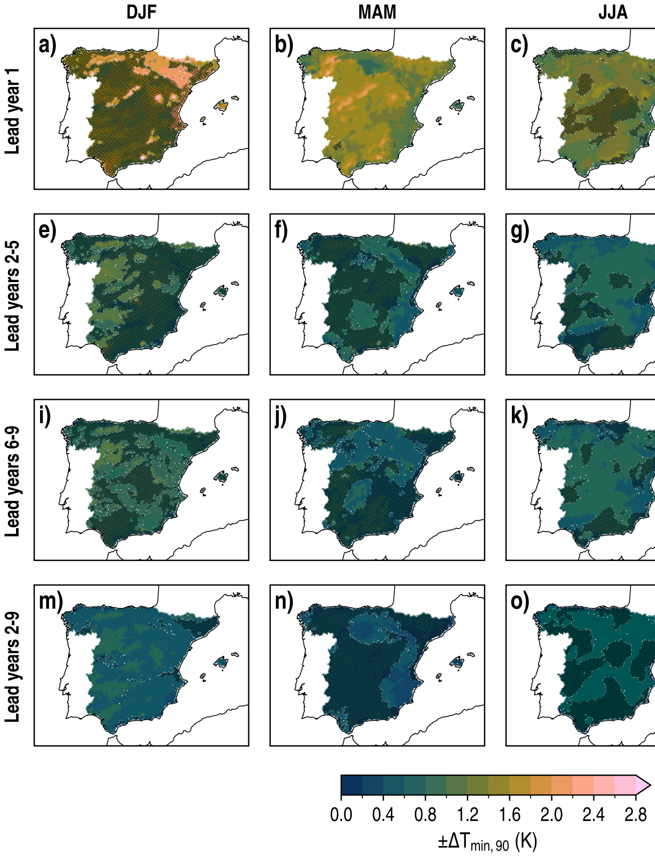

Figure B.66: Spatial distributions of the confidence intervals of T_{min} at the 90 % level for a single WRF-DPLE$_4$ member ($\pm\Delta T_{\mathrm{min,90}}$) for lead years 1, 2–5, 6–9 and 2–9 (rows) in DJF, MAM, JJA and SON (columns). The absence (presence) of black dots indicates the locations where the confidence intervals represent the forecast uncertainty at the 90 % confidence level.

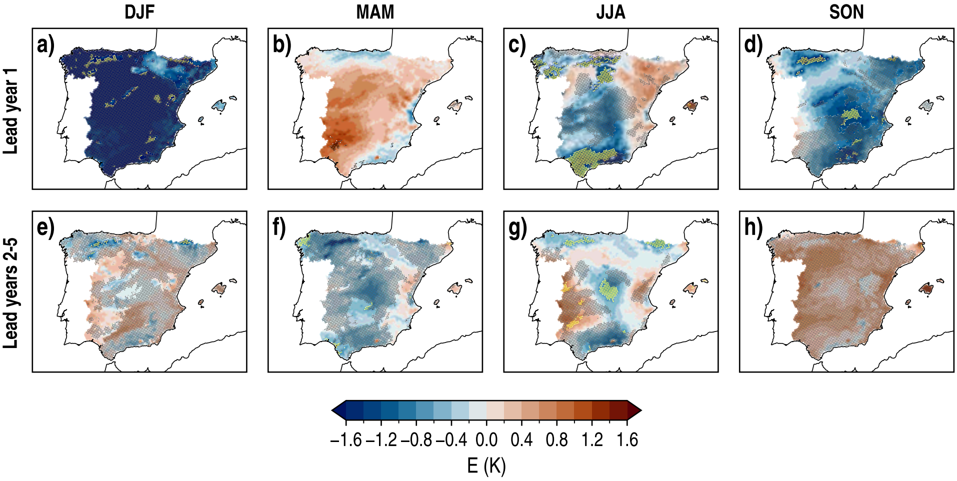

Figure B.67 : As Figure B.66 but for the WRF-DPLE$_4$ anomaly error of $T_{\min}$ (E) in lead years 1 and 2–5 (rows). Yellow triangles indicate the locations where the forecast uncertainty is represented by the confidence intervals but the observational anomalies fall outside them.

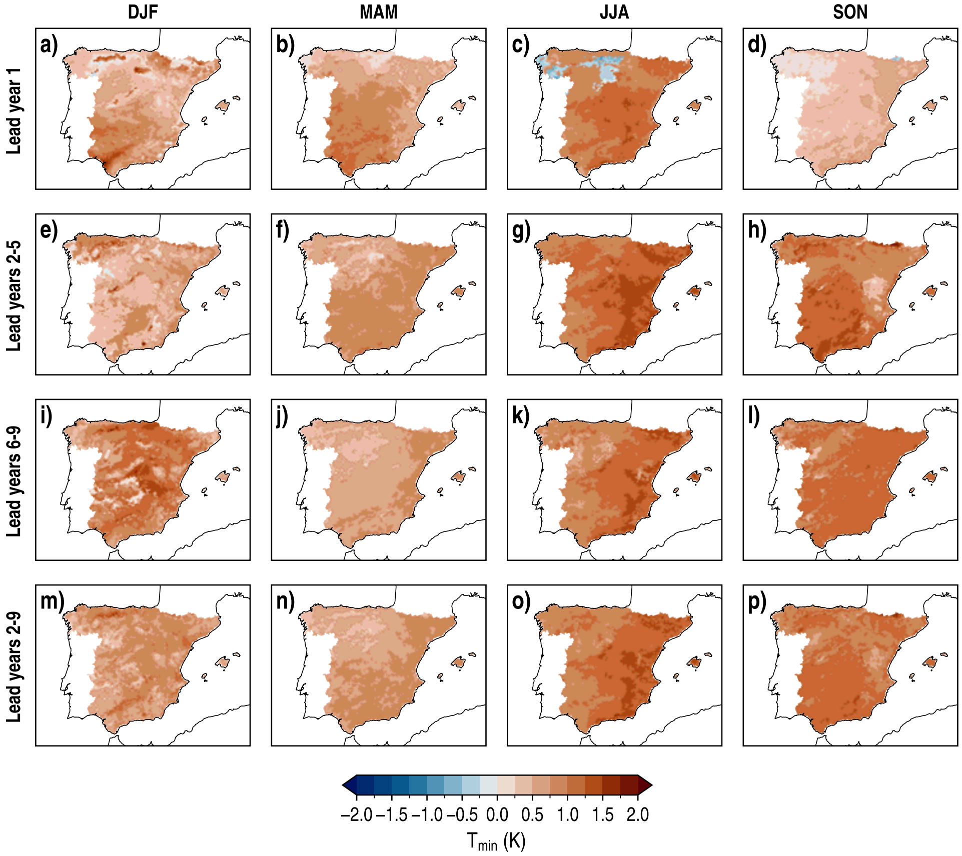

Figure B.68: Spatial distributions of the WRF-DPLE$_{10}$ multiannual mean anomalies of T_{min} for lead years 1, 2–5, 6–9 and 2–9 (rows) in DJF, MAM, JJA and SON (columns).

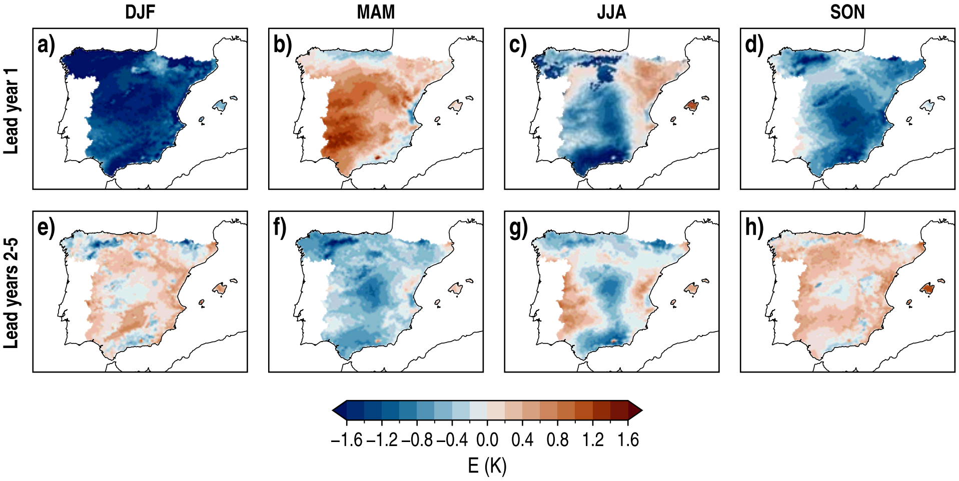

Figure B.69: Spatial distributions of the WRF-DPLE$_{10}$ anomaly error of T_{min} (E) for lead years 1 and 2–5 (rows) in DJF, MAM, JJA and SON (columns).

B.3.4. *Daily mean near-surface air temperature*

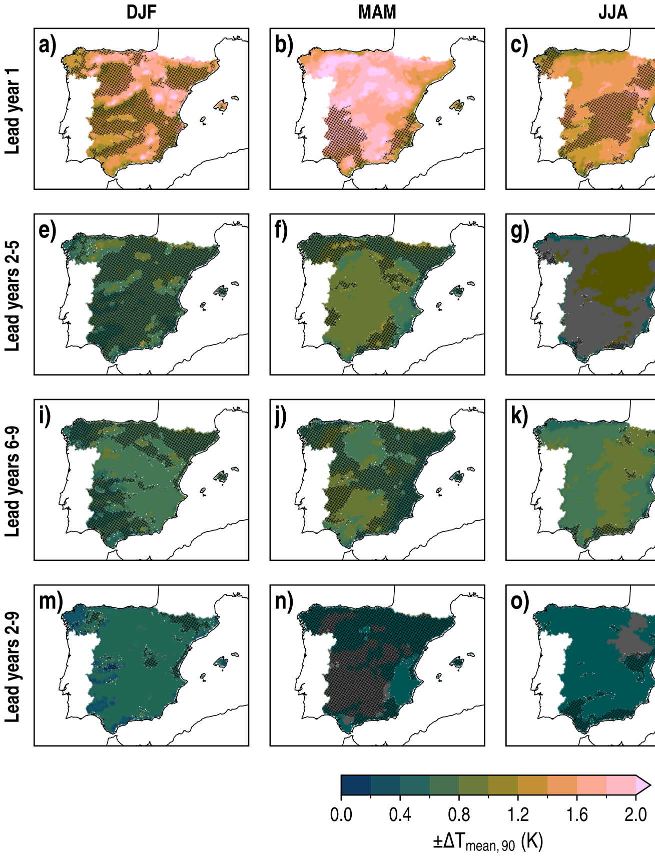

Figure B.70 : Spatial distributions of the confidence intervals of T_{mean} at the 90 % level for a single WRF-DPLE$_4$ member ($\pm\Delta T_{\mathrm{mean},90}$) for lead years 1, 2–5, 6–9 and 2–9 (rows) in DJF, MAM, JJA and SON (columns). The absence (presence) of black dots indicates the locations where the confidence intervals represent the forecast uncertainty at the 90 % confidence level.

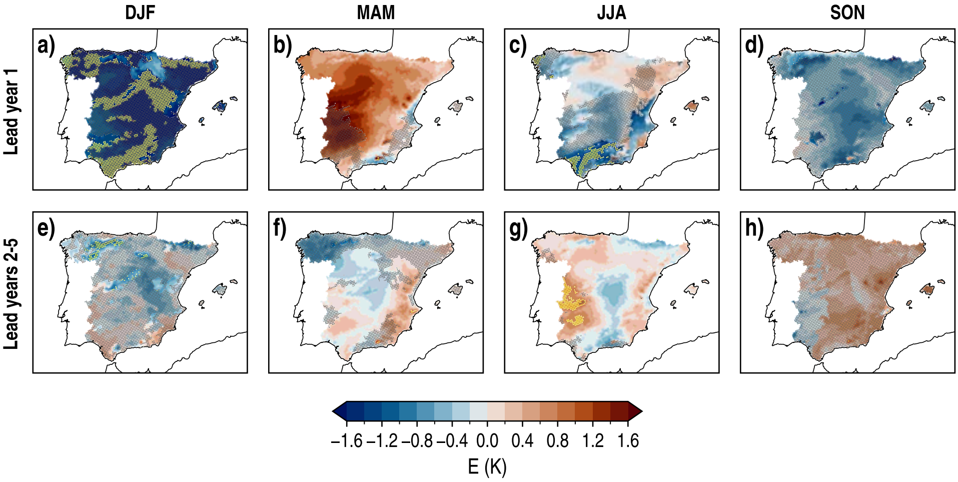

Figure B.71: As Figure B.70 but for the WRF-DPLE$_4$ anomaly error of T_{mean} (E) in lead years 1 and 2–5 (rows). Yellow triangles indicate the locations where the forecast uncertainty is represented by the confidence intervals but the observational anomalies fall outside them.

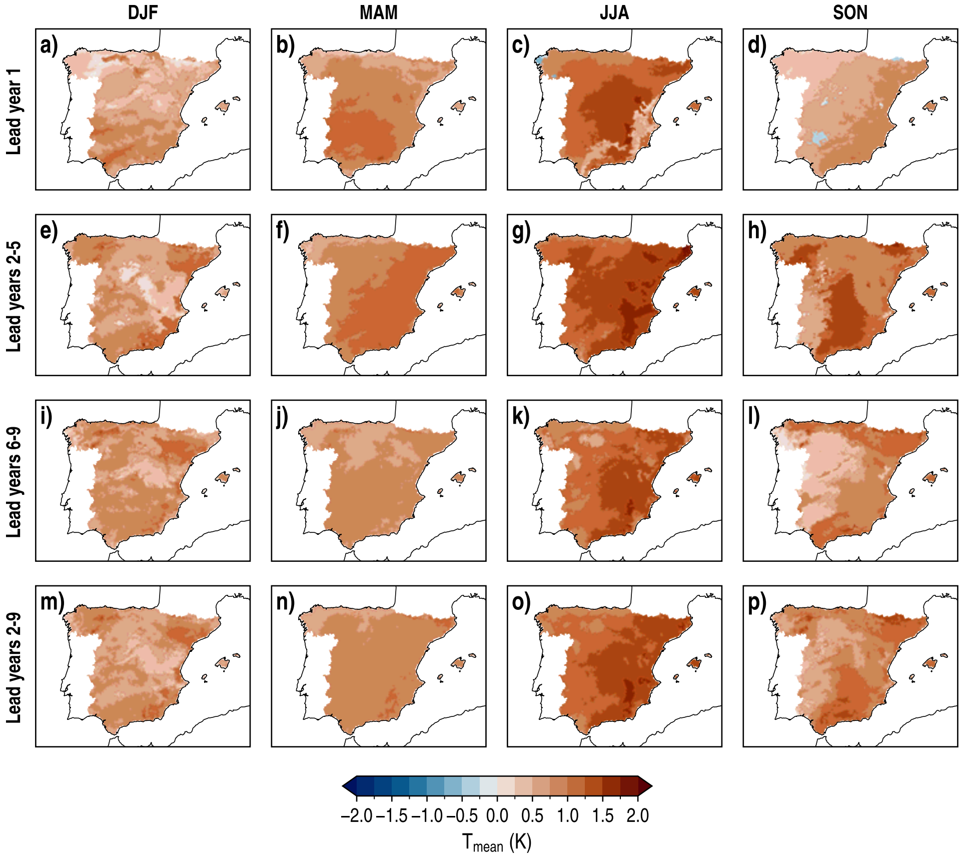

Figure B.72: Spatial distributions of the WRF-DPLE$_{10}$ multiannual mean anomalies of T_{mean} for lead years 1, 2–5, 6–9 and 2–9 (rows) in DJF, MAM, JJA and SON (columns).

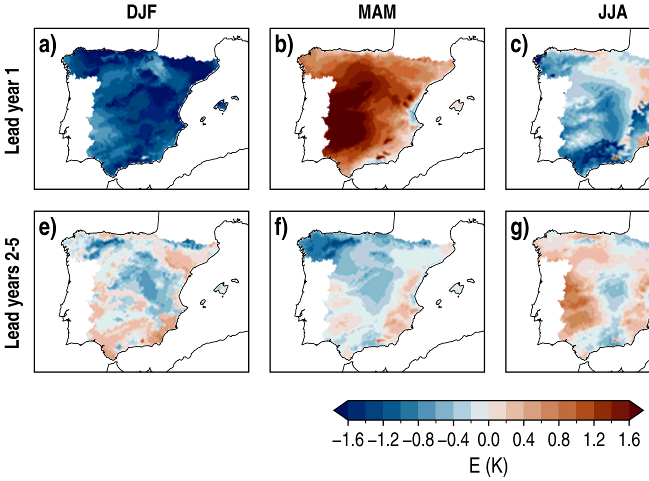

Figure B.73: Spatial distributions of the WRF-DPLE$_{10}$ anomaly error of T_{mean} (E) for lead years 1 and 2–5 (rows) in DJF, MAM, JJA and SON (columns).

B.4. Drit correction techniques for decadal climate predictions

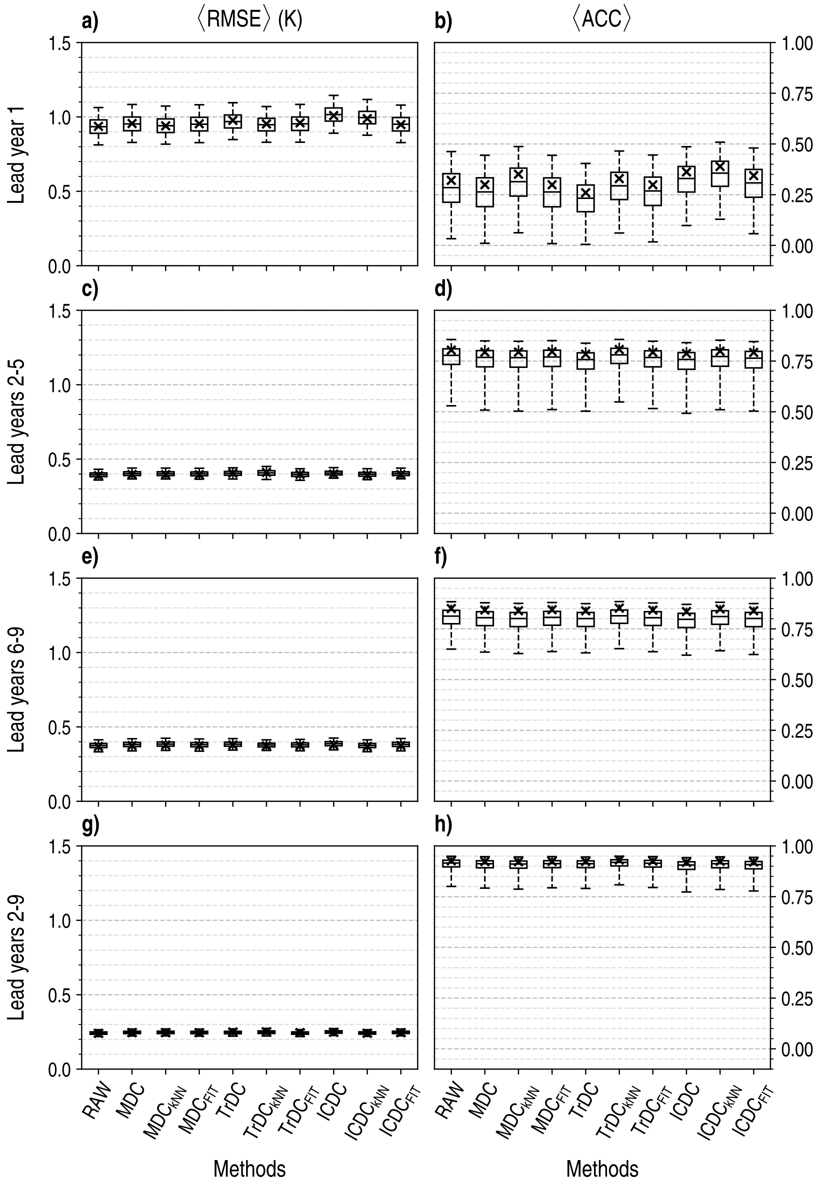

Figure B.74: Spatially averaged RMSE (⟨RMSE⟩, left column) and ACC (⟨ACC⟩, right column) for the ENS40 NSAT anomaly along lead time in the EUR domain. The results are presented for each drift correction method. Crosses denote the spatial averages. Box plots show the results of a bootstrapping (see Section 8.2.3) for which lines indicate the median value and boxes and whiskers enclose the confidence intervals at the 50 % and 95 % levels, respectively.

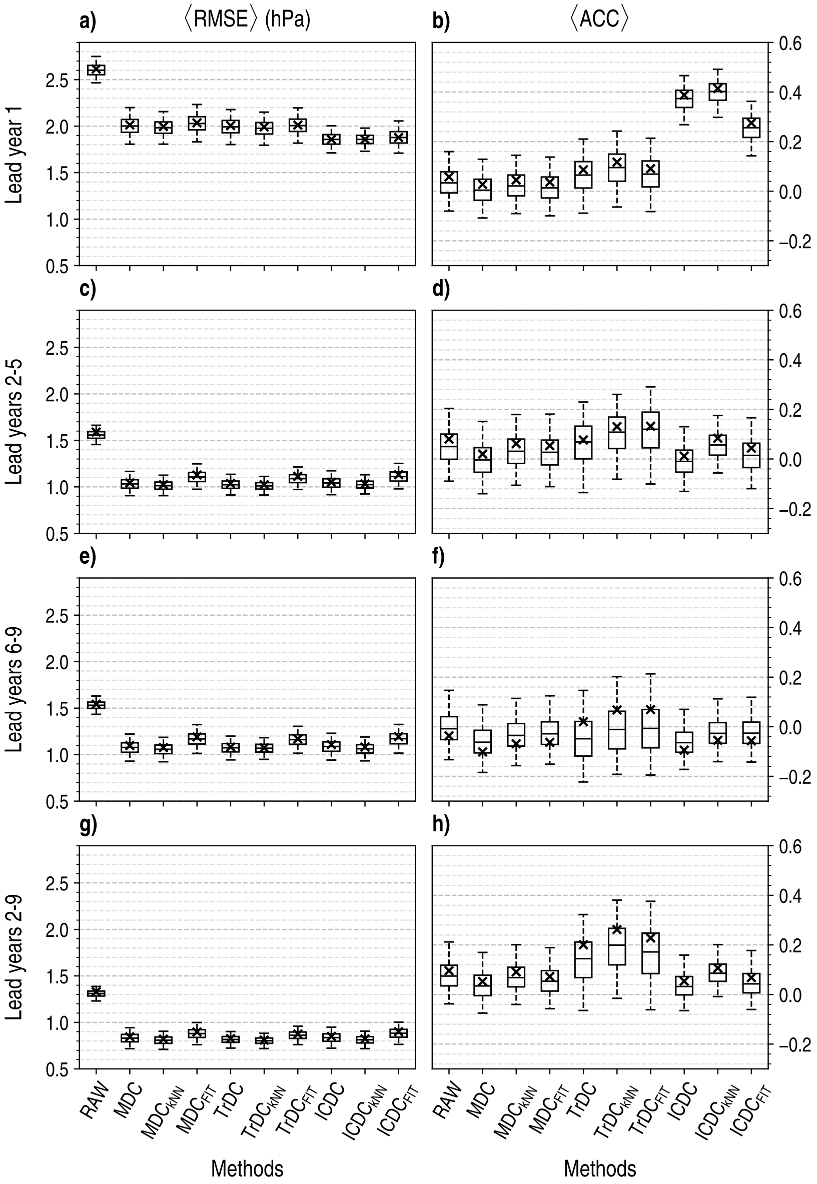

Figure B.75: As Figure B.74 but for SLP.

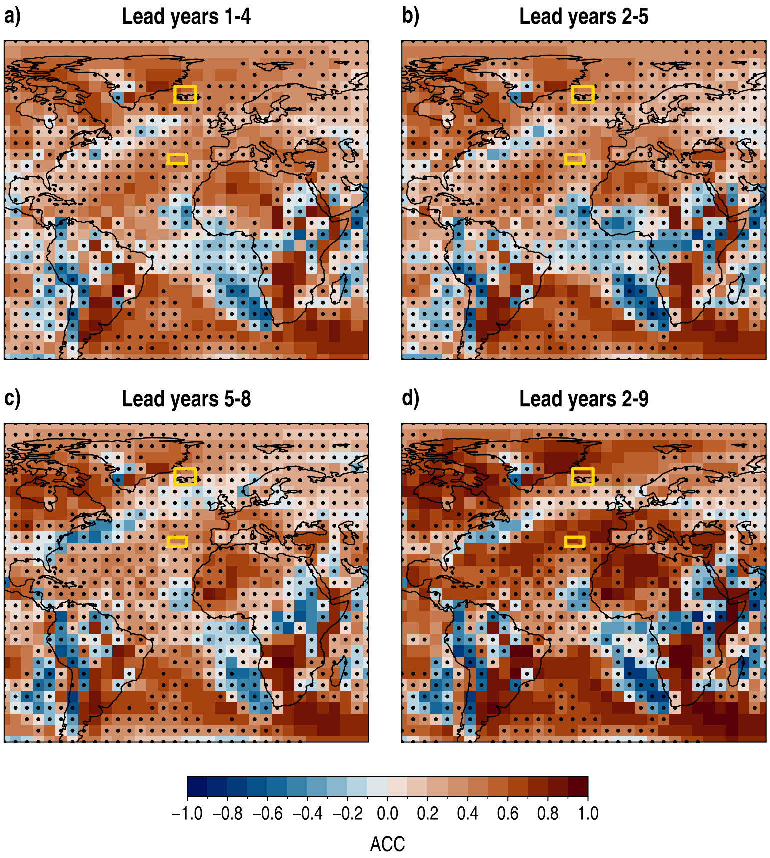

Figure B.76: Spatial distributions of ACC for the ENS40 multiannual means of SLP in lead years 1–4, 2–5, 5–8 and 2–9 in DJF. The data have been drift corrected with the $\mathrm{TrDC}_{\mathrm{kNN}}$ method. The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 95 % confidence level. Yellow boxes denote the regions considered in the calculation of the NAO index.

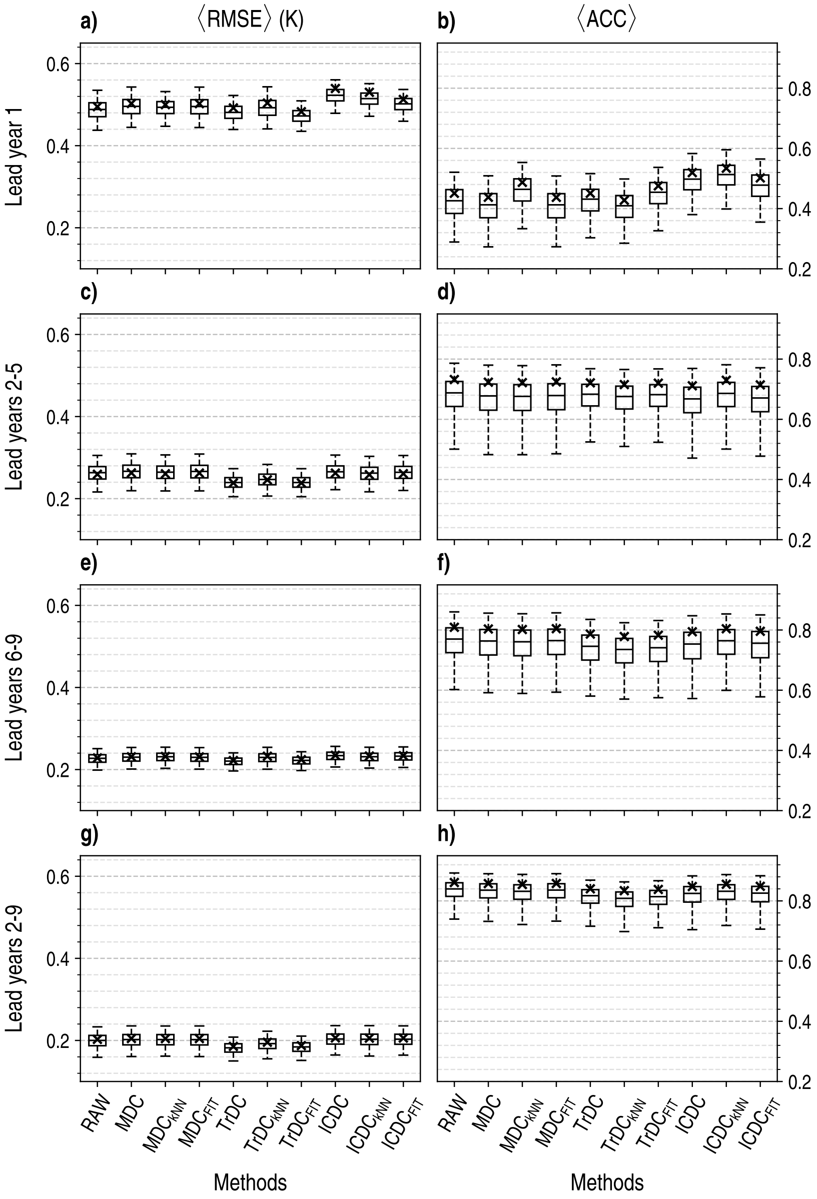

Figure B.77: As Figure B.74 but for the SA domain.

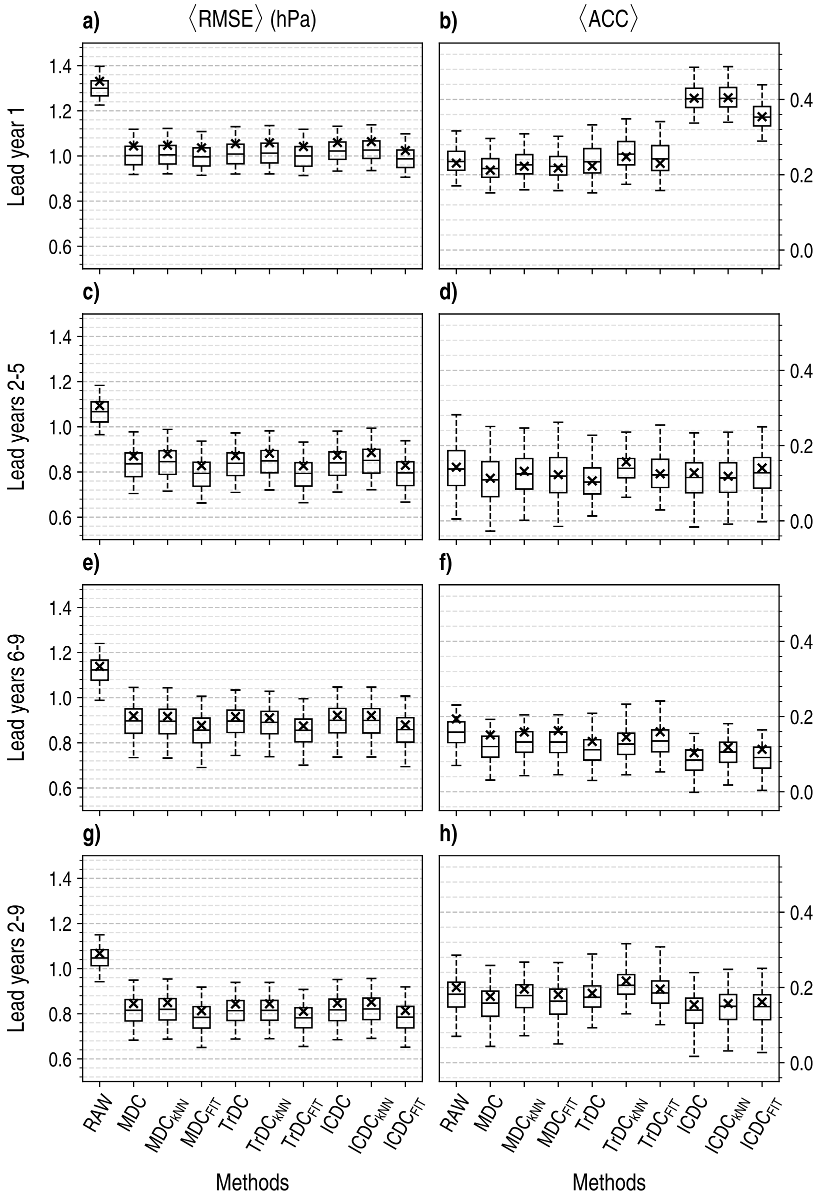

Figure B.78: As Figure B.74 but for SLP and the SA domain.

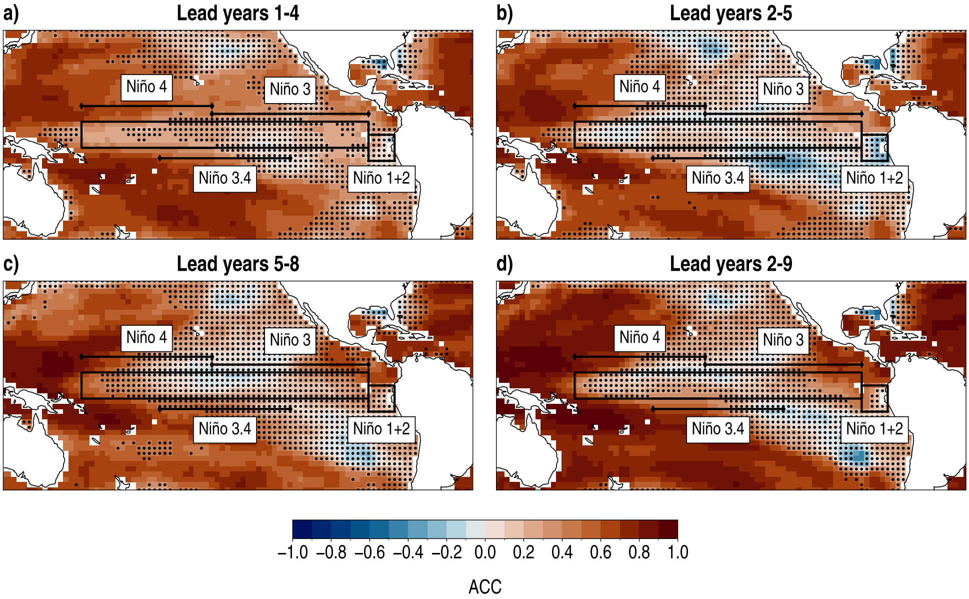

Figure B.79: Spatial distributions of ACC for the ENS40 multiannual means of SST in lead years 1–4, 2–5, 5–8 and 2–9 in DJF. Black boxes denote the regions considered in the calculation of the ENSO indices. For the Niño 3, 3.4 and 4 indices, there is a box which encompasses the common latitudes for their respective regions, whereas the straight lines delimit the range of longitudes for each one. The absence (presence) of black dots indicates (not) statistically significant results different from zero at the 95 % confidence level.

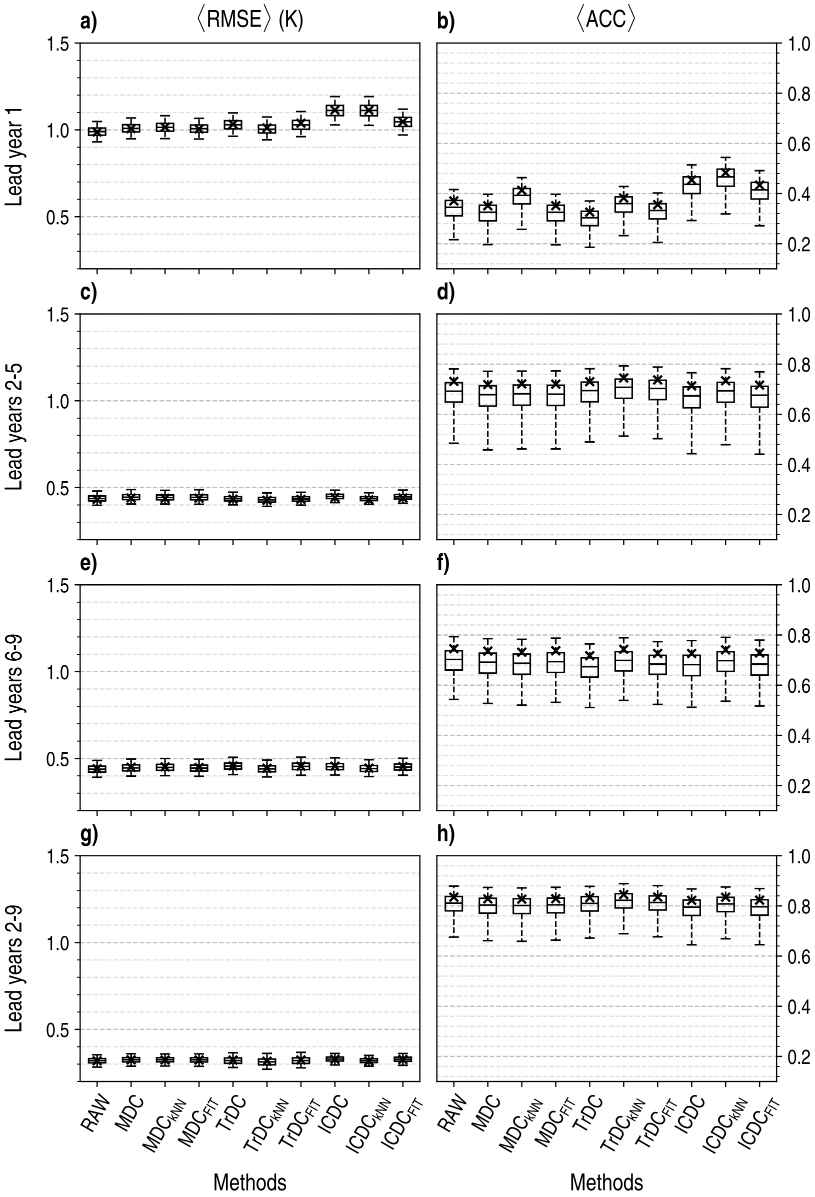

Figure B.80: As Figure B.74 but for the NA domain.

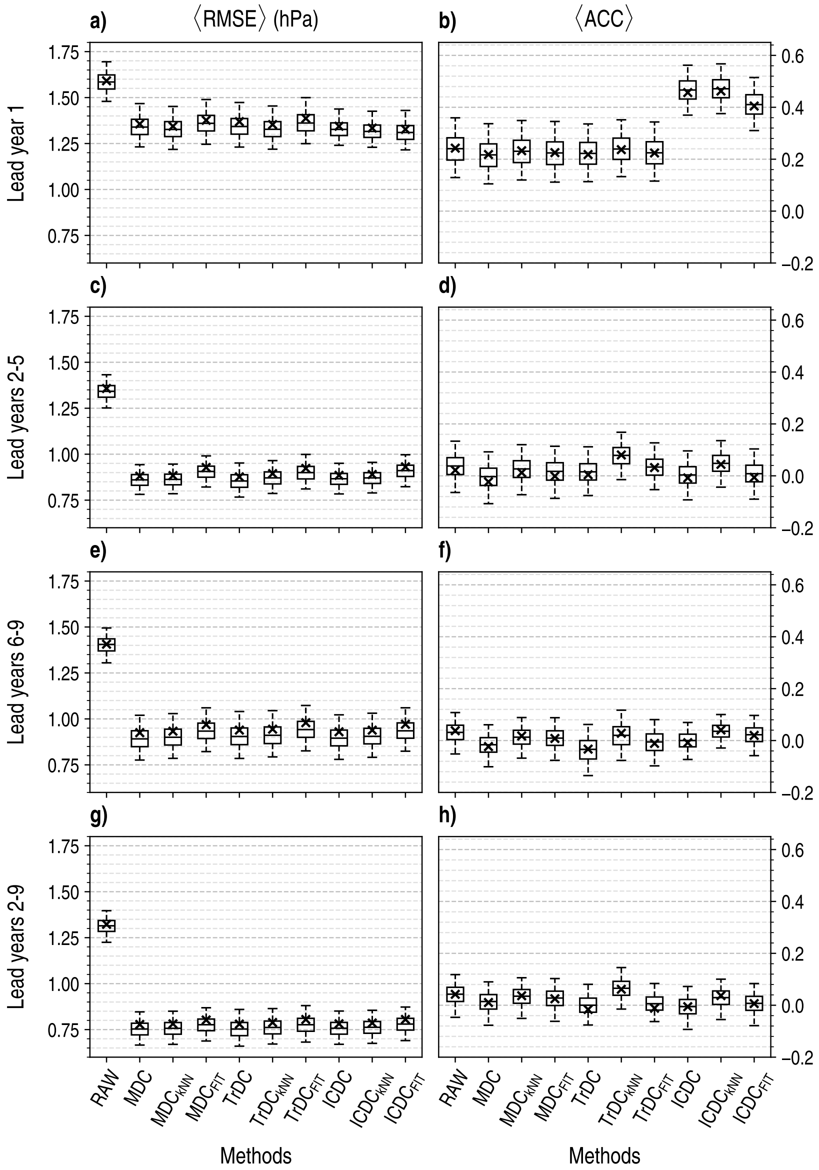

Figure B.81: As Figure B.74 but for SLP and the NA domain.

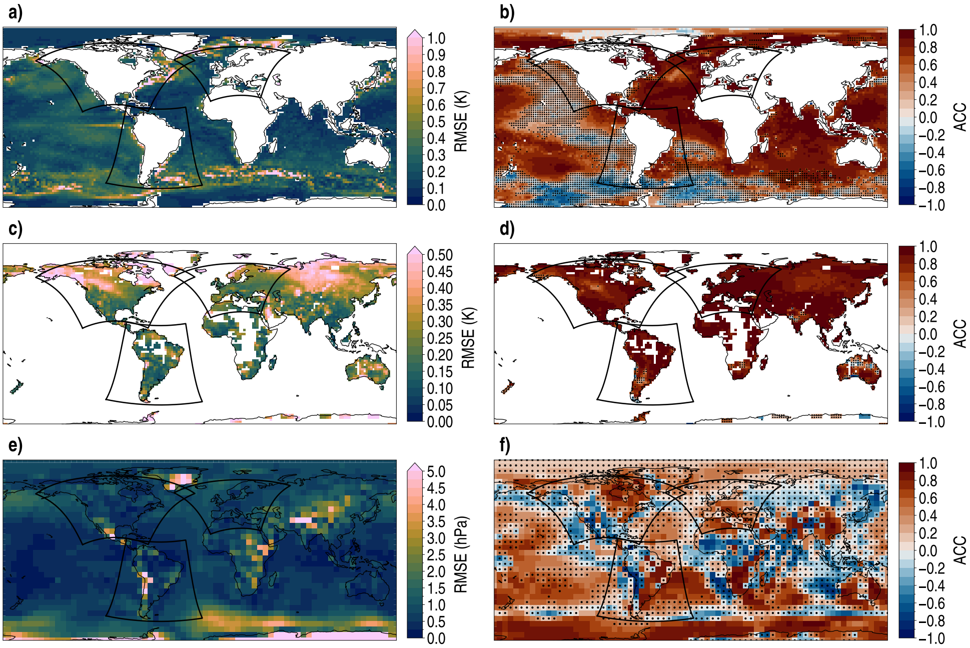

Figure B.82: Spatial distributions of RMSE (left column) and ACC (right column) for the ENS40 multiannual means of SST (top row), NSAT anomaly (middle row) and SLP (bottom row), dritf-corrected with TrDC$_{\mathrm{kNN}}$, in lead years 2–9 at annual scale. The absence (presence) of black dots in ACC panels indicates (not) statistically significant results different from zero at the 95 % confidence level. Black lines denote the boundaries of the EUR, SA and NA domains.

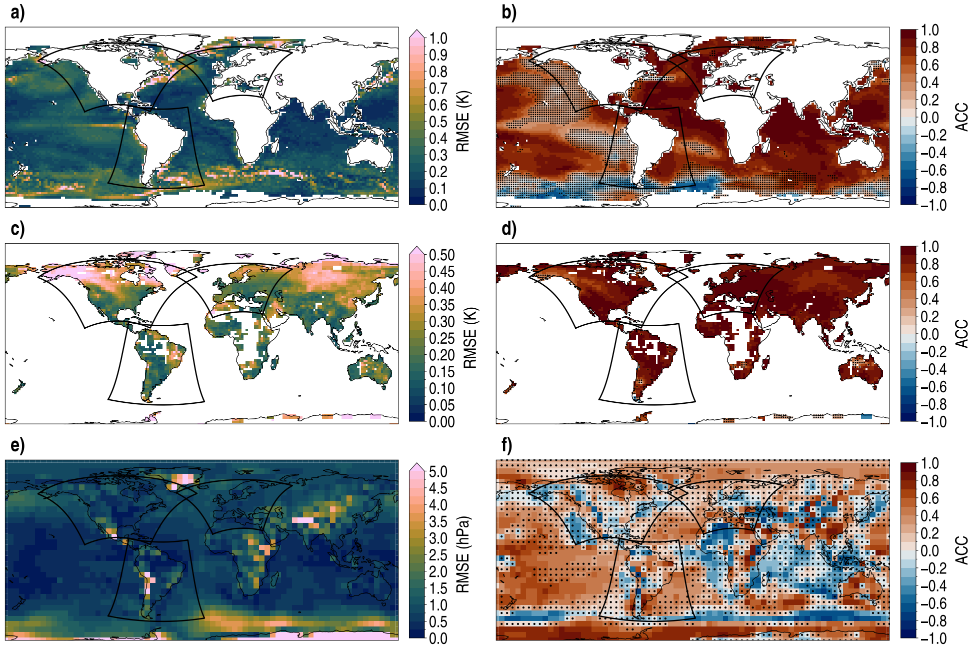

Figure B.83: As Figure B.82 but for ICDC$_{\mathrm{kNN}}$ as the drift correction method.

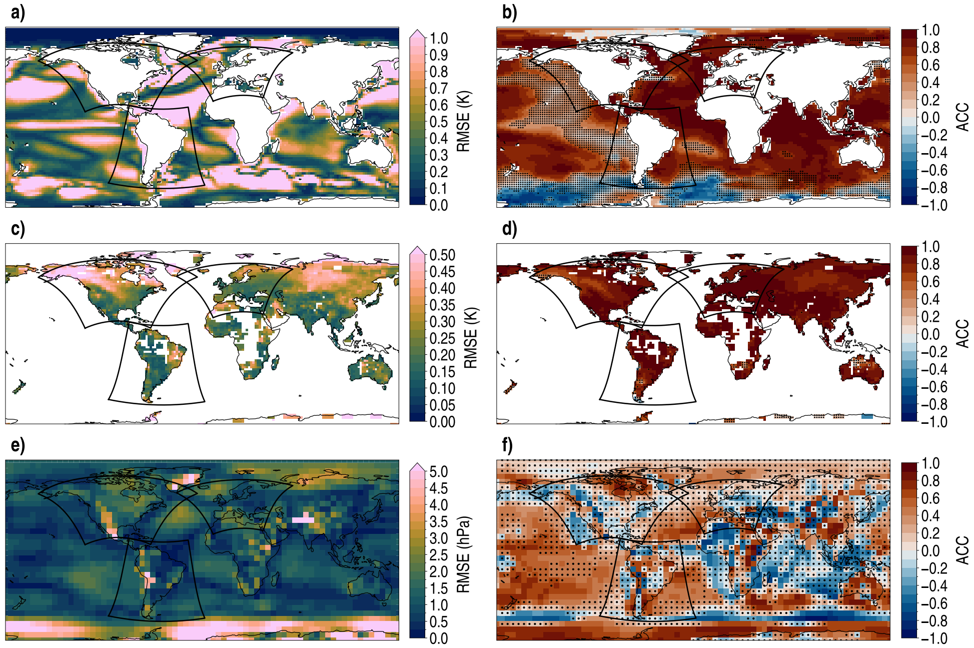

Figure B.84: As Figure B.82 but for RAW (uncorrected data).

References

AEMET. (2019). AEMET high-resolution (0.05 deg) daily gridded precipitation dataset for Peninsular Spain and Balearic Islands [Version 2.0]. Retrieved October 5, 2023, from https://www.aemet.es/es/serviciosclimaticos/cambio_climat/datos_diarios?w=2&w2=0

AEMET. (2020a). AEMET high-resolution (0.05 deg) daily gridded maximum temperature dataset for Peninsular Spain and Balearic Islands [Version 1.0]. Retrieved October 5, 2023, from https://www.aemet.es/es/serviciosclimaticos/cambio_climat/datos_diarios?w=2&w2=0

AEMET. (2020b). AEMET high-resolution (0.05 deg) daily gridded minimum temperature dataset for Peninsular Spain and Balearic Islands [Version 1.0]. Retrieved October 5, 2023, from https://www.aemet.es/es/serviciosclimaticos/cambio_climat/datos_diarios?w=2&w2=0

Ahrens, C. D. (2009). Weather Forecasting. In *Meteorology today: An introduction to weather, climate, and the environment* (9th ed., pp. 338–369). Brooks/Cole.

Allan, R., and Ansell, T. (2006). A New Globally Complete Monthly Historical Gridded Mean Sea Level Pressure Dataset (HadSLP2): 1850–2004. *Journal of Climate*, *19*(22), 5816–5842. https://doi.org/10.1175/JCLI3937.1

Amblar-Francés, M. P., Ramos-Calzado, P., Sanchis-Lladó, J., Hernanz-Lázaro, A., Peral-García, M. C., Navascués, B., Dominguez-Alonso, M., Pastor-Saavedra, M. A., and Rodríguez-Camino, E. (2020). High resolution climate change projections for the Pyrenees region. *Advances in Science and Research*, *17*, 191–208. https://doi.org/10.5194/asr-17-191-2020

Argüeso, D., Hidalgo-Muñoz, J. M., Gámiz-Fortis, S. R., Esteban-Parra, M. J., Dudhia, J., and Castro-Díez, Y. (2011). Evaluation of WRF Parameterizations for Climate Studies over Southern Spain Using a Multistep Regionalization. *Journal of Climate*, *24*(21), 5633–5651. https://doi.org/10.1175/JCLI-D-11-00073.1

Aristotle. (1952). *Meteorologica* (H. D. P. Lee, Trans.; 1st ed.). Harvard University Press.

Beck, A., Ahrens, B., and Stadlbacher, K. (2004). Impact of nesting strategies in dynamical downscaling of reanalysis data. *Geophysical Research Letters*, *31*(19), 2004GL020115. https://doi.org/10.1029/2004GL020115

Beck, H. E., McVicar, T. R., Vergopolan, N., Berg, A., Lutsko, N. J., Dufour, A., Zeng, Z., Jiang, X., Van Dijk, A. I. J. M., and Miralles, D. G. (2023). High-resolution (1 km) Köppen-Geiger maps for 1901–2099 based on constrained CMIP6 projections. *Scientific Data*, *10*(1), 724. https://doi.org/10.1038/s41597-023-02549-6

Bellucci, A., Haarsma, R., Bellouin, N., Booth, B., Cagnazzo, C., Van Den Hurk, B., Keenlyside, N., Koenigk, T., Massonnet, F., Materia, S., and Weiss, M. (2015). Advancements in decadal climate predictability: The role of nonoceanic drivers. *Reviews of Geophysics*, *53*(2), 165–202. https://doi.org/10.1002/2014RG000473

Betts, A. K. (1986). A new convective adjustment scheme. Part I: Observational and theoretical basis. *Quarterly Journal of the Royal Meteorological Society*, *112*(473), 677–691. https://doi.org/10.1002/qj.49711247307

Betts, A. K., and Miller, M. J. (1986). A new convective adjustment scheme. Part II: Single column tests using GATE wave, BOMEX, ATEX and arctic air-mass data sets. *Quarterly Journal of the Royal Meteorological Society*, *112*(473), 693–709. https://doi.org/10.1002/qj.49711247308

Betts, A. K. (2004). Understanding Hydrometeorology Using Global Models. *Bulletin of the American Meteorological Society*, *85*(11), 1673–1688. https://doi.org/10.1175/BAMS-85-11-1673

Boer, G. J., Smith, D. M., Cassou, C., Doblas-Reyes, F., Danabasoglu, G., Kirtman, B., Kushnir, Y., Kimoto, M., Meehl, G. A., Msadek, R., Mueller, W. A., Taylor, K. E., Zwiers, F., Rixen, M., Ruprich-Robert, Y., and Eade, R. (2016). The Decadal Climate Prediction Project (DCPP) contribution to CMIP6. *Geoscientific Model Development*, *9*(10), 3751–3777. https://doi.org/10.5194/gmd-9-3751-2016

Borge, R., Alexandrov, V., José Del Vas, J., Lumbreras, J., and Rodríguez, E. (2008). A comprehensive sensitivity analysis of the WRF model for air quality applications over the Iberian Peninsula. *Atmospheric Environment*, *42*(37), 8560–8574. https://doi.org/10.1016/j.atmosenv.2008.08.032

Brasseur, G. P., and Gallardo, L. (2016). Climate services: Lessons learned and future prospects. *Earth's Future*, *4*(3), 79–89. https://doi.org/10.1002/2015EF000338

Brönnimann, S., Xoplaki, E., Casty, C., Pauling, A., and Luterbacher, J. (2006). ENSO influence on Europe during the last centuries. *Climate Dynamics*, *28*(2), 181–197. https://doi.org/10.1007/s00382-006-0175-z

Bruyère, C. L., Done, J. M., Holland, G. J., and Fredrick, S. (2014). Bias corrections of global models for regional climate simulations of high-impact weather. *Climate Dynamics*, *43*(7), 1847–1856. https://doi.org/10.1007/s00382-013-2011-6

Cai, W., McPhaden, M. J., Grimm, A. M., Rodrigues, R. R., Taschetto, A. S., Garreaud, R. D., Dewitte, B., Poveda, G., Ham, Y.-G., Santoso, A., Ng, B., Anderson, W., Wang, G., Geng, T., Jo, H.-S., Marengo, J. A., Alves, L. M., Osman, M., Li, S., . . . Vera, C. (2020). Climate impacts of the El Niño–Southern Oscillation on South America. *Nature Reviews Earth & Environment*, *1*(4), 215–231. https://doi.org/10.1038/s43017-020-0040-3

Calinski, T., and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, *3*(1), 1–27. https://doi.org/10.1080/03610927408827101

Cammalleri, C., Naumann, G., Mentaschi, L., Formetta, G., Forzieri, G., Gosling, S., Bisselink, B., De Roo, A., and Feyen, L. (2020). *Global warming and drought impacts in the EU*. Publications Office of the European Union. https://doi.org/10.2760/597045,

Cardoso Pereira, S., Marta-Almeida, M., Carvalho, A. C., and Rocha, A. (2020). Extreme precipitation events under climate change in the Iberian Peninsula. *International Journal of Climatology*, *40*(2), 1255–1278. https://doi.org/10.1002/joc.6269

Cassou, C., Kushnir, Y., Hawkins, E., Pirani, A., Kucharski, F., Kang, I.-S., and Caltabiano, N. (2018). Decadal Climate Variability and Predictability: Challenges and Opportunities. *Bulletin of the American Meteorological Society*, *99*(3), 479–490. https://doi.org/10.1175/BAMS-D-16-0286.1

CESM1-CAM5-DP [HPSS archive]. (n.d.). https://portal.nersc.gov/archive/home/c/ccsm/www/CESM1-CAM5-DP/

Chen, D., Rojas, M., Samset, B., Cobb, K., Diongue Niang, A., Edwards, P., Emori, S., Faria, S., Hawkins, E., Hope, P., Huybrechts, P., Meinshausen, M., Mustafa, S., Plattner, G.-K., and Tréguier, A.-M. (2021). Framing, Context, and Methods [Type: Book Section]. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. Matthews, T. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (Eds.), *Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change* (pp. 147–286). Cambridge University Press. https://doi.org/10.1017/9781009157896.003

Chen, F., and Dudhia, J. (2001). Coupling an Advanced Land Surface–Hydrology Model with the Penn State–NCAR MM5 Modeling System. Part I: Model Implementation and Sensitivity. *Monthly Weather Review*, *129*(4), 569–585. https://doi.org/10.1175/1520-0493(2001)129<0569:CAALSH>2.0.CO;2

Choudhury, D., Sen Gupta, A., Sharma, A., Mehrotra, R., and Sivakumar, B. (2017). An Assessment of Drift Correction Alternatives for CMIP5 Decadal Predictions. *Journal of Geophysical Research: Atmospheres*, *122*(19). https://doi.org/10.1002/2017JD026900

CLIVAR. (2011). *Data and bias correction for decadal climate predictions* (CLIVAR Publication Series No. 150). International CLIVAR Office Project.

Collins, W., Rasch, P., Boville, B., McCaa, J., Williamson, D., Kiehl, J., Briegleb, B., Bitz, C., Lin, S.-J., Zhang, M., and Dai, Y. (2004). *Description of the NCAR Community Atmosphere Model (CAM 3.0)* (Artwork Size: 12360 KB Medium: application/pdf). UCAR/NCAR. https://doi.org/10.5065/D63N21CH

CORDEX. (2015). *CORDEX domains for model integrations* (Version updated in 23/10/2015). https://cordex.org/wp-content/uploads/2012/11/CORDEX-domain-description_231015.pdf

Crameri, F. (2023, October 5). *Scientific colour maps* (Version 8.0.1). Zenodo. https://doi.org/10.5281/ZENODO.1243862

Daley, R. (1991). *Atmospheric data analysis* (1st ed.). Cambridge University Press.

Danabasoglu, G., Bates, S. C., Briegleb, B. P., Jayne, S. R., Jochum, M., Large, W. G., Peacock, S., and Yeager, S. G. (2012). The CCSM4 Ocean Component. *Journal of Climate*, *25*(5), 1361–1389. https://doi.org/10.1175/JCLI-D-11-00091.1

Data Sets Available to the Community [Large ensemble community project (LENS)]. (n.d.). https://www.cesm.ucar.edu/community-projects/lens/data-sets

Decadal Prediction Large Ensemble Project output fields list [Decadal prediction large ensemble project]. (n.d.). https://www2.cesm.ucar.edu/projects/community-projects/DPLE/DPLE_output_fields/

Decommissioning of ECMWF Public Datasets Service [ECMWF confluence wiki]. (n.d.). https://confluence.ecmwf.int/display/DAC/Decommissioning+of+ECMWF+Public+Datasets+Service

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., Van De Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., … Vitart, F. (2011). The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597. https://doi.org/10.1002/qj.828

Demory, M.-E., Berthou, S., Fernández, J., Sørland, S. L., Brogli, R., Roberts, M. J., Beyerle, U., Seddon, J., Haarsma, R., Schär, C., Buonomo, E., Christensen, O. B., Ciarlo', J. M., Fealy, R., Nikulin, G., Peano, D., Putrasahan, D., Roberts, C. D., Senan, R., … Vautard, R. (2020). European daily precipitation according to EURO-CORDEX regional climate models (RCMs) and high-resolution global climate models (GCMs) from the High-Resolution Model Intercomparison Project (HighResMIP). *Geoscientific Model Development*, *13*(11), 5485–5506. https://doi.org/10.5194/gmd-13-5485-2020

Denis, B., Laprise, R., and Caya, D. (2003). Sensitivity of a regional climate model to the resolution of the lateral boundary conditions. *Climate Dynamics*, *20*(2), 107–126. https://doi.org/10.1007/s00382-002-0264-6

Denis, B., Laprise, R., Caya, D., and Côté, J. (2002). Downscaling ability of one-way nested regional climate models: the Big-Brother Experiment. *Climate Dynamics*, *18*(8), 627–646. https://doi.org/10.1007/s00382-001-0201-0

Doblas-Reyes, F., Sörensson, A., Almazroui, M., Dosio, A., Gutowski, W., Haarsma, R., Hamdi, R., Hewitson, B., Kwon, W.-T., Lamptey, B., Maraun, D., Stephenson, T., Takayabu, I., Terray, L., Turner, A., and Zuo, Z. (2021). Linking Global to Regional Climate Change. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. Matthews, T. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (Eds.), *Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change* (pp. 1363–1512). Cambridge University Press. https://doi.org/10.1017/9781009157896.012

Dunstone, N., Smith, D., Scaife, A., Hermanson, L., Eade, R., Robinson, N., Andrews, M., and Knight, J. (2016). Skilful predictions of the winter North Atlantic Oscillation one year ahead. *Nature Geoscience*, *9*(11), 809–814. https://doi.org/10.1038/ngeo2824

Earth Resources Observation And Science Center. (2017). Global 30 Arc-Second Elevation (GTOPO30). https://doi.org/10.5066/F7DF6PQS

Ek, M. B., Mitchell, K. E., Lin, Y., Rogers, E., Grunmann, P., Koren, V., Gayno, G., and Tarpley, J. D. (2003). Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *Journal of Geophysical Research: Atmospheres*, *108*, 2002JD003296. https://doi.org/10.1029/2002JD003296

Esteban-Parra, M. J., García-Valdecasas, M., Peinó-Calero, E., Romero-Jiménez, E., Yeste, P., Rosa-Cánovas, J. J., Rodríguez-Brito, A., Gámiz-Fortis, S. R., and Castro-Díez, Y. (2022). Climate Variability and Trends. In R. Zamora and M. Oliva (Eds.), *The landscape of the sierra nevada* (pp. 129–148). Springer International Publishing. https://doi.org/10.1007/978-3-030-94219-9_9

Eyring, V., Arblaster, J. M., Cionni, I., Sedláček, J., Perlwitz, J., Young, P. J., Bekki, S., Bergmann, D., Cameron-Smith, P., Collins, W. J., Faluvegi, G., Gottschaldt, K.-D., Horowitz, L. W., Kinnison, D. E., Lamarque, J.-F., Marsh, D. R., Saint-Martin, D., Shindell, D. T., Sudo, K., . . . Watanabe, S. (2013). Long-term ozone changes and associated climate impacts in CMIP5 simulations. *Journal of Geophysical Research: Atmospheres*, *118*(10), 5029–5060. https://doi.org/10.1002/jgrd.50316

Eyring, V., Gillett, N., Achuta Rao, K., Barimalala, R., Barreiro Parrillo, M., Bellouin, N., Cassou, C., Durack, P., Kosaka, Y., McGregor, S., Min, S., Morgenstern, O., and Sun, Y. (2021). Human Influence on the Climate System [Type: Book Section]. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. Matthews, T. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (Eds.), *Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change* (pp. 423–552). Cambridge University Press. https://doi.org/10.1017/9781009157896.005

FAO/UNESCO. (1978). *Soil map of the world* (Vol. 1-10). United Nations Educational, Scientific; Cultural Organization.

FAO/USDA. (2002). Hybrid STATSGO/FAO (30-second for CONUS /5-minute elsewhere) Soil Texture. Retrieved July 10, 2024, from https://ral.ucar.edu/model/noah-multiparameterization-land-surface-model-noah-mp-lsm

Feldmann, H., Pinto, J. G., Laube, N., Uhlig, M., Moemken, J., Pasternack, A., Früh, B., Pohlmann, H., and Kottmeier, C. (2019). Skill and added value of the MiKlip regional decadal prediction system for temperature over Europe. *Tellus A:*

Dynamic Meteorology and Oceanography, *71*(1), 1618678. https://doi.org/10.1080/16000870.2019.1618678

Friedl, M. A., and Land Team/EMC/NCEP. (2008). Modified IGBP MODIS 20-category vegetation (land-use) data. Retrieved July 10, 2024, from https://ral.ucar.edu/model/noah-multiparameterization-land-surface-model-noah-mp-lsm

Friedl, M. A., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A., and Huang, X. (2010). MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. *Remote Sensing of Environment*, *114*(1), 168–182. https://doi.org/10.1016/j.rse.2009.08.016

Fučkar, N. S., Volpi, D., Guemas, V., and Doblas-Reyes, F. J. (2014). A posteriori adjustment of near-term climate predictions: Accounting for the drift dependence on the initial conditions. *Geophysical Research Letters*, *41*(14), 5200–5207. https://doi.org/10.1002/2014GL060815

Fujiwara, M., Wright, J. S., Manney, G. L., Gray, L. J., Anstey, J., Birner, T., Davis, S., Gerber, E. P., Harvey, V. L., Hegglin, M. I., Homeyer, C. R., Knox, J. A., Krüger, K., Lambert, A., Long, C. S., Martineau, P., Molod, A., Monge-Sanz, B. M., Santee, M. L., . . . Zou, C.-Z. (2017). Introduction to the SPARC Reanalysis Intercomparison Project (S-RIP) and overview of the reanalysis systems. *Atmospheric Chemistry and Physics*, *17*(2), 1417–1452. https://doi.org/10.5194/acp-17-1417-2017

Fyfe, J., Fox-Kemper, B., Kopp, R., and Garner, G. (2021). Summary for Policymakers of the Working Group I Contribution to the IPCC Sixth Assessment Report - data for Figure SPM.8 (v20210809). https://doi.org/10.5285/98AF2184E13E4B91893AB72F301790DB

Gailly, J.-l., and Adler, M. (1995). *zlib: A Massively Spiffy Yet Delicately Unobtrusive Compression Library*. https://zlib.net/

Gangstø, R., Weigel, A., Liniger, M., and Appenzeller, C. (2013). Methodological aspects of the validation of decadal predictions. *Climate Research*, *55*(3), 181–200. https://doi.org/10.3354/cr01135

García-Valdecasas, M. (2018). *Climate-change projections in the Iberian Península: a study on the hydrological impacts* [Doctoral dissertation, Universidad de Granada]. https://hdl.handle.net/10481/51890

García-Valdecasas, M., Gámiz-Fortis, S. R., Romero-Jiménez, E., Rosa-Cánovas, J. J., Yeste, P., Castro-Díez, Y., and Esteban-Parra, M. J. (2021). Projected changes in the Iberian Peninsula drought characteristics. *Science of The Total Environment*, *757*, 143702. https://doi.org/10.1016/j.scitotenv.2020.143702

García-Valdecasas, M., Rosa-Cánovas, J. J., Romero-Jiménez, E., Yeste, P., Gámiz-Fortis, S. R., Castro-Díez, Y., and Esteban-Parra, M. J. (2020a). The role of the surface evapotranspiration in regional climate modelling: Evaluation and near-term future changes. *Atmospheric Research*, *237*, 104867. https://doi.org/10.1016/j.atmosres.2020.104867

García-Valdecasas, M., Yeste, P., Gámiz-Fortis, S. R., Castro-Díez, Y., and Esteban-Parra, M. J. (2020b). Future changes in land and atmospheric variables: An analysis of their couplings in the Iberian Peninsula. *Science of The Total Environment*, *722*, 137902. https://doi.org/10.1016/j.scitotenv.2020.137902

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., Da Silva, A. M., Gu, W., . . . Zhao, B. (2017). The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *Journal of Climate*, *30*(14), 5419–5454. https://doi.org/10.1175/JCLI-D-16-0758.1

Giorgi, F. (2019). Thirty Years of Regional Climate Modeling: Where Are We and Where Are We Going next? *Journal of Geophysical Research: Atmospheres*, *124*(11), 5696–5723. https://doi.org/10.1029/2018JD030094

Giorgi, F., and Gutowski, W. J. (2015). Regional Dynamical Downscaling and the CORDEX Initiative. *Annual Review of Environment and Resources*, *40*(1), 467–490. https://doi.org/10.1146/annurev-environ-102014-021217

Giorgi, F., Jones, C., and Asrar, G. R. (2009). Addressing climate information needs at the regional level: the CORDEX framework. *WMO Bulletin*, *58*(3), 175–183.

Giorgi, F., and Mearns, L. O. (1999). Introduction to special section: Regional Climate Modeling Revisited. *Journal of Geophysical Research: Atmospheres*, *104*, 6335–6352. https://doi.org/10.1029/98JD02072

Giorgi, F., Solmon, F., and Giuliani, G. (2023). *Regional Climatic Model RegCM User's Guide Version 5.0.0*. The Abdus Salam International Centre for Theoretical Physics. Trieste, Italy. https://doi.org/10.5281/zenodo.7548172

Gneiting, T., and Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437

Goddard, L. (2016). From science to service. *Science*, *353*(6306), 1366–1367. https://doi.org/10.1126/science.aag3087

Goddard, L., Kumar, A., Solomon, A., Smith, D., Boer, G., Gonzalez, P., Kharin, V., Merryfield, W., Deser, C., Mason, S. J., Kirtman, B. P., Msadek, R., Sutton, R., Hawkins, E., Fricker, T., Hegerl, G., Ferro, C. A. T., Stephenson, D. B., Meehl, G. A., . . . Delworth, T. (2013). A verification framework for interannual-to-decadal predictions experiments. *Climate Dynamics*, *40*(1), 245–272. https://doi.org/10.1007/s00382-012-1481-2

Gómez, B., and Miguez-Macho, G. (2017). The impact of wave number selection and spin-up time in spectral nudging. *Quarterly Journal of the Royal Meteorological Society*, *143*(705), 1772–1786. https://doi.org/10.1002/qj.3032

Gómez-Navarro, J. J., Raible, C. C., Bozhinova, D., Martius, O., García Valero, J. A., and Montávez, J. P. (2018). A new region-aware bias-correction method for simulated precipitation in areas of complex orography. *Geoscientific Model Development*, *11*(6), 2231–2247. https://doi.org/10.5194/gmd-11-2231-2018

Gomis, M. I., Pidcock, R., Connors, S., Harold, J., Hawkins, E., Johansen, T. G., Delmotte, V. M., Morelli, A., Nicolai, M., Pean, C., Pirani, A., and Skea, J. (2018). *IPCC Visual Style Guide for WGI Authors* [Updated June 2022]. Intergovernmental Panel on Climate Change.

Gonzalez, P. L. M., and Goddard, L. (2016). Long-lead ENSO predictability from CMIP5 decadal hindcasts. *Climate Dynamics*, *46*(9), 3127–3147. https://doi.org/10.1007/s00382-015-2757-0

Graham, R., Yun, W., Kim, J., Kumar, A., Jones, D., Bettio, L., Gagnon, N., Kolli, R., and Smith, D. (2011). Long-range forecasting and the Global Framework for Climate Services. *Climate Research*, *47*(1), 47–55. https://doi.org/10.3354/cr00963

Grell, G., Dudhia, J., and Stauffer, D. R. (1994). A description of the fifth-generation Penn State/NCAR Mesoscale Model (MM5). https://api.semanticscholar.org/CorpusID:130948085

Hamed, K. H., and Rao, A. R. (1998). A modified Mann-Kendall trend test for autocorrelated data. *Journal of Hydrology*, *204*(1), 182–196. https://doi.org/10.1016/S0022-1694(97)00125-X

Hansen, J., Ruedy, R., Sato, M., and Lo, K. (2010). Global surface temperature change. *Reviews of Geophysics*, *48*(4), RG4004. https://doi.org/10.1029/2010RG000345

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Van Kerkwijk, M. H., Brett, M., Haldane, A., Del Río, J. F., Wiebe, M., Peterson, P., . . . Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hawkins, E., and Sutton, R. (2009). The Potential to Narrow Uncertainty in Regional Climate Predictions. *Bulletin of the American Meteorological Society*, *90*(8), 1095–1108. https://doi.org/10.1175/2009BAMS2607.1

Hazeleger, W., Guemas, V., Wouters, B., Corti, S., Andreu–Burillo, I., Doblas–Reyes, F. J., Wyser, K., and Caian, M. (2013). Multiyear climate predictions using two initialization strategies. *Geophysical Research Letters*, *40*(9), 1794–1798. https://doi.org/10.1002/grl.50355

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., . . . Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Hillel, D. (1998). *Environmental soil physics*. Academic Press.

Holland, G., Done, J., Bruyere, C., Cooper, C. K., and Suzuki, A. (2010). Model Investigations of the Effects of Climate Variability and Change on Future Gulf of Mexico Tropical Cyclone Activity. *All Days*, OTC–20690–MS. https://doi.org/10.4043/20690-MS

Hong, S.-Y., Dudhia, J., and Chen, S.-H. (2004). A Revised Approach to Ice Microphysical Processes for the Bulk Parameterization of Clouds and Precipitation. *Monthly Weather Review*, *132*(1), 103–120. https://doi.org/10.1175/1520-0493(2004)132<0103:ARATIM>2.0.CO;2

Hoyer, S., and Hamman, J. (2017). xarray: N-D labeled Arrays and Datasets in Python. *Journal of Open Research Software*, *5*(1), 10. https://doi.org/10.5334/jors.148

Hu, W., Ma, W., Yang, Z.-L., Ma, Y., and Xie, Z. (2023). Sensitivity Analysis of the Noah-MP Land Surface Model for Soil Hydrothermal Simulations Over the Tibetan Plateau. *Journal of Advances in Modeling Earth Systems*, *15*(3), e2022MS003136. https://doi.org/10.1029/2022MS003136

Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., Menne, M. J., Smith, T. M., Vose, R. S., and Zhang, H.-M. (2017). Extended Reconstructed Sea Surface Temperature, Version 5 (ERSSTv5): Upgrades, Validations, and Intercomparisons. *Journal of Climate*, *30*(20), 8179–8205. https://doi.org/10.1175/JCLI-D-16-0836.1

Hunke, E. C., and Lipscomb, W. H. (2008). *CICE: the Los Alamos Sea Ice Model Documentation and Software User's Manual Version 4.0.* (Los Alamos National Laboratory Technical Report No. LA-CC-06-012). Los Alamos National Laboratory.

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Hurrell, J. W., Holland, M. M., Gent, P. R., Ghan, S., Kay, J. E., Kushner, P. J., Lamarque, J.-F., Large, W. G., Lawrence, D., Lindsay, K., Lipscomb, W. H., Long, M. C., Mahowald, N., Marsh, D. R., Neale, R. B., Rasch, P., Vavrus, S., Vertenstein, M., Bader, D., . . . Marshall, S. (2013). The Community Earth System Model: A Framework for Collaborative Research. *Bulletin of the American Meteorological Society*, *94*(9), 1339–1360. https://doi.org/10.1175/BAMS-D-12-00121.1

Hurrell, J. W., Kushnir, Y., Ottersen, G., and Visbeck, M. (2003). An overview of the North Atlantic Oscillation. In J. W. Hurrell, Y. Kushnir, G. Ottersen, and M. Visbeck (Eds.), *Geophysical monograph series* (pp. 1–35, Vol. 134). American Geophysical Union. https://doi.org/10.1029/134GM01

Hussain, M., and Mahmud, I. (2019). pyMannKendall: a python package for non parametric Mann Kendall family of trend tests. *Journal of Open Source Software*, *4*(39), 1556. https://doi.org/10.21105/joss.01556

Infanti, J. M., and Kirtman, B. P. (2016). North American rainfall and temperature prediction response to the diversity of ENSO. *Climate Dynamics*, *46*(9), 3007–3023. https://doi.org/10.1007/s00382-015-2749-0

Instituto Geográfico Nacional. (2019). *España en mapas. Una síntesis geográfica* (Centro Nacional de Información Geográfica, Ed.; 2nd ed.). Centro Nacional de Información Geográfica. Retrieved January 11, 2024, from https://doi.org/10.7419/162.06.2018

IPCC. (2021a). Annex II: Models [Gutiérrez, J M., A.-M. Tréguier (eds.)] [Type: Book Section]. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. Matthews, T. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (Eds.), *Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change* (pp. 2087–2138). Cambridge University Press. https://doi.org/10.1017/9781009157896.016

IPCC. (2021b). Annex VII: Glossary [Matthews, J.B.R., V. Möller, R. van Diemen, J.S. Fuglestvedt, V. Masson-Delmotte, C. Méndez, S. Semenov, A. Reisinger (eds.)] In V. Masson-Delmotte, P. Zhai, A. Pirani, S. Connors, C. Péan, S.

Berger, N. Caud, Y. Chen, L. Goldfarb, M. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. Matthews, T. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (Eds.), *Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change* (pp. 2215–2256). Cambridge University Press. https://doi.org/10.1017/9781009157896.022

IPCC. (2022). *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (H. O. Pörtner, D. C. Roberts, M. Tignor, E. S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, and B. Rama, Eds.). Cambridge University Press. https://doi.org/10.1017/9781009325844

Jach, L., Warrach-Sagi, K., Ingwersen, J., Kaas, E., and Wulfmeyer, V. (2020). Land Cover Impacts on Land-Atmosphere Coupling Strength in Climate Simulations With WRF Over Europe. *Journal of Geophysical Research: Atmospheres*, *125*(18), e2019JD031989. https://doi.org/10.1029/2019JD031989

Jacob, D., and Podzun, R. (1997). Sensitivity studies with the regional climate model REMO. *Meteorology and Atmospheric Physics*, *63*(1), 119–129. https://doi.org/10.1007/BF01025368

Jacob, D., Petersen, J., Eggert, B., Alias, A., Christensen, O. B., Bouwer, L. M., Braun, A., Colette, A., Déqué, M., Georgievski, G., Georgopoulou, E., Gobiet, A., Menut, L., Nikulin, G., Haensler, A., Hempelmann, N., Jones, C., Keuler, K., Kovats, S., . . . Yiou, P. (2014). EURO-CORDEX: new high-resolution climate change projections for European impact research. *Regional Environmental Change*, *14*(2), 563–578. https://doi.org/10.1007/s10113-013-0499-2

Jaeger, E. B., and Seneviratne, S. I. (2011). Impact of soil moisture–atmosphere coupling on European climate extremes and trends in a regional climate model. *Climate Dynamics*, *36*(9), 1919–1939. https://doi.org/10.1007/s00382-010-0780-8

Janjić, Z. I. (1994). The Step-Mountain Eta Coordinate Model: Further Developments of the Convection, Viscous Sublayer, and Turbulence Closure Schemes. *Monthly Weather Review*, *122*(5), 927–945. https://doi.org/10.1175/1520-0493(1994)122<0927:TSMECM>2.0.CO;2

Janjić, Z. I. (2000). Comments on "Development and Evaluation of a Convection Scheme for Use in Climate Models". *Journal of the Atmospheric Sciences*, *57*(21), 3686–3686. https://doi.org/10.1175/1520-0469(2000)057<3686:CODAEO>2.0.CO;2

Jerez, S., López-Romero, J. M., Turco, M., Lorente-Plazas, R., Gómez-Navarro, J. J., Jiménez-Guerrero, P., and Montávez, J. P. (2020). On the Spin-Up Period in WRF Simulations Over Europe: Trade-Offs Between Length and Seasonality. *Journal of Advances in Modeling Earth Systems*, *12*(4), e2019MS001945. https://doi.org/10.1029/2019MS001945

Jiménez, P. A., Dudhia, J., González-Rouco, J. F., Navarro, J., Montávez, J. P., and García-Bustamante, E. (2012). A Revised Scheme for the WRF Surface Layer Formulation. *Monthly Weather Review*, *140*(3), 898–918. https://doi.org/10.1175/MWR-D-11-00056.1

Kadow, C., Illing, S., Kunst, O., Rust, H. W., Pohlmann, H., Müller, W. A., and Cubasch, U. (2016). Evaluation of forecasts by accuracy and spread in the MiKlip decadal climate prediction system. *Meteorologische Zeitschrift*, *25*(6), 631–643. https://doi.org/10.1127/metz/2015/0639

Kaplan, A., Kushnir, Y., and Cane, M. A. (2000). Reduced Space Optimal Interpolation of Historical Marine Sea Level Pressure: 1854–1992*. *Journal of Climate*, *13*(16), 2987–3002. https://doi.org/10.1175/1520-0442(2000)013<2987:RSOIOH>2.0.CO;2

Katragkou, E., García-Díez, M., Vautard, R., Sobolowski, S., Zanis, P., Alexandri, G., Cardoso, R. M., Colette, A., Fernandez, J., Gobiet, A., Goergen, K., Karacostas, T., Knist, S., Mayer, S., Soares, P. M. M., Pytharoulis, I., Tegoulias, I., Tsikerdekis, A., and Jacob, D. (2015). Regional climate hindcast simulations within EURO-CORDEX: evaluation of a WRF multi-physics ensemble. *Geoscientific Model Development*, *8*(3), 603–618. https://doi.org/10.5194/gmd-8-603-2015

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J. M., Bates, S. C., Danabasoglu, G., Edwards, J., Holland, M., Kushner, P., Lamarque, J.-F., Lawrence, D., Lindsay, K., Middleton, A., Munoz, E., Neale, R., Oleson, K., . . . Vertenstein, M. (2015). The Community Earth System Model (CESM) Large Ensemble Project: A Community Resource for Studying Climate Change in the Presence of Internal Climate Variability. *Bulletin of the American Meteorological Society*, *96*(8), 1333–1349. https://doi.org/10.1175/BAMS-D-13-00255.1

Kharin, V. V., Boer, G. J., Merryfield, W. J., Scinocca, J. F., and Lee, W.-S. (2012). Statistical adjustment of decadal predictions in a changing climate [_eprint:

https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2012GL052647]. *Geophysical Research Letters*, *39*(19). https://doi.org/https://doi.org/10.1029/2012GL052647

Khodayar, S., Sehlinger, A., Feldmann, H., and Kottmeier, C. (2015). Sensitivity of soil moisture initialization for decadal predictions under different regional climatic conditions in Europe. *International Journal of Climatology*, *35*(8), 1899–1915. https://doi.org/10.1002/joc.4096

Kirtman, B., Power, S. B., Adedoyin, G. J., Boer, G. J., Bojariu, R., Camilloni, I., Doblas-Reyes, F. J., Fiore, A. M., Kimoto, M., Meehl, G. A., Prather, M., Sarr, A., Schär, C., Sutton, R., van Oldenborgh, G. J., Vecchi, G., and Wang, H. J. (2013). Near-term Climate Change: Projections and Predictability. In T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley (Eds.), *Climate change 2013: The physical science basis. contribution of working group i to the fifth assessment report of the intergovernmental panel on climate change* (pp. 953–1028). Cambridge University Press. https://doi.org/10.1017/CBO9781107415324.023

Knight, J. R., Folland, C. K., and Scaife, A. A. (2006). Climate impacts of the Atlantic Multidecadal Oscillation. *Geophysical Research Letters*, *33*(17), L17706. https://doi.org/10.1029/2006GL026242

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., Onogi, K., Kamahori, H., Kobayashi, C., Endo, H., Miyaoka, K., and Takahashi, K. (2015). The JRA-55 Reanalysis: General Specifications and Basic Characteristics. *Journal of the Meteorological Society of Japan. Ser. II*, *93*(1), 5–48. https://doi.org/10.2151/jmsj.2015-001

Kotamarthi, R., Hayhoe, K., Mearns, L., Wuebbles, D., Jacobs, J., and Jurado, J. (2021, February 11). *Downscaling Techniques for High-Resolution Climate Projections: From Global Change to Local Impacts* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781108601269

Kothe, S., Tödter, J., and Ahrens, B. (2016). Strategies for soil initialization of regional decadal climate predictions. *Meteorologische Zeitschrift*, *25*(6), 775–794. https://doi.org/10.1127/metz/2016/0729

Kotlarski, S., Keuler, K., Christensen, O. B., Colette, A., Déqué, M., Gobiet, A., Goergen, K., Jacob, D., Lüthi, D., Van Meijgaard, E., Nikulin, G., Schär, C., Teichmann, C., Vautard, R., Warrach-Sagi, K., and Wulfmeyer, V. (2014). Regional climate modeling on European scales: a joint standard evaluation of the EURO-

CORDEX RCM ensemble. *Geoscientific Model Development*, *7*(4), 1297–1333. https://doi.org/10.5194/gmd-7-1297-2014

Kruschke, T., Rust, H. W., Kadow, C., Müller, W. A., Pohlmann, H., Leckebusch, G. C., and Ulbrich, U. (2016). Probabilistic evaluation of decadal prediction skill regarding Northern Hemisphere winter storms. *Meteorologische Zeitschrift*, *25*(6), 721–738. https://doi.org/10.1127/metz/2015/0641

Lamarque, J.-F., Kyle, G. P., Meinshausen, M., Riahi, K., Smith, S. J., Van Vuuren, D. P., Conley, A. J., and Vitt, F. (2011). Global and regional evolution of short-lived radiatively-active gases and aerosols in the Representative Concentration Pathways. *Climatic Change*, *109*(1), 191–212. https://doi.org/10.1007/s10584-011-0155-0

Lamarque, J.-F., Bond, T. C., Eyring, V., Granier, C., Heil, A., Klimont, Z., Lee, D., Liousse, C., Mieville, A., Owen, B., Schultz, M. G., Shindell, D., Smith, S. J., Stehfest, E., Van Aardenne, J., Cooper, O. R., Kainuma, M., Mahowald, N., McConnell, J. R., . . . Van Vuuren, D. P. (2010). Historical (1850–2000) gridded anthropogenic and biomass burning emissions of reactive gases and aerosols: methodology and application. *Atmospheric Chemistry and Physics*, *10*(15), 7017–7039. https://doi.org/10.5194/acp-10-7017-2010

Laprise, R. (1992). The Euler Equations of Motion with Hydrostatic Pressure as an Independent Variable. *Monthly Weather Review*, *120*(1), 197–207. https://doi.org/10.1175/1520-0493(1992)120<0197:TEEOMW>2.0.CO;2

Laprise, R. (2008). Regional climate modelling. *Journal of Computational Physics*, *227*(7), 3641–3666. https://doi.org/10.1016/j.jcp.2006.10.024

Lawrence, D. M., Oleson, K. W., Flanner, M. G., Thornton, P. E., Swenson, S. C., Lawrence, P. J., Zeng, X., Yang, Z.-L., Levis, S., Sakaguchi, K., Bonan, G. B., and Slater, A. G. (2011). Parameterization improvements and functional and structural advances in Version 4 of the Community Land Model. *Journal of Advances in Modeling Earth Systems*, *3*(3), M03001. https://doi.org/10.1029/2011MS000045

Lee, J.-Y., Marotzke, J., Bala, G., Cao, L., Corti, S., Dunne, J., Engelbrecht, F., Fischer, E., Fyfe, J., Jones, C., Maycock, A., Mutemi, J., Ndiaye, O., Panickal, S., and Zhou, T. (2021). Future Global Climate: Scenario-Based Projections and Near-Term Information [Type: Book Section]. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. Matthews, T. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (Eds.), *Climate change 2021: The* *physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change* (pp. 553–672). Cambridge University Press. https://doi.org/10.1017/9781009157896.006

Lenssen, N. J. L., Schmidt, G. A., Hansen, J. E., Menne, M. J., Persin, A., Ruedy, R., and Zyss, D. (2019). Improvements in the GISTEMP Uncertainty Model. *Journal of Geophysical Research: Atmospheres*, *124*(12), 6307–6326. https://doi.org/10.1029/2018JD029522

Lorenz, P., and Jacob, D. (2005). Influence of regional scale information on the global circulation: A two-way nesting climate simulation. *Geophysical Research Letters*, *32*(18), 2005GL023351. https://doi.org/10.1029/2005GL023351

Lorenz, R., Davin, E. L., and Seneviratne, S. I. (2012). Modeling land-climate coupling in Europe: Impact of land surface representation on climate variability and extremes. *Journal of Geophysical Research: Atmospheres*, *117*, 2012JD017755. https://doi.org/10.1029/2012JD017755

Lorenzo, M. N., and Alvarez, I. (2022). Future changes of hot extremes in Spain: towards warmer conditions. *Natural Hazards*, *113*(1), 383–402. https://doi.org/10.1007/s11069-022-05306-x

Lorenzo, M., and Alvarez, I. (2020). Climate change patterns in precipitation over Spain using CORDEX projections for 2021–2050. *Science of The Total Environment*, *723*, 138024. https://doi.org/10.1016/j.scitotenv.2020.138024

Marbaix, P., Gallée, H., Brasseur, O., and Van Ypersele, J.-P. (2003). Lateral Boundary Conditions in Regional Climate Models: A Detailed Study of the Relaxation Procedure. *Monthly Weather Review*, *131*(3), 461–479. https://doi.org/10.1175/1520-0493(2003)131<0461:LBCIRC>2.0.CO;2

Marotzke, J., Müller, W. A., Vamborg, F. S. E., Becker, P., Cubasch, U., Feldmann, H., Kaspar, F., Kottmeier, C., Marini, C., Polkova, I., Prömmel, K., Rust, H. W., Stammer, D., Ulbrich, U., Kadow, C., Köhl, A., Kröger, J., Kruschke, T., Pinto, J. G., . . . Ziese, M. (2016). MiKlip: A National Research Project on Decadal Climate Prediction. *Bulletin of the American Meteorological Society*, *97*(12), 2379–2394. https://doi.org/10.1175/BAMS-D-15-00184.1

Marsh, D. R., Mills, M. J., Kinnison, D. E., Lamarque, J.-F., Calvo, N., and Polvani, L. M. (2013). Climate Change from 1850 to 2005 Simulated in CESM1(WACCM). *Journal of Climate*, *26*(19), 7372–7391. https://doi.org/10.1175/JCLI-D-12-00558.1

Matei, D., Pohlmann, H., Jungclaus, J., Müller, W., Haak, H., and Marotzke, J. (2012). Two Tales of Initializing Decadal Climate Prediction Experiments with the

ECHAM5/MPI-OM Model. *Journal of Climate*, *25*(24), 8502–8523. https://doi.org/10.1175/JCLI-D-11-00633.1

Matte, D., Laprise, R., Thériault, J. M., and Lucas-Picher, P. (2017). Spatial spin-up of fine scales in a regional climate model simulation driven by low-resolution boundary conditions. *Climate Dynamics*, *49*(1), 563–574. https://doi.org/10.1007/s00382-016-3358-2

McFarlane, N. (2011). Parameterizations: representing key processes in climate models without resolving them. *WIREs Climate Change*, *2*(4), 482–497. https://doi.org/10.1002/wcc.122

Meehl, G. A., Goddard, L., Boer, G., Burgman, R., Branstator, G., Cassou, C., Corti, S., Danabasoglu, G., Doblas-Reyes, F., Hawkins, E., Karspeck, A., Kimoto, M., Kumar, A., Matei, D., Mignot, J., Msadek, R., Navarra, A., Pohlmann, H., Rienecker, M., . . . Yeager, S. (2014). Decadal Climate Prediction: An Update from the Trenches. *Bulletin of the American Meteorological Society*, *95*(2), 243–267. https://doi.org/10.1175/BAMS-D-12-00241.1

Meehl, G. A., Goddard, L., Murphy, J., Stouffer, R. J., Boer, G., Danabasoglu, G., Dixon, K., Giorgetta, M. A., Greene, A. M., Hawkins, E., Hegerl, G., Karoly, D., Keenlyside, N., Kimoto, M., Kirtman, B., Navarra, A., Pulwarty, R., Smith, D., Stammer, D., and Stockdale, T. (2009). Decadal Prediction: Can It Be Skillful? *Bulletin of the American Meteorological Society*, *90*(10), 1467–1486. https://doi.org/10.1175/2009BAMS2778.1

Meehl, G. A., Richter, J. H., Teng, H., Capotondi, A., Cobb, K., Doblas-Reyes, F., Donat, M. G., England, M. H., Fyfe, J. C., Han, W., Kim, H., Kirtman, B. P., Kushnir, Y., Lovenduski, N. S., Mann, M. E., Merryfield, W. J., Nieves, V., Pegion, K., Rosenbloom, N., . . . Xie, S.-P. (2021). Initialized Earth System prediction from subseasonal to decadal timescales. *Nature Reviews Earth & Environment*, *2*(5), 340–357. https://doi.org/10.1038/s43017-021-00155-x

Meinshausen, M., Smith, S. J., Calvin, K., Daniel, J. S., Kainuma, M. L. T., Lamarque, J.-F., Matsumoto, K., Montzka, S. A., Raper, S. C. B., Riahi, K., Thomson, A., Velders, G. J. M., and Van Vuuren, D. P. (2011). The RCP greenhouse gas concentrations and their extensions from 1765 to 2300. *Climatic Change*, *109*(1), 213–241. https://doi.org/10.1007/s10584-011-0156-z

Messmer, M., Gómez-Navarro, J. J., and Raible, C. C. (2017). Sensitivity experiments on the response of Vb cyclones to sea surface temperature and soil moisture changes. *Earth System Dynamics*, *8*(3), 477–493. https://doi.org/10.5194/esd-8-477-2017

Met Office. (2010). *Cartopy: a cartographic python library with a Matplotlib interface*. https://scitools.org.uk/cartopy

Miguez-Macho, G., Stenchikov, G. L., and Robock, A. (2004). Spectral nudging to eliminate the effects of domain position and geometry in regional climate model simulations. *Journal of Geophysical Research: Atmospheres*, *109*, 2003JD004495. https://doi.org/10.1029/2003JD004495

Miller, D. A., and White, R. A. (1998). A Conterminous United States Multilayer Soil Characteristics Dataset for Regional Climate and Hydrology Modeling. *Earth Interactions*, *2*(2), 1–26. https://doi.org/10.1175/1087-3562(1998)002<0001:ACUSMS>2.3.CO;2

Mittelbach, F., and Schöpf, R. (1989). With LaTeX into the Nineties. *TUGboat Conference Proceedings*, *10*(4), 681–690. https://tug.org/TUGboat/Articles/tb10-4/tb26mitt.pdf

Model [CORDEX-AustralAsia wikipage]. (n.d.). http://cordex-australasia.wikidot.com/models

Monaghan, A., Steinhoff, D., Bruyere, C., and Yates, D. (2014). NCAR CESM Global Bias-Corrected CMIP5 Output to Support WRF/MPAS Research [Artwork Size: 13.150 Tbytes Pages: 13.150 Tbytes]. https://doi.org/10.5065/D6DJ5CN4

Monin, A. S., and Obukhov, A. M. (1954). Basic laws of turulent mixing in the atmosphere near the ground [in Russian]. *Tr. Inst. Teor. Geofiz. Akad. Nauk SSSR*, *24*, 1963–1987.

Murphy, A. H. (1988). Skill Scores Based on the Mean Square Error and Their Relationships to the Correlation Coefficient. *Monthly Weather Review*, *116*(12), 2417–2424. https://doi.org/10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2

Murphy, A. H., and Epstein, E. S. (1989). Skill Scores and Correlation Coefficients in Model Verification. *Monthly Weather Review*, *117*(3), 572–582. https://doi.org/10.1175/1520-0493(1989)117<0572:SSACCI>2.0.CO;2

Navascués, B., Calvo, J., Morales, G., Santos, C., Callado, A., Cansado, A., Cuxart, J., Díez, M., Del Río, P., Escribà, P., García-Colombo, O., García-Moya, J., Geijo, C., Gutiérrez, E., Hortal, M., Martínez, I., Orfila, B., Parodi, J., Rodríguez, E., . . . Simarro, J. (2013). Long-term verification of HIRLAM and ECMWF forecasts over Southern Europe. *Atmospheric Research*, *125-126*, 20–33. https://doi.org/10.1016/j.atmosres.2013.01.010

Navascués, B., Rodríguez, E., Ayuso, J., and Järvenoja, S. (2003). *Analysis of surface variables and parameterization of surface processes in HIRLAM. Part II: Seasonal assimilation experiment* (No. 58). Sveriges meteorologiska och hydrologiska institut (SMHI). Norrköping, Sweden. https://hdl.handle.net/20.500.11765/12004

Nelder, J. A., and Mead, R. (1965). A Simplex Method for Function Minimization. *The Computer Journal*, *7*(4), 308–313. https://doi.org/10.1093/comjnl/7.4.308

North, G. R., Bell, T. L., Cahalan, R. F., and Moeng, F. J. (1982). Sampling Errors in the Estimation of Empirical Orthogonal Functions. *Monthly Weather Review*, *110*(7), 699–706. https://doi.org/10.1175/1520-0493(1982)110<0699:SEITEO>2.0.CO;2

Omrani, H., Drobinski, P., and Dubos, T. (2012). Spectral nudging in regional climate modelling: how strongly should we nudge? *Quarterly Journal of the Royal Meteorological Society*, *138*(668), 1808–1813. https://doi.org/10.1002/qj.1894

O'Neill, B. C., Tebaldi, C., Van Vuuren, D. P., Eyring, V., Friedlingstein, P., Hurtt, G., Knutti, R., Kriegler, E., Lamarque, J.-F., Lowe, J., Meehl, G. A., Moss, R., Riahi, K., and Sanderson, B. M. (2016). The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6. *Geoscientific Model Development*, *9*(9), 3461–3482. https://doi.org/10.5194/gmd-9-3461-2016

Paeth, H., Li, J., Pollinger, F., Müller, W. A., Pohlmann, H., Feldmann, H., and Panitz, H.-J. (2019). An effective drift correction for dynamical downscaling of decadal global climate predictions. *Climate Dynamics*, *52*(3), 1343–1357. https://doi.org/10.1007/s00382-018-4195-2

Paeth, H., Paxian, A., Sein, D. V., Jacob, D., Panitz, H.-J., Warscher, M., Fink, A. H., Kunstmann, H., Breil, M., Engel, T., Krause, A., Toedter, J., and Ahrens, B. (2017). Decadal and multi-year predictability of the West African monsoon and the role of dynamical downscaling. *Meteorologische Zeitschrift*, *26*(4), 363–377. https://doi.org/10.1127/metz/2017/0811

Pasternack, A., Bhend, J., Liniger, M. A., Rust, H. W., Müller, W. A., and Ulbrich, U. (2018). Parametric decadal climate forecast recalibration (DeFoReSt 1.0). *Geoscientific Model Development*, *11*(1), 351–368. https://doi.org/10.5194/gmd-11-351-2018

Pasternack, A., Grieger, J., Rust, H. W., and Ulbrich, U. (2021). Recalibrating decadal climate predictions – what is an adequate model for the drift? *Geoscientific Model Development*, *14*(7), 4335–4355. https://doi.org/10.5194/gmd-14-4335-2021

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peral-García, C., Navascués Fernández-Victorio, B., and Ramos-Calzado, P. (2017). *Serie de precipitación diaria en rejilla con fines climáticos* (Nota técnica de AEMET No. 24). Agencia Estatal de Meteorología (AEMET)). https://doi.org/10.31978/014-17-009-5

Pleim, J. E. (2007). A Combined Local and Nonlocal Closure Model for the Atmospheric Boundary Layer. Part I: Model Description and Testing. *Journal of Applied Meteorology and Climatology*, *46*(9), 1383–1395. https://doi.org/10.1175/JAM2539.1

Prein, A. F., Gobiet, A., Truhetz, H., Keuler, K., Goergen, K., Teichmann, C., Fox Maule, C., Van Meijgaard, E., Déqué, M., Nikulin, G., Vautard, R., Colette, A., Kjellström, E., and Jacob, D. (2016). Precipitation in the EURO-CORDEX 0.11 and 0.44 simulations: high resolution, high benefits? *Climate Dynamics*, *46*(1), 383–412. https://doi.org/10.1007/s00382-015-2589-y

Preisendorfer, R. W. (1988). *Principal component analysis in meteorology and oceanography* (C. D. Mobley, Ed.). Elsevier.

Qualitative colour schemes [Paul tol's notes]. (n.d.). https://personal.sron.nl/~pault/

Queralt, S., Hernández, E., Barriopedro, D., Gallego, D., Ribera, P., and Casanova, C. (2009). North Atlantic Oscillation influence and weather types associated with winter total and extreme precipitation events in Spain. *Atmospheric Research*, *94*(4), 675–683. https://doi.org/10.1016/j.atmosres.2009.09.005

Quintana-Seguí, P., Peral, C., Turco, M., Llasat, M., and Martin, E. (2016). Meteorological Analysis Systems in North-East Spain: Validation of SAFRAN and SPAN. *Journal of Environmental Informatics*. https://doi.org/10.3808/jei.201600335

Radu, R., Déqué, M., and Somot, S. (2008). Spectral nudging in a spectral regional climate model. *Tellus A: Dynamic Meteorology and Oceanography*, *60*(5), 898. https://doi.org/10.1111/j.1600-0870.2008.00341.x

Randall, D. A., Bitz, C. M., Danabasoglu, G., Denning, A. S., Gent, P. R., Gettelman, A., Griffies, S. M., Lynch, P., Morrison, H., Pincus, R., and Thuburn, J. (2019). 100 Years of Earth System Model Development. *Meteorological Monographs*, *59*, 12.1–12.66. https://doi.org/10.1175/AMSMONOGRAPHS-D-18-0018.1

Rew, R., and Davis, G. (1990). NetCDF: an interface for scientific data access [Publisher: Institute of Electrical and Electronics Engineers (IEEE)]. *IEEE Computer Graphics and Applications*, *10*(4), 76–82. https://doi.org/10.1109/38.56302

Reyers, M., Feldmann, H., Mieruch, S., Pinto, J. G., Uhlig, M., Ahrens, B., Früh, B., Modali, K., Laube, N., Moemken, J., Müller, W., Schädler, G., and Kottmeier, C. (2019). Development and prospects of the regional MiKlip decadal prediction system over Europe: predictive skill, added value of regionalization, and ensemble size dependency. *Earth System Dynamics*, *10*(1), 171–187. https://doi.org/10.5194/esd-10-171-2019

Ríos-Cornejo, D., Penas, Á., Álvarez-Esteban, R., and Del Río, S. (2015a). Links between teleconnection patterns and precipitation in Spain. *Atmospheric Research*, *156*, 14–28. https://doi.org/10.1016/j.atmosres.2014.12.012

Ríos-Cornejo, D., Penas, Á., Álvarez-Esteban, R., and Del Río, S. (2015b). Links between teleconnection patterns and mean temperature in Spain. *Theoretical and Applied Climatology*, *122*(1), 1–18. https://doi.org/10.1007/s00704-014-1256-2

Rodríguez, E., Navascués, B., Ayuso, J., and Järvenoja, S. (2003). *Analysis of surface variables and parameterization of surface processes in HIRLAM. Part I: Approach and verification by parallel runs* (No. 58). Sveriges meteorologiska och hydrologiska institut (SMHI). Norrköping, Sweden. https://hdl.handle.net/20.500.11765/12003

Rohli, R. V., and Vega, A. J. (2018). *Climatology* (4th ed.). Jones & Bartlett Learning.

Rosa-Cánovas, J. J., García-Valdecasas, M., Romero-Jiménez, E., Yeste, P., Gámiz-Fortis, S. R., Castro-Díez, Y., and Esteban-Parra, M. J. (2023). Drift Correction and Sub-Ensemble Predictive Skill Evaluation of the Decadal Prediction Large Ensemble With Application to Regional Studies. *Journal of Geophysical Research: Atmospheres*, *128*(22), e2023JD039709. https://doi.org/10.1029/2023JD039709

San-Miguel-Ayanz, J., Durrant, T., Boca, R., Libertà, G., Branco, A., de Rigo, D., Ferrari, D., Mianti, P., Artés-Vivancos, T., Costa, H., Lana, F., Löffler, P., Nuijten, D., Ahlgren, A. C., and Leray, T. (2018). *Forest fires in Europe, Middle East and North Africa 2017*. Publications Office of the European Union. https://doi.org/10.2760/663443

Santer, B. D., Wigley, T. M. L., Boyle, J. S., Gaffen, D. J., Hnilo, J. J., Nychka, D., Parker, D. E., and Taylor, K. E. (2000). Statistical significance of trends and trend differences in layer-average atmospheric temperature time series. *Journal of Geophysical Research: Atmospheres*, *105*, 7337–7356. https://doi.org/10.1029/1999JD901105

Santos-Alamillos, F. J., Pozo-Vázquez, D., Ruiz-Arias, J. A., Lara-Fanego, V., and Tovar-Pescador, J. (2013). Analysis of WRF Model Wind Estimate Sensitivity to

Physics Parameterization Choice and Terrain Representation in Andalusia (Southern Spain). *Journal of Applied Meteorology and Climatology*, *52*(7), 1592–1609. https://doi.org/10.1175/JAMC-D-12-0204.1

Scaife, A. A., and Smith, D. (2018). A signal-to-noise paradox in climate science. *npj Climate and Atmospheric Science*, *1*(1), 28. https://doi.org/10.1038/s41612-018-0038-4

Schalnat, G. E., Dilger, A., Randers-Pehrson, G., Truta, C., and The PNG Reference Library Authors. (1995). *PNG Reference Library: libpng*. http://www.libpng.org/

Schulzweida, U. (2023). CDO User Guide [Publisher: Zenodo Version Number: 2.3.0]. https://doi.org/10.5281/ZENODO.10020800

Seck, A., Welty, C., and Maxwell, R. M. (2015). Spin-up behavior and effects of initial conditions for an integrated hydrologic model. *Water Resources Research*, *51*(4), 2188–2210. https://doi.org/10.1002/2014WR016371

Sen, P. K. (1968). Estimates of the Regression Coefficient Based on Kendall's Tau. *Journal of the American Statistical Association*, *63*(324), 1379–1389. https://doi.org/10.1080/01621459.1968.10480934

Seneviratne, S. I., Corti, T., Davin, E. L., Hirschi, M., Jaeger, E. B., Lehner, I., Orlowsky, B., and Teuling, A. J. (2010). Investigating soil moisture–climate interactions in a changing climate: A review. *Earth-Science Reviews*, *99*(3), 125–161. https://doi.org/10.1016/j.earscirev.2010.02.004

Sienz, F., Müller, W. A., and Pohlmann, H. (2016). Ensemble size impact on the decadal predictive skill assessment. *Meteorologische Zeitschrift*, *25*(6), 645–655. https://doi.org/10.1127/metz/2016/0670

Skamarock, W., Klemp, J., Dudhia, J., Gill, D., Barker, D., Wang, W., Huang, X.-Y., and Duda, M. (2008, June). *A Description of the Advanced Research WRF Version 3* (NCAR technical note No. NCAR/TN-475+STR) (Artwork Size: 1002 KB Medium: application/pdf). National Center for Atmospheric Research. https://doi.org/10.5065/D68S4MVH

Smith, D. M., Eade, R., Scaife, A. A., Caron, L.-P., Danabasoglu, G., DelSole, T. M., Delworth, T., Doblas-Reyes, F. J., Dunstone, N. J., Hermanson, L., Kharin, V., Kimoto, M., Merryfield, W. J., Mochizuki, T., Müller, W. A., Pohlmann, H., Yeager, S., and Yang, X. (2019). Robust skill of decadal climate predictions. *npj Climate and Atmospheric Science*, *2*(1), 13. https://doi.org/10.1038/s41612-019-0071-y

Smith, D. M., Scaife, A. A., Eade, R., Athanasiadis, P., Bellucci, A., Bethke, I., Bilbao, R., Borchert, L. F., Caron, L.-P., Counillon, F., Danabasoglu, G., Delworth, T., Doblas-Reyes, F. J., Dunstone, N. J., Estella-Perez, V., Flavoni, S., Hermanson, L., Keenlyside, N., Kharin, V., . . . Zhang, L. (2020). North Atlantic climate far more predictable than models imply. *Nature*, *583*(7818), 796–800. https://doi.org/10.1038/s41586-020-2525-0

Soil Survey Division Staff. (1993). *Soil Survey Manual*. United States Department of Agriculture.

SPARC. (2022). *SPARC Reanalysis Intercomparison Project (S-RIP) Final Report* (Series: SPARC report). https://elib.dlr.de/148623/

Stevens, B., Giorgetta, M., Esch, M., Mauritsen, T., Crueger, T., Rast, S., Salzmann, M., Schmidt, H., Bader, J., Block, K., Brokopf, R., Fast, I., Kinne, S., Kornblueh, L., Lohmann, U., Pincus, R., Reichler, T., and Roeckner, E. (2013). Atmospheric component of the MPI-M Earth System Model: ECHAM6. *Journal of Advances in Modeling Earth Systems*, *5*(2), 146–172. https://doi.org/10.1002/jame.20015

Storch, H. v., Langenberg, H., and Feser, F. (2000). A Spectral Nudging Technique for Dynamical Downscaling Purposes. *Monthly Weather Review*, *128*(10), 3664–3673. https://doi.org/10.1175/1520-0493(2000)128<3664:ASNTFD>2.0.CO;2

Strobach, E., and Bel, G. (2019). Regional Decadal Climate Predictions Using an Ensemble of WRF Parameterizations Driven by the MIROC5 GCM. *Journal of Applied Meteorology and Climatology*, *58*(3), 527–549. https://doi.org/10.1175/JAMC-D-18-0051.1

Taylor, K. E., Stouffer, R. J., and Meehl, G. A. (2012). An Overview of CMIP5 and the Experiment Design. *Bulletin of the American Meteorological Society*, *93*(4), 485–498. https://doi.org/10.1175/BAMS-D-11-00094.1

Teutschbein, C., and Seibert, J. (2012). Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods. *Journal of Hydrology*, *456-457*, 12–29. https://doi.org/10.1016/j.jhydrol.2012.05.052

The Board of Trustees of the University of Illinois and The HDF Group. (1998). *Hierarchical Data Format 5: HDF5*. https://www.hdfgroup.org/

The Inkscape Team. (2003). *Inkscape*. https://inkscape.org/

The Open MPI Team. (2004). *Open MPI: Open Source High Performance Computing*. https://www.open-mpi.org

Trenberth, K. E., and Shea, D. J. (2006). Atlantic hurricanes and natural variability in 2005. *Geophysical Research Letters*, *33*(12), L12704. https://doi.org/10.1029/2006GL026894

Trenberth, K. E., and Stepaniak, D. P. (2001). Indices of El Niño Evolution. *Journal of Climate*, *14*(8), 1697–1701. https://doi.org/10.1175/1520-0442(2001)014<1697:LIOENO>2.0.CO;2

Trigo, R. M., Pozo-Vázquez, D., Osborn, T. J., Castro-Díez, Y., Gámiz-Fortis, S., and Esteban-Parra, M. J. (2004). North Atlantic oscillation influence on precipitation, river flow and water resources in the Iberian Peninsula. *International Journal of Climatology*, *24*(8), 925–944. https://doi.org/10.1002/joc.1048

Turco, M., Rosa-Cánovas, J. J., Bedia, J., Jerez, S., Montávez, J. P., Llasat, M. C., and Provenzale, A. (2018). Exacerbated fires in Mediterranean Europe due to anthropogenic warming projected with non-stationary climate-fire models. *Nature Communications*, *9*(1), 3821. https://doi.org/10.1038/s41467-018-06358-z

Turco, M., Sanna, A., Herrera, S., Llasat, M.-C., and Gutiérrez, J. M. (2013). Large biases and inconsistent climate change signals in ENSEMBLES regional projections. *Climatic Change*, *120*(4), 859–869. https://doi.org/10.1007/s10584-013-0844-y

Undén, P., Rontu, L., Järvinen, H., Lynch, P., Calvo, J., Cats, G., Cuxart, J., Eerola, K., Fortelius, C., García-Moya, J.-A., Jones, C., Lenderlink, G., McDonald, A., McGrath, R., Navascués, B., Nielsen, N. W., Odegaard, V., Rodríguez, E., Rummukainen, M., . . . Tijm, A. (2002). *HIRLAM-5 Scientific Documentation*. Sveriges meteorologiska och hydrologiska institut (SMHI). Norrköping, Sweden. http://hdl.handle.net/20.500.11765/6323

Unified Noah LSM [Research applications laboratory]. (n.d.). https://ral.ucar.edu/model/unified-noah-lsm

University of British Columbia, Image Power, Inc., and Adams, M. D. (1999). *JasPer*. https://www.ece.uvic.ca/~frodo/jasper/

Vidale, P. L., Lüthi, D., Wegmann, R., and Schär, C. (2007). European summer climate variability in a heterogeneous multi-model ensemble. *Climatic Change*, *81*, 209–232. https://doi.org/10.1007/s10584-006-9218-z

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., Van Der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., . . . Vázquez-Baeza, Y. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wallace, J. M., and Gutzler, D. S. (1981). Teleconnections in the Geopotential Height Field during the Northern Hemisphere Winter. *Monthly Weather Review*, *109*(4), 784–812. https://doi.org/10.1175/1520-0493(1981)109<0784:TITGHF>2.0.CO;2

Wang, W., Bruyère, C., Duda, M., Dudhia, J., Gill, D., Kavulich, M., Keene, K., Chen, M., Lin, H.-C., Michalakes, J., Rizvi, S., Zhang, X., Berner, J., Ha, S., and Fossell, K. (2008). *WRF Version 3 Modeling System User's Guide*. National Center for Atmospheric Research. https://www2.mmm.ucar.edu/wrf/users/docs/user_guide_V3/user_guide_V3.9/ARWUsersGuideV3.9.pdf

Wang, Z., Zeng, X., and Decker, M. (2010). Improving snow processes in the Noah land model. *Journal of Geophysical Research: Atmospheres*, *115*, 2009JD013761. https://doi.org/10.1029/2009JD013761

Wicker, L. J., and Skamarock, W. C. (2002). Time-Splitting Methods for Elastic Models Using Forward Time Schemes. *Monthly Weather Review*, *130*(8), 2088–2097. https://doi.org/10.1175/1520-0493(2002)130<2088:TSMFEM>2.0.CO;2

Wilks, D. S. (2006). *Statistical methods in the atmospheric sciences* (2nd ed). Academic Press.

WPS Developers. (2017). *WRF Pre-Processing System (WPS) source code* (Version 3.9.1). Github. https://github.com/wrf-model/WPS/releases/tag/v3.9.1

WRF Developers. (2017). *Weather Research and Forecasting (WRF) model source code* (Version 3.9.1.1). Github. https://github.com/wrf-model/WRF/releases/tag/V3.9.1.1

WRF V3 Geographical Static Data Downloads Page [WRF users' page]. (n.d.). https://www2.mmm.ucar.edu/wrf/users/download/get_sources_wps_geog_V3.html

Yang, Y., Uddstrom, M., and Duncan, M. (2011). Effects of short spin-up periods on soil moisture simulation and the causes over New Zealand: LSM SPIN-UP ISSUES FOR SHORT RECORD DATA. *Journal of Geophysical Research: Atmospheres*, *116*, n/a–n/a. https://doi.org/10.1029/2011JD016121

Yeager, S. G., Danabasoglu, G., Rosenbloom, N. A., Strand, W., Bates, S. C., Meehl, G. A., Karspeck, A. R., Lindsay, K., Long, M. C., Teng, H., and Lovenduski, N. S. (2018). Predicting Near-Term Changes in the Earth System: A Large Ensemble of Initialized Decadal Prediction Simulations Using the Community Earth System Model. *Bulletin of the American Meteorological Society*, *99*(9), 1867–1886. https://doi.org/10.1175/BAMS-D-17-0098.1

Yeager, S. G., Karspeck, A., Danabasoglu, G., Tribbia, J., and Teng, H. (2012). A Decadal Prediction Case Study: Late Twentieth-Century North Atlantic Ocean Heat Content. *Journal of Climate*, *25*(15), 5173–5189. https://doi.org/10.1175/JCLI-D-11-00595.1

Yoo, A. B., Jette, M. A., and Grondona, M. (2003). SLURM: Simple Linux Utility for Resource Management [ISSN: 0302-9743, 1611-3349]. In *Lecture notes in computer science* (pp. 44–60). Springer Berlin Heidelberg. https://doi.org/10.1007/10968987_3

Yuan, N., Fu, Z., and Liu, S. (2013). Long-term memory in climate variability: A new look based on fractional integral techniques. *Journal of Geophysical Research: Atmospheres*, *118*(23). https://doi.org/10.1002/2013JD020776

Zender, C. S. (2008). Analysis of self-describing gridded geoscience data with netCDF Operators (NCO) [Publisher: Elsevier BV]. *Environmental Modelling & Software*, *23*(10), 1338–1342. https://doi.org/10.1016/j.envsoft.2008.03.004