JGR Atmospheres

RESEARCH ARTICLE
10.1029/2023JD039709

Key Points:

- Generally, the initial condition-based drift correction (ICDC) shows the best performance for all the domains, climate variables and indices analyzed
- The ICDC contributes to improve the predictive skill of the North Atlantic Oscillation index and reduces the loss of predictive skill observed in El Niño/Southern Oscillation indices
- Despite the loss of predictive skill after ensemble size reduction, ENS3 performs similar to ENS10 and ENS40 for sea surface temperature (SST) at lead years 2–9

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

J. J. Rosa-Cánovas,
jjrc@ugr.es

Citation:

Rosa-Cánovas, J. J., García-Valdecasas Ojeda, M., Romero-Jiménez, E., Yeste, P., Gámiz-Fortis, S. R., Castro-Díez, Y., & 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*, e2023JD039709. https://doi.org/10.1029/2023JD039709

Received 25 JUL 2023
Accepted 26 OCT 2023

Drift Correction and Sub-Ensemble Predictive Skill Evaluation of the Decadal Prediction Large Ensemble With Application to Regional Studies

J. J. Rosa-Cánovas[1,2], M. García-Valdecasas Ojeda[1,2], E. Romero-Jiménez[1], P. Yeste[1,2], S. R. Gámiz-Fortis[1,2], Y. Castro-Díez[1,2], and M. J. Esteban-Parra[1,2]

[1]Department of Applied Physics, University of Granada, Granada, Spain, [2]Andalusian Institute for Earth System Research (IISTA-CEAMA), Granada, Spain

Abstract A large ensemble of experiments is required to reveal the predictable climate signal masked by the background noise in the decadal climate prediction (DCP). This is one of the main obstacles which complicates the generation of high-resolution decadal climate information at regional scale, given the computing cost of the task. In this study, a set of representative sub-ensembles of three members (ENS3) from the Decadal Prediction Large Ensemble has been selected to produce dynamically downscaled DCPs in future studies, minimizing the amount of computing resources required to conduct the regionalization while reducing as much as possible the loss of predictive skill with respect to the full ensemble (ENS40). The procedure to follow comprises two steps: first, an analysis to choose the most appropriate method of drift correction to remove the model drift; second, the selection of three members to build ENS3 and the evaluation of its performance and the impact of ensemble size on sub-ensemble performance. The study has been focused on sea surface temperature (SST), near-surface temperature anomaly and sea level pressure over some Coordinated Regional Climate Downscaling Experiment regions: Europe, South America and North America. The initial condition-based approach has been shown to be the most suitable method in the three domains. Although there is an inevitable loss of predictive skill when reducing the ensemble size, ENS3 has shown to be a relatively good alternative to ENS10 and ENS40 when facing computing constraints and the analysis is focused on SST.

Plain Language Summary The decadal climate prediction (DCP) has been shown to be particularly useful in developing mitigation and adaptation strategies to the short-term climate change. When climate information at a very high resolution is needed over a certain region of interest, global model outputs are used to conduct regional simulations. Since the regionalization of a large ensemble of global decadal predictions requires a huge amount of computational resources, a set of representative sub-ensembles of three members from the Decadal Prediction Large Ensemble has been selected to produce regional DCPs in future studies. Prior to building these sub-ensembles, several drift correction methods have been evaluated by assessing their contribution to the predictive skill of some variables and climate indices.

1. Introduction

The decadal climate prediction (DCP) has progressively received more attention during the last years since climate information at this time scale has been shown to be particularly relevant in order to develop mitigation and adaptation strategies to the short-term climate change (Kushnir et al., 2019; Smith et al., 2018). DCP aims at filling the gap between seasonal-to-interannual forecasting and multidecadal-to-century climate projections. While the former relies on model initialization to properly forecast the natural climate variability, the latter are only externally forced (e.g., with information about anthropogenic changes in concentration of greenhouse gases and no observation-based initialization) to capture the long-term signal of climate change. In this context, the nature of DCP is twofold: it is not only an initial-condition problem, but also a boundary-condition one (Meehl et al., 2021). DCP represents a confluence point where the initialization stage may contribute to enhancing the magnitude of the predictable signal and reducing the uncertainty in climate change information (Meehl et al., 2009, 2014). The model components initialized in DCP are those which have longer climate memory, such as the ocean (and potentially sea ice cover and soil moisture content).

Very important advancements in climate modeling have been achieved over the last 30 years in terms of model complexity, accuracy and reliability (Randall et al., 2019). However, all dynamical models are in essence approximations that contain inherent biases (i.e., errors) which arise from different aspects, such as the selected type and order of numerical approximation, coarse spatial and temporal resolution, subgrid-scale parameterizations and uncertainties in boundary and forcing data (D. Chen et al., 2021). The way of addressing this issue in the framework of DCP has some particularities. There are two main strategies to initialize decadal experiments: the "full-field" and the "anomaly" initialization schemes. When a model is initialized following the "full-field" approach (Yeager et al., 2012, 2018), that is, using full observational data as initial conditions, the simulation always drifts away from the observed climate toward the state systematically preferred by that model given its configuration, that is, the model climatology. On the other hand, the "anomaly" approach (Matei et al., 2012) attempts to avoid this drift by only using observed anomalies added to the model climatology as initial inputs. Later, the model climatology is subtracted from the output to obtain the predicted anomalies. Notwithstanding, 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. No strategy consistently performs the best, so both are considered in the Decadal Climate Prediction Project (DCPP) contribution to CMIP6 (Boer et al., 2016). Regardless of the initialization method, an assessment of the model drift is firmly recommended by Boer et al. (2016) prior to carrying out any analysis in order to reduce the lead time-dependent bias in output data.

Although all drift correction procedures share the same fundamental principle, some differences exist among them. The simplest method considers removing the lead time-dependent drift in terms of the observed mean climate (Boer et al., 2016). A more comprehensive approach may be required when discrepancies between observed and modeled trends exist (Kharin et al., 2012). In some cases, including information about the observed climate at the initialization date might improve the drift detection given the importance of initial conditions in this type of experiments (Choudhury et al., 2017; Fučkar et al., 2014). More complex methods opt to consider non-monotonous lead time dependencies in the evolving drift (Gangstø et al., 2013; Kruschke et al., 2016) or make use of Principal Component Analysis to perform the correction (Paeth et al., 2019). In all cases, using a wide set of retrospective forecasts initialized every year is suggested to avoid sampling biases in the drift assessment (Choudhury et al., 2016) and to properly evaluate the skill of the product (Goddard et al., 2013). Choudhury et al. (2017) carried out an extensive study to evaluate the performance of some of those methodologies in different CMIP5 decadal data sets. They concluded that the choice of a specific approach to remove the drift can depend on multiple factors, such as the metric chosen for evaluation, the region or even the climate model.

The level of trustworthiness in DCP is linked to the amplitude of climate signal which can be predicted versus the amplitude of background noise related to both model biases and the chaotic nature of weather and climate. This relationship is known as the signal-to-noise ratio. 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 bias due to sensitivity of predictions to uncertainties in observational initial state (Meehl et al., 2014). A considerable amount of ensemble members is needed to reveal the predictable signal masked by the unpredictable noise (Smith et al., 2019). By using a conceptual model to analyze the effect of the ensemble size on 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 yearly initialized from 1960 onwards. When information at finer spatial resolution is requested over a specific domain, outputs from global models are regionalized by using downscaling techniques (Giorgi & Mearns, 1991; Kotamarthi et al., 2021). Dynamical downscaling (DD) has shown to be skilful at improving the representation of physical processes whose impact is more remarkable at local scale (e.g., S. Chen et al., 2019; García-Valdecasas Ojeda et al., 2017, 2020a, 2020b; Posada-Marín et al., 2019). In the context of DCP, DD has also been applied to produce and assess the skill of large ensembles of high-resolution decadal predictions over Europe (e.g., Ehmele et al., 2020; Feldmann et al., 2019; Reyers et al., 2019). The requirement of large ensembles makes the generation of operational decadal predictions at regional scale only affordable for those research groups or institutions which have substantial computing resources to carry out the simulations, leaving out most of the scientific community and significantly delaying advancements in this branch of the climate science.

Given these computing constraints and taking into account that a reduction of the ensemble size inevitably leads to the loss of predictive skill, the main aim of this study is to provide guidance for the selection of a 3-member sub-ensemble (ENS3) from the Decadal Prediction Large Ensemble (DPLE; Yeager et al., 2018) which is representative of the performance of the whole 40-member ensemble (ENS40) to some extent for several areas of interest. The selected sub-ensemble could be used to conduct DD simulations in future studies, minimizing the computing requirements to conduct the downscaling 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. DPLE has been chosen in this analysis since it is the only decadal prediction system which, to the best of our knowledge, publicly provide all the input fields required to run a regional model for decadal experiments initialized every year and several members of the ensemble (see https://www2.cesm.ucar.edu/projects/community-projects/DPLE/DPLE_output_fields/). Although the DPLE is composed of 40 members, only 10 are available as input for DD simulations (ENS10). DD simulations require a large set of variables from a global model with specific time aggregations (e.g., 6-hourly frequency) and at several height/soil levels to run the regional model. Since the storage requirements to save all these variables are highly expensive, global models generally save only a selection of output fields which exclude data with shorter time aggregation.

The procedure to follow in this research 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 sub-ensemble performance.

Three climate fields have been evaluated: sea surface temperature (SST), near-surface temperature anomaly (NSTA) and sea level pressure (SLP). Moreover, several climate indices computed with these fields have also been analyzed. Since the skill of correction methods could vary through the global domain, the study has been addressed by considering several regions of interest in the context of DD: the European (EUR), South-American (SA), and North-American (NA) domains from the Coordinated Regional Climate Downscaling Experiment (CORDEX; Giorgi et al., 2009).

This document is organized as follows. Section 2 contains an overview of the data sets considered in this study, the drift correction techniques used to remove model drift, the description of the climate indices analyzed and the predictive skill evaluation methodology. Section 3 presents the results obtained from drift correction, sub-ensemble selection, impact of ensemble size on performance and the discussion. Finally, the conclusions of this work are summarized in Section 4.

2. Data and Methodology

2.1. The CESM Decadal Prediction Large Ensemble

The DPLE (Yeager et al., 2018) is a set of global near-term climate simulations conducted by using the Community Earth System Model version 1.1 (CESM), developed by the National Center for Atmospheric Research (NCAR). The project encompasses a collection of 62 experiments, spanning 122 months, initialized every year in November from 1954 to 2015, for each ensemble member. The ensemble is composed of 40 members generated by randomly perturbing the initial atmospheric conditions. The model configuration in DPLE is the same as in the previous Large Ensemble Project (LE; Kay et al., 2015), primarily focused on historical and long-term climate simulations. The Community Atmosphere Model version 5 (CAM5; Hurrell et al., 2013), run at nominal 1° horizontal resolution and 30 vertical levels, is the atmosphere component. The ocean component is the Parallel Ocean Program, version 2 (POP2; Danabasoglu et al., 2012), run at nominal 1° horizontal resolution and 60 vertical levels. Los Alamos National Laboratory Community Ice Code, version 4 (CICE4; Hunke & Lipscomb, 2008) is the sea ice component and was run at the same resolution as the ocean component. Finally, the land component is the Community Land Model, version 4 (CLM4; Lawrence et al., 2011). The fundamental difference between CESM-LE and CESM-DPLE is the "full-field" initialization of the ocean and sea ice components. The initial states for both components were obtained from a forced ocean-sea ice (FOSI) simulation driven by atmospheric reanalysis data (Yeager et al., 2018). The radiative forcing used to drive the DPLE simulations is historical up to 2005 (Lamarque et al., 2010) and follows the representative concentration pathway 8.5 (RCP8.5; Meinshausen et al., 2011) from 2006 onwards. Further details on the experimental design and initialization methodology can be found in Kay et al. (2015) and Yeager et al. (2018).

DPLE data used in this study to address the drift correction comprise the whole ENS40 for the decadal experiments, initialized every year, starting from 1960 to 2009. However, as mentioned in Section 1, only the members of ENS10 provide all variables required with the most appropriate time frequency to conduct DD simulations. For this reason, only those 10 members have been considered to build ENS3 and assess the impact of ensemble size on the sub-ensemble performance. The data sets used in this study have been downloaded from the Earth System Grid Federation (ESGF) platform (Danabasoglu, 2019).

2.2. Reference Data Sets

The reference data used in this study can be divided into two categories: the data used in drift correction and the data used in skill evaluation. The ERA5 reanalysis (Hersbach et al., 2020, 2023) has been chosen as reference data set to carry out the drift correction. ERA5 is the latest state-of-the-art reanalysis produced by the European Center for Medium-Range Weather Forecasts (ECMWF). All fields are supplied at 0.25° × 0.25° grid resolution, although they have been interpolated on the DPLE grid (1° horizontal resolution) before applying the drift correction techniques. As stated in Section 1, the analysis has been focused on SST, NSTA, SLP and several climate indices. However, in the context of DD, the adjustment of the drift would be applied to all input fields in the simulations to keep the physical coherence between them (Paeth et al., 2019). Thus, ERA5 has been considered as the reference data set because it provides enough information to correct all input variables in DD simulations at multiple height and soil levels. To evaluate the SST predictive skill, the Extended Reconstructed SST version 5 (ERSSTv5; Huang et al., 2017a, 2017b) has been used. The Goddard Institute for Space Studies Surface Temperature Analysis version 4 (GISTEMP4; GISTEMP Team, 2023; Lenssen et al., 2019) has been chosen in the evaluation of NSTA over land. Since there are gaps with unavailable data in NSTA time series over some grid points, the locations considered in the analysis of NSTA are at least 70% completed to obtain the multi-year lead time series (see Section 2.5). For SLP, the near-real-time update of Hadley Center's monthly historical mean SLP (HadSLP2r; Allan & Ansell, 2006) has been selected. These data sets are provided with different spatial resolution. The DPLE adjusted data have been interpolated on the reference data set grids before predictive skill evaluation. While SST and NSTA have been interpolated on a 2° × 2° resolution grid, SLP has been interpolated on a grid of 5° × 5° resolution. Both drift correction and evaluation have been performed in the period 1960–2019.

2.3. Drift Correction Methods

The drift correction methods used in this study are described in this section. Although they have been applied to correct the drift only for SST, NSTA, and SLP variables, the method which performs the best is intended to be used to correct the drift in all input fields of DD simulations. A multivariate correction approach is expected to minimize the artificial drift in DD input data while maintaining the physical coherence between the fields involved (Paeth et al., 2019). All methods described in the following have 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.

2.3.1. Mean Drift Correction

The mean drift correction method (MDC; Boer et al., 2016) aims at removing the lead time-dependent model drift with respect to the reference mean fields. Let $Y_{kj\tau}$ be a decadal climate forecast, where k is the member of the ensemble, j stands for the initial date and τ denotes the lead time. Consider the ensemble mean $\{Y\}_{j\tau}$, calculated as follows:

$$\{Y\}_{j\tau} = \frac{1}{N_{ens}} \sum_{k=1}^{N_{ens}} Y_{kj\tau}, \tag{1}$$

where N_{ens} is the number of ensemble members. The lead time-dependent forecast climatology is calculated as

$$\{\overline{Y}\}_{\tau} = \frac{1}{N_d} \sum_{j=\delta_1}^{\delta_2} \{Y\}_{j\tau}, \tag{2}$$

where δ_1 is the initial date for the chronologically first decadal forecast of the set, δ_2 identifies the initial date for the last forecast and N_d is the number of initial dates between δ_1 and δ_2.

On the other hand, let $X_{j\tau}$ be the reference data used to correct the model drift. Note that this data set is a continuous time series which spans the whole period considered in the correction (from 1960 to 2019). Thus, the reference data set cannot be associated with specific initial dates j or lead times τ in the same way as the decadal forecasts. Here, $X_{j\tau}$ does not denotes the reference information at lead time τ in an experiment initialized on the date j, but the reference information in that continuous series at the time which corresponds to the initial date j and lead time τ in the ensemble mean $\{Y\}_{j\tau}$. Thus, the procedure to obtain the lead time-dependent reference climatology is

$$\overline{X}_{\tau} = \frac{1}{N_d} \sum_{j=\delta_1}^{\delta_2} X_{j\tau} \tag{3}$$

The model drift at lead time τ is given by

$$d_{\tau}^{MDC} = \left\{\overline{Y}\right\}_{\tau} - \overline{X}_{\tau} \tag{4}$$

Finally, the corrected forecast for the ensemble member k and initial date j at lead time τ is calculated by removing the drift in $Y_{kj\tau}$:

$$Y_{kj\tau}^{MDC} = Y_{kj\tau} - d_{\tau}^{MDC} \tag{5}$$

2.3.2. Trend-Based Drift Correction

A model which does not properly capture long-term climate trends, such as the global warming, could produce decadal hindcasts which drift away from the observations showing dependency on the initialization time. Since the MDC cannot be suitable for addressing this sort of bias, Kharin et al. (2012) proposed a new trend-based drift correction method (TrDC) which also considers climate trends in the assessment of the model drift. Following their approach, the ensemble mean $\{Y\}_{j\tau}$ in Equation 1 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:

$$\begin{aligned}\{Y\}_{j\tau} &= \alpha_{\tau}^{Y} + \beta_{\tau}^{Y} j + \epsilon_{j\tau}^{Y} \\ X_{j\tau} &= \alpha_{\tau}^{X} + \beta_{\tau}^{X} j + \epsilon_{j\tau}^{X}\end{aligned}$$

where $\alpha_{\tau}^{Y} = \left\{\overline{Y}\right\}_{\tau} - \overline{j}\beta_{\tau}^{Y}$ and $\beta_{\tau}^{Y} = \mathrm{Cov}\left[\{Y\}_{j\tau}, j\right]/\mathrm{Var}[j]$ denote the intercept and slope coefficients for the decadal forecast, respectively, while $\alpha_{\tau}^{X} = \overline{X}_{\tau} - \overline{j}\beta_{\tau}^{X}$ and $\beta_{\tau}^{X} = \mathrm{Cov}\left[X_{j\tau}, j\right]/\mathrm{Var}[j]$ identify such coefficients for reference data. The higher-order errors are represented by $\epsilon_{j\tau}^{Y}$ and $\epsilon_{j\tau}^{X}$. In this case, the model drift is given by

$$\begin{aligned}d_{j\tau}^{TrDC} &= \alpha_{\tau}^{Y} + \beta_{\tau}^{Y} j - \left(\alpha_{\tau}^{X} + \beta_{\tau}^{X} j\right) = \\ &= d_{\tau}^{MDC} + \left(\beta_{\tau}^{Y} - \beta_{\tau}^{X}\right)\left(j - \overline{j}\right)\end{aligned} \tag{6}$$

where d_{τ}^{MDC} is the model drift calculated in Equation 4. Note that if there are no differences between the slope coefficients, the trend-based drift is reduced to the mean drift previously calculated. The corrected forecast for the ensemble member k, initial date j and at lead time τ is calculated by removing the drift from $Y_{kj\tau}$:

$$Y_{kj\tau}^{TrDC} = Y_{kj\tau} - d_{j\tau}^{TrDC} \tag{7}$$

2.3.3. Initial Condition-Based Drift Correction

As TrDC, the initial condition-based drift correction (ICDC; Fučkar et al., 2014) attempts to correct long-term climate trends. Additionally, this method also attempts to take advantage of the relevance that an accurate representation of the climate state at the initialization stage exhibits 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 data set $\eta_j = X_{j,\tau=1}$ at the initial date j. Note that η_j does not necessarily correspond to the same observed 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 (see Section 2.2). The equations of the ensemble mean $\{Y\}_{j\tau}$ and reference data $X_{j\tau}$ can be written as follows:

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

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

$$\begin{aligned} d_{j\tau}^{ICDC} &= \tilde{\alpha}_{\tau}^{Y} + \tilde{\beta}_{\tau}^{Y}\eta_j - \left(\tilde{\alpha}_{\tau}^{X} + \tilde{\beta}_{\tau}^{X}\eta_j\right) = \\ &= d_{\tau}^{MDC} + \left(\tilde{\beta}_{\tau}^{Y} - \beta_{\tau}^{Y}\right)\left(\eta_j - \overline{\eta}\right) \end{aligned}$$

where d^{MDC} is the model drift calculated in the Equation 4. The corrected forecast for the ensemble member k, initial date j and at lead time τ is calculated by removing the drift from $Y_{kj\tau}$:

$$Y_{kj\tau}^{ICDC} = Y_{kj\tau} - d_{j\tau}^{ICDC} \tag{8}$$

2.4. Complements for the Conventional Methods

2.4.1. Polynomial Fitting

In addition to the conventional methods described above, there are other techniques which can be used to complement drift correction conducted through those methods. One of them is the polynomial fitting (FIT; Gangstø et al., 2013; Kruschke et al., 2016). FIT is expected 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

$$d_{j\tau}^{FIT} = a_{0,j\tau} + a_{1,j\tau}\tau + a_{2,j\tau}\tau^2 + a_{3,j\tau}\tau^3, \tag{9}$$

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:

$$\begin{aligned} d_{j\tau}^{FIT} &= (b_0 + b_1 t_{j\tau}) + (b_2 + b_3 t_{j\tau})\tau + \\ &+ (b_4 + b_5 t_{j\tau})\tau^2 + (b_6 + b_7 t_{j\tau})\tau^3 \end{aligned} \tag{10}$$

The coefficients in Equation 10 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 $Y_{kj\tau}$:

$$Y_{kj\tau}^{FIT} = Y_{kj\tau} - d_{j\tau}^{FIT} \tag{11}$$

2.4.2. K-Nearest Neighbors

The k-nearest neighbors approach (kNN) 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 composite 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 the m closest $\eta_{i\neq j}$ values to η_j. As in Choudhury et al. (2017), $m = 60\%$ has been chosen.
3. These m 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 $Y_{kj\tau}$:

$$Y_{kj\tau}^{kNN} = Y_{kj\tau} - d_{j\tau}^{kNN} \tag{12}$$

2.5. Skill Evaluation

Since the ultimate goal of the study is to build ENS3 with three drift-corrected DPLE members to carry out DD simulations, the outcomes from drift correction have been evaluated within some of the regional domains defined by the Coordinated Regional Climate Downscaling Experiment (CORDEX; Giorgi et al., 2009) which are considered of special interest for our purposes: the European (EUR), South American (SA), and North American (NA) domains. Several lead time windows spanning 1, 4, and 8 years have been considered in the analysis to evaluate the dependence of the predictive skill on the lead time, as suggested by Goddard et al. (2013).

2.5.1. 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 the predictive skill of SST, NSTA, SLP and climate indices for ENS40. DCPs are skilful inasmuch as they can reproduce not only the observed climate variability, but also the magnitude of that change. These features can be quantified by using two deterministic metrics (Jolliffe & Stephenson, 2012): the anomaly correlation coefficient (ACC) and the root squared mean error (RMSE), respectively. Considering the ensemble average of forecasts $\{Y\}_{j\tau}^{DC}$, drift-corrected by using one of the aforementioned procedures, and the reference data $X_{j\tau}$, the averages along a given lead time period are

$$\begin{aligned} \{\hat{Y}\}_{j}^{DC} &= \frac{1}{\theta_2 - \theta_1 + 1} \sum_{\tau=\theta_1}^{\theta_2} \{Y\}_{j\tau}^{DC} \\ \hat{X}_j &= \frac{1}{\theta_2 - \theta_1 + 1} \sum_{\tau=\theta_1}^{\theta_2} X_{j\tau} \end{aligned} \tag{13}$$

where θ_1 and θ_2 are the first and last time steps, respectively, of the lead time period. The RMSE is calculated as follows:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{j=\delta_1}^{\delta_2} \left(\{\hat{Y}\}_{j}^{DC} - \hat{X}_j \right)^2}{N_d}} \tag{14}$$

The ACC is obtained by using

$$\mathrm{ACC} = \frac{\sum_{j=\delta_1}^{\delta_2} \left(\{\hat{Y}\}_{j}^{DC} - \overline{\{\hat{Y}\}}^{DC} \right) \left(\hat{X}_j - \overline{\hat{X}} \right)}{\sqrt{\sum_{j=\delta_1}^{\delta_2} \left(\{\hat{Y}\}_{j}^{DC} - \overline{\{\hat{Y}\}}^{DC} \right)^2 \sum_{j=\delta_1}^{\delta_2} \left(\hat{X}_j - \overline{\hat{X}} \right)^2}} \tag{15}$$

After calculating both metrics for each grid point, the spatially weighted averages ⟨RMSE⟩ and ⟨ACC⟩ have been computed over the CORDEX domains. The best qualified methods have been selected to build ENS3 afterward. A non-parametric bootstrap approach has been applied to address the statistical significance of these results. In order to check if ACC outputs significantly positive (or negative) values at the 95% confidence level, bootstrapped distributions of 5,000 scores have been computed at each grid point by resampling with replacement forecasts $\{Y\}_{j\tau}^{DC}$ across initialization date and ensemble member dimensions, following instructions outlined by Goddard et al. (2013). The same procedure has been followed to assess the statistical significance of the ACC scores for the climate indices and the confidence intervals for the averaged scores ⟨RMSE⟩ and ⟨ACC⟩.

2.5.2. Climate Indices

In addition to the analysis of SST, NSTA, and SLP, ACC scores for several climate indices have been also used as a decision factor when assessing the performance of the drift correction methods. Climate indices allow us to include in the analysis the representation of large-scale patterns of climate variability which influence on local climate, enabling a broader assessment which is not only constrained to the CORDEX regions. Several El Niño/Southern Oscillation (ENSO) indices have been considered alongside the North Atlantic Oscillation (NAO) and the Atlantic Multidecadal 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 & Kirtman, 2016), although its effect is 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: Niño 1 + 2, Niño 3, Niño 3.4, Niño 4, and Trans-Niño indices (Trenberth & Stepaniak, 2001).

Table 1
Definition of Regions to Calculate the Climate Indices

Index	Region
Niño 1 + 2	0°S–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°S–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°N–60°N, 80°W–0°W, (North Atlantic) and 60°S–60°N, 180°E−180°W (global average)

On the other hand, NAO is one of the most important atmospheric circulation modes in the Northern Hemisphere, described by changes in geopotential height or SLP over the action centers located at 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 an 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 for boreal winter months (December, January, and February) by using SST, excepting the NAO index, for which SLP is used instead. The regions from which the indices have been calculated are shown in Table 1. Niño 1 + 2, Niño 3, Niño 3.4, and Niño 4 indices have been computed by calculating the anomalies of the SST averaged 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 & Stepaniak, 2001). To calculate the NAO index, the anomalies of the spatially-averaged SLP in the Iceland region has been subtracted to the anomalies of the spatial average in the Azores region, applying the definition used in other studies which assessed the predictive skill of different decadal prediction systems (Smith et al., 2019, 2020). Following the same approach of Trenberth and Shea (2006) and Smith et al. (2020), the AMV index have been calculated as the difference between the SST averaged in the North Atlantic region and the SST global average.

2.5.3. Evaluation of Single Members and Sub-Ensemble Predictive Skill

To select the members which have been used to build ENS3, the analysis has been focused on ⟨ACC⟩ calculated for SST at lead years 2–9 for each domain separately. This lead time period has been chosen to examine the performance at decadal time scale (when the trend component of the skill becomes more important) rather than accounting for the skill arising from multi-year variability. Moreover, the skill which arises from seasonal to annual variability is avoided by removing the first year from the assessed period. This assessment has been centered on SST because of the particularly relevant role the ocean component of the climate system plays in the predictive skill of DCPs. As stated in Section 2.1, only 10 out of the 40 members of 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”), the member showing the lowest skill (the “worst”) and a member with an intermediate behavior. 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 sub-ensemble of the whole ensemble can be constructed. By selecting members with heterogeneous skill levels, we expect to retain to some extent part of the spread of ENS40, or at least ENS10, since these members cover the whole range of possible single performances among the 10 members available for DD.

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

1. How much does ⟨ACC⟩ of SST depend on ensemble size?
2. Does confidence interval of the sub-ensemble ⟨ACC⟩ contain the result obtained for ENS10 (the maximum ensemble size attainable by dynamically-downscaled DPLE) and ENS40 (the maximum ensemble size attainable by DPLE)?

3. Is the spread of members in the sub-ensemble suitable to quantify the uncertainty in the sub-ensemble predictive skill?

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, ⟨RMSE⟩ and ⟨ACC⟩ have been calculated for ensemble sizes ranging from 3 to 10. These ensembles have been constructed by the same bootstrap approach described in Section 2.5.1, but considering only the 10 members available for DD. Second, the results have been compared against 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 predictive skills which can be potentially achieved by ENS10 and ENS40 are covered by the confidence intervals obtained for sub-ensembles with different sizes. Again, a bootstrap strategy has been applied for ensemble sizes ranging from 3 to 10, alongside our ENS3, to calculate ⟨RMSE⟩ and ⟨ACC⟩. For every sub-ensemble, the bootstrapping has been repeated 5,000 times in order to calculate the percentage of ENS10 and ENS40 skill score coverage that the confidence intervals get. This methodology was also used by Sienz et al. (2016) to explore the skill of small sub-ensembles, 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 sub-ensemble spread is appropriate to represent the range of possible individual predictions over time, that is, how reliable the sub-ensemble predictions are. Following Goddard et al. (2013), reliability of decadal predictions can be analyzed through the Continuous Ranked Probability Skill Score (CRPSS):

$$CRPSS = 1 - \frac{\sum_{j=\delta_1}^{\delta_2} CRPS_{H_j}}{\sum_{j=\delta_1}^{\delta_2} CRPS_{R_j}}, \tag{16}$$

where $CRPS_{H_j}$ and $CRPS_{R_j}$ are the Continuous Ranked Probability Score for hindcast and reference distributions. The CRPS can be interpreted as the mean square error in the probabilistic space and is given by:

$$\text{CRPS}\left(\hat{Y}_{kj}^{DC}, \hat{X}_j\right) = \int_{-\infty}^{\infty} \left[\mathcal{G}\left(\hat{Y}_{kj}^{DC}\right) - \mathcal{H}\left(\hat{X}_j\right)\right]^2 dy, \tag{17}$$

where $\mathcal{G}$ and $\mathcal{H}$ are the cumulative distribution functions of the hindcasts (as cumulative Gaussian distribution) and the observations (as the Heaviside function), respectively. For a $\mathcal{G}\left(\hat{Y}_{kj}^{DC}\right)$ as a Gaussian distribution of mean $\left\{\hat{Y}_j\right\}_{N_{ens}}^{DC}$ (the ensemble mean of an ensemble with size N_{ens}) and variance σ^2, Equation 17 turns into (Gneiting & Raftery, 2007):

$$\begin{aligned} &CRPS\left(N\left(\left\{\hat{Y}_j\right\}_{N_{ens}}^{DC}, \sigma^2\right), \hat{X}_j\right)\right) = \\ &= \sigma\left[\frac{1}{\sqrt{\pi}} - 2\varphi\left(\frac{\hat{X}_j - \left\{\hat{Y}_j\right\}_{N_{ens}}^{DC}}{\sigma}\right) - \left(\frac{\hat{X}_j - \left\{\hat{Y}_j\right\}_{N_{ens}}^{DC}}{\sigma}\right)\left(2\phi\left(\frac{\hat{X}_j - \left\{\hat{Y}_j\right\}_{N_{ens}}^{DC}}{\sigma}\right) - 1\right)\right] \end{aligned} \tag{18}$$

where φ and ϕ denote the Gaussian probability distribution function and cumulative distribution function, respectively. In Equation 16, the mean of the hindcast and reference distributions is $\left\{\hat{Y}_j\right\}_{N_{ens}}^{DC}$, while the variance of the hindcast distribution is given by Goddard et al. (2013):

$$\sigma_H^2 = \frac{1}{N_d} \sum_{j=\delta_1}^{\delta_2} \frac{1}{N_{ens} - 1} \sum_{k=1}^{N_{ens}} \left(\hat{Y}_{kj}^{DC} - \left\{\hat{Y}_j\right\}_{N_{ens}}^{DC}\right)^2 \tag{19}$$

and the variance of the reference distribution is:

$$\sigma_R^2 = \frac{\sum_{j=\delta_1}^{\delta_2} \left(\left\{\hat{Y}_j\right\}_{N_{ens}}^{DC} - \hat{X}_j\right)^2}{N_d - 2} \tag{20}$$

The optimal value in Equation 16 is CRPSS = 0. It is attained for $\sigma_H = \sigma_R$, when the hindcast and reference distributions are equal and, therefore, the ensemble spread is certainly adequate to quantify the uncertainty. The sub-ensembles of different sizes from 3 to 10 members have been constructed by the same bootstrap approach described above, applied only for initial dates, while the members of the sub-ensembles 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.

3. Results

3.1. Evaluation of the Drift Correction Methods

As detailed in Section 2.5.1, this Section is devoted to the analysis of the added value of the drift correction methods to the predictive skill of SST, NSTA, SLP and several climate indices for ENS40. The assessment of the method performances has been focused on three CORDEX regions: the EUR, the SA and the NA domains.

3.1.1. European Domain

The spatially averaged scores for SST over the EUR domain per drift correction method along lead time are depicted in Figure 1. Red lines within boxes denote averages, boxes identify the 50% confidence interval and whiskers correspond to the 95% confidence interval. All methods have performed very well in terms of ⟨RMSE⟩ and none substantially have achieved better results than the others. The highest ⟨RMSE⟩ values have been found during the first year. ⟨RMSE⟩ for uncorrected data (RAW) is around 0.76 K, whereas values below 0.45 K are depicted when a drift correction method is applied. At lead years 2–5 and 6–9, a decrement in ⟨RMSE⟩ can be observed, while the lowest values are shown at decadal scale. In the results for ⟨ACC⟩, ICDC-like methods have outperformed the rest at lead year 1 and have shown the narrower confidence intervals. No differences are found among the ICDC-like methods. At this time scale, the initialization fingerprint is more prominent than afterward, so techniques which incorporate information about initial conditions in drift correction are candidates to perform the best. On the other hand, only $\mathrm{MDC}_{\mathrm{kNN}}$ and $\mathrm{TrDC}_{\mathrm{kNN}}$ give slightly better results than RAW, apart from ICDC-like methods. Although MDC is not expected to improve RAW performance in terms of ⟨ACC⟩, it is so for TrDC. Nevertheless, the introduction of the trend component leads to slightly poorer scores with respect to RAW and MDC. According to Equation 6, 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. Since MDC neither improves nor worsens RAW ⟨ACC⟩, 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 data set in drift correction) and ERSSTv5 (reference data set in skill evaluation) SST trends. At multi-year scale, ICDC-like methods keep performing the best although less differences are found among other methods. ICDC-like methods give ⟨ACC⟩ scores around 0.9. At lead years 2–9, the situation is similar with ⟨ACC⟩ scores near 0.95 for ICDC-like methods. The results for NSTA are depicted in Figure S1 in Supporting Information S1. Correction methods generally do not provide an added value to prediction skill, except for lead year 1. For the rest of lead times, skill scores between the methods and RAW are very similar. Since the field corrected is an anomaly, the mean field along the analyzed period in reference data set 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, ⟨ACC⟩ scores for RAW are very high at multi-year and decadal scales, so the added value of correction methods is minimal for this variable. In the analysis of SLP, whose results are depicted in Figure S2 in Supporting Information S1, ICDC-like methods continue giving the best results, although ⟨ACC⟩ scores are lower than for SST and NSTA. In addition, the highest ⟨ACC⟩ scores are found at lead year 1 for ICDC-like methods and they decrease along lead time. Since the climate change trend is stronger for temperature fields than for SLP, these outcomes were expected.

In Europe, it is also interesting to evaluate how well these techniques perform in the prediction of the NAO and AMV indices. ⟨ACC⟩ scores for each drift correction method and several lead times (multi-year and decadal) are depicted in Figure 2. In terms of NAO, results for ICDC-like methods are very promising at all multi-year lead times and also highly optimistic when comparing with other studies at lead years 2–9. While ⟨ACC⟩ of 0.85 are found for ICDC and $\mathrm{ICDC}_{\mathrm{kNN}}$, Smith et al. (2019) found an ⟨ACC⟩ of 0.49, using a multi-model ensemble composed by 71 members, and an ⟨ACC⟩ of 0.79 was found by Smith et al. (2020) for a higher ensemble size of 169 members after post-processing the NAO time series to rescale the signal. In this study, the predictive skill is higher at the beginning of the decade and decrease with lead time, as happened for SLP. The contrast between the high ACC scores for NAO and the low ⟨ACC⟩ scores for SLP is due to the fact that ⟨ACC⟩ for SLP has been calculated with annual averages, while ACC for NAO has been calculated in boreal winter months, when results for SLP are much more optimistic (Figure S3 in Supporting Information S1). TrDC-like methods only show

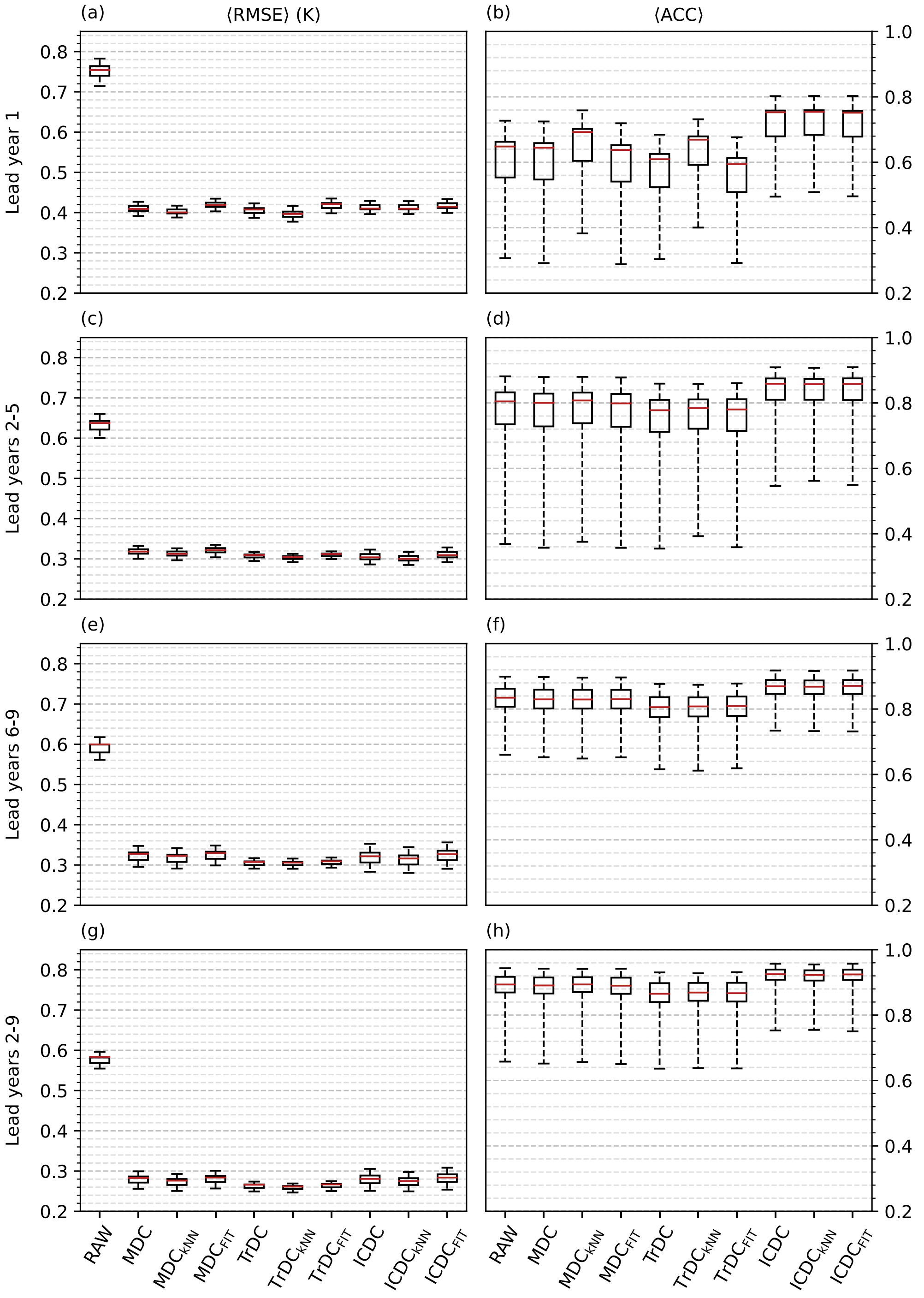

Figure 1. Spatially averaged root squared mean error in (a), (c), (e), and (g), and anomaly correlation coefficient in (b), (d), (f), and (h), for sea surface temperature along lead time over the EUR domain. Results are represented for each drift correction method. Red lines within boxes denote averages, boxes identify the 50% confidence interval of the spatial sample average and whiskers enclose the 95% confidence interval.

statistically-significant results in the earliest years at multi-year scale and at decadal scale too. ⟨ACC⟩ scores are too low for MDC-like methods and barely show statistical significance. Much higher correlations have been obtained in the analysis of AMV, as expected given that this index depends entirely on SST. Although generally ⟨ACC⟩ is above 0.70 almost for all methods and lead times, the most promising scores have been found for ICDC-like methods, with values above 0.90 for all lead times from 2 to 5 years ongoing and at decadal scale.

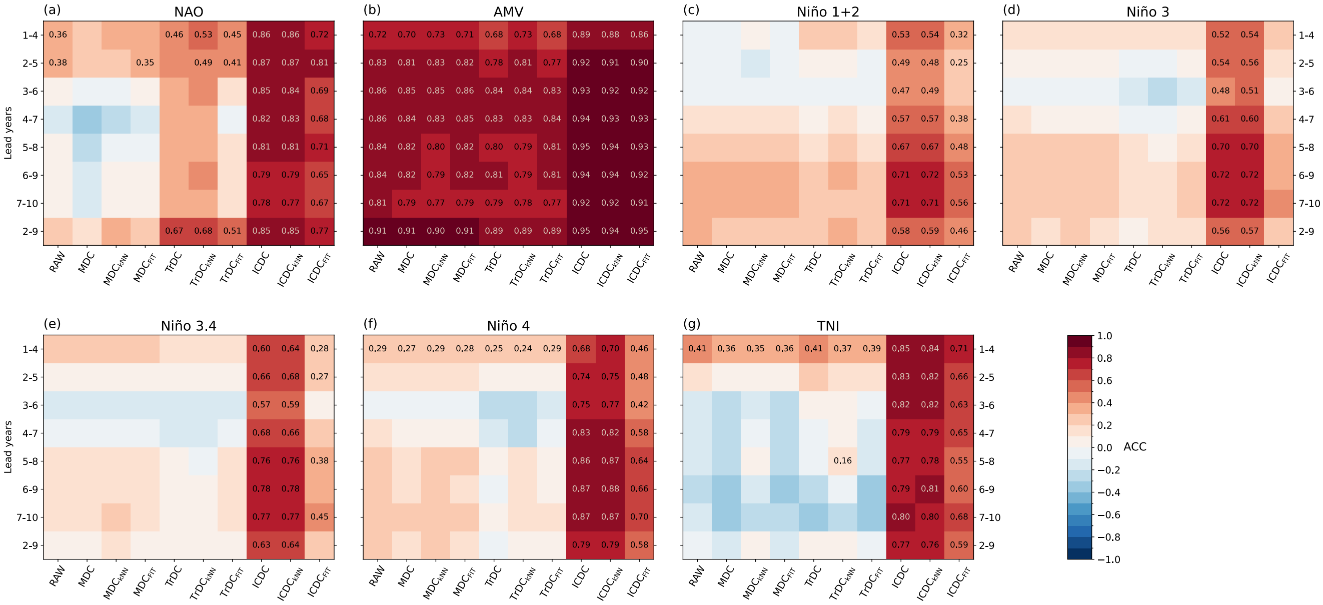

Figure 2. Anomaly correlation coefficient (ACC) computed for climate indices along lead time for each drift correction method. Numerical ACC values are shown when correlation is statistically-significant at the 95% confidence level.

The dominance of ICDC-like methods leads to considering them as the most appropriate candidates for the EUR domain. Among them, ICDC and ICDC$_{kNN}$ give the best results for NAO, although the three techniques perform similarly in the analysis of SST, NSTA, and SLP. Bearing in mind that the application of the conventional ICDC requires less effort and computing resources, ICDC may be the preferred choice in this area.

3.1.2. South American Domain

Results for averaged skill scores over the SA domain for SST are depicted in Figure 3. As for the EUR domain, drift correction improves ⟨RMSE⟩ with respect to RAW regardless of the correction method. At lead year 1, maximum improvements are nearly 0.20 K, but techniques using polynomial fitting give higher ⟨RMSE⟩. Differences among methods are less evident at the other lead times, when reductions in ⟨RMSE⟩ from 0.80 to 0.30 K are found with respect to RAW, approximately. Relevant disparities among method performances are depicted for ⟨ACC⟩ scores. As expected, ICDC-like procedures contribute to better capturing the climate variability at lead year 1 with outcomes about 0.65. 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. At lead years 2–5 and 6–9, ICDC-like methods continue performing better than the others, with scores around 0.54 and 0.52, respectively, followed by MDC-like and TrDC-like methods, but the correlations decrease with respect to lead year 1. Additionally, ⟨ACC⟩ for TrDC-like methods is lower than for RAW, as happened for the EUR domain. At decadal scale the situation is similar, with ⟨ACC⟩ scores higher than at multi-year scale, but with lower scores than for the first year. In the analysis of the predictive skill for NSTA (Figure S4 in Supporting Information S1), all methods give outcomes similar to RAW in ⟨RMSE⟩, as happened for the EUR domain. Maximum values around 0.5 K at lead year 1 and minimum around 0.22 K at lead years 2–9 are observed. The performance in terms of ⟨ACC⟩ is also very similar for all lead times, excepting for lead year 1, when ICDC-like methods get the highest correlations around 0.6. At the other lead times, correlations are higher but differences between methods are lower. Nevertheless, ICDC-like methods continue performing the best. After examining the results for SLP (Figure S5 in Supporting Information S1), similar conclusions have been obtained, although ICDC-like methods perform slightly better also at multi-year and decadal scale in terms of ⟨ACC⟩. As happened in the EUR domain, again ⟨ACC⟩ scores are much lower for SLP than for temperature variables.

The skill to capture the variability of some of the most relevant ENSO indices is also assessed in terms of the drift correction method and lead time (Figure 2). In general, the best performance is given by ICDC and ICDC$_{kNN}$ methods since very high ACC scores are obtained when they are used to remove the drift. Indeed, these methods are the only which consistently show statistically-significant positive ACC scores for the Niño 1, 2, 3, 3.4, 4, and

10.1029/2023JD039709

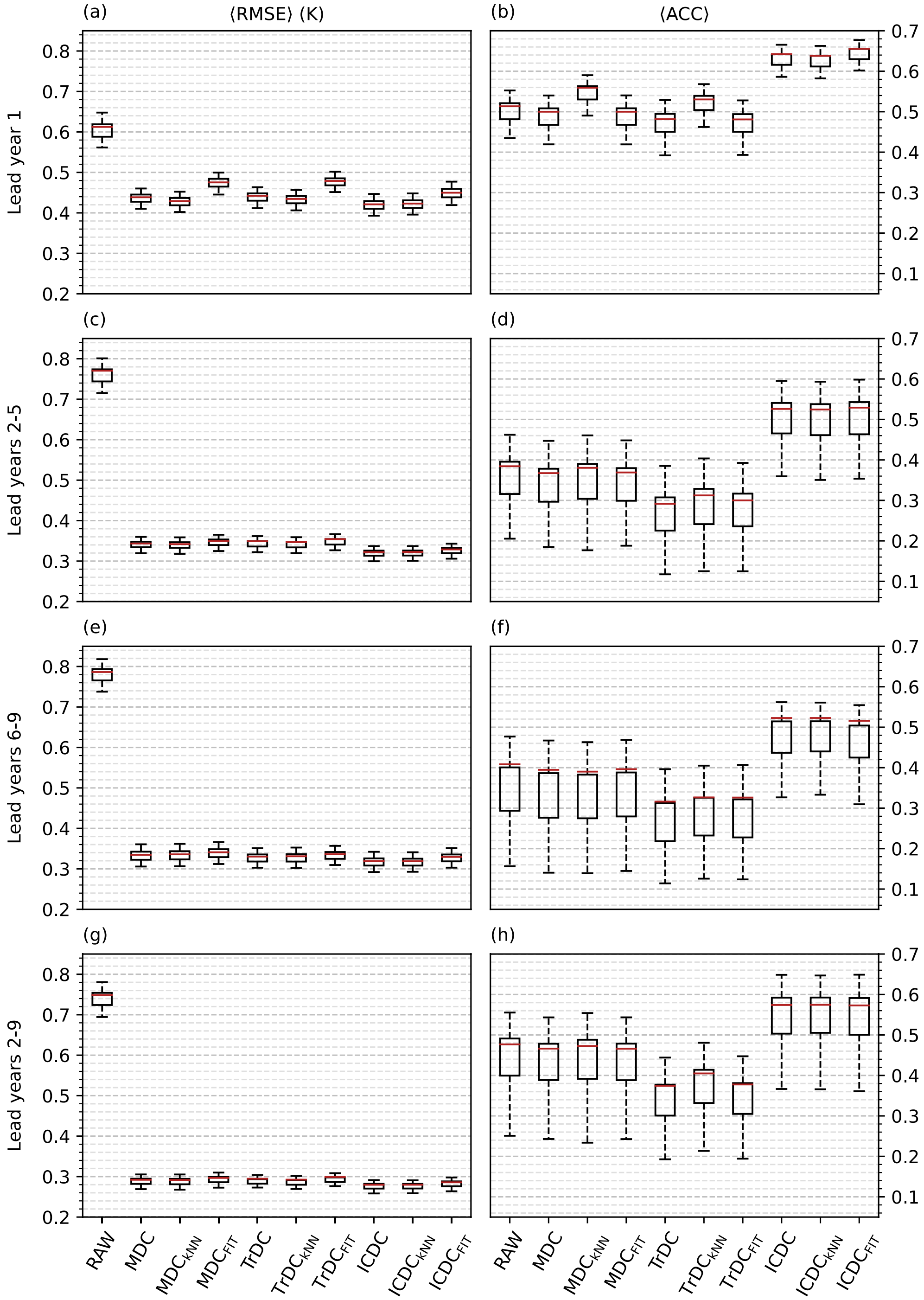

Figure 3. As Figure 1 but for the SA domain.

TNI indices. For instance, correlations between 0.57 and 0.78 have been found for Niño 3.4 index with the highest values at the last lead time windows of the decade, when the externally forced component become more important in the simulation. Even more optimistic results have been found for Niño 4 index and TNI, where values above 0.8 are found at several lead times throughout the decade. Niño 4 correlations are higher from lead years 4–7 onwards, while TNI results are better at the beginning of the decade. In both cases, ACC is above 0.7 for almost all lead times. These results are very promising in favor of ICDC and ICDC$_{kNN}$, since they prevent the severe loss of the predictive skill of ENSO along lead time highlighted by previous studies (Choi et al., 2016; Gonzalez & Goddard, 2016).

Given the performance of ICDC and ICDC$_{\mathrm{kNN}}$, and since there are no large differences between them, ICDC has been chosen to carry out the drift correction over the SA domain.

3.1.3. North American Domain

Spatially averaged scores for SST in NA domain are depicted in Figure 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 no big differences among them. ⟨RMSE⟩ for corrected data is higher at lead year 1 with values around 0.45 K, while the lowest values are found at lead years 2–9, ranging from 0.30 to 0.35 K, approximately. Again, the analysis of the ⟨ACC⟩ scores is necessary to choose the most appropriate correction procedure for this region. At lead year 1, ICDC-like methods outperform the rest, with values near 0.70. There are no robust differences between MDC-like and TrDC-like methods, although kNN gives slightly better results than the conventional and polynomial approaches. At lead years 2–5, the gap between ICDC-like and the other methods is larger and averages are about 0.72 for ICDC-like techniques, whereas the others do not improve RAW performance. At lead years 6–9, the situation is the same but with slightly higher ⟨ACC⟩ scores. The improvement in predictive skill of ICDC-like over the other methods continues existing at lead years 2–9, but differences slightly decrease with respect to lead years 6–9. While ⟨ACC⟩ values are nearly 0.76 for ICDC-like methods, they are below 0.6 for the rest. In the analysis of NSTA (Figure S6 in Supporting Information S1), results are similar to those for the EUR and SA domains. The methods generally give the same results for ⟨RMSE⟩, although ICDC-like techniques are lightly better. While the highest errors have been found at lead year 1, with the lowest value around 1.1 K for ICDC-like methods, the lowest is around 0.35 K at decadal scale. In the context of ⟨ACC⟩ ICDC-like methods outperform again the others at lead year 1 with correlations near 0.6, but the differences among them are smaller at the other lead times. The highest ⟨ACC⟩ values are shown at lead years 2–9 for ICDC-like methods and are about 0.84. In terms of SLP (Figure S7 in Supporting Information S1), ICDC-like methods also give the best results for ⟨ACC⟩, where the largest scores are observed at lead year 1 around 0.46. 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.

Considering the performances of ICDC and ICDC$_{\mathrm{kNN}}$ in the analysis of the climate indices (Figure 2) described in the previous sections and taking into account that there is no distinction among performances of both procedures, ICDC has been selected as the most suitable method to remove the drift over this domain.

RMSE and ACC maps for SST, NSTA, and SLP at lead years 2–9, with ICDC used as the drift correction method, are shown in Figure S8 in Supporting Information S1. In terms of ACC, the results obtained for the three domains considered in the analysis are maintained at global scale. The highest scores are observed for NSTA, with statistically significant positive values in the whole domain. As observed in Figure S2, S5, and S7 in Supporting Information S1, SLP is the field which get the lowest scores, especially in the south of the global domain, where significant negative correlations are observed. Results for RMSE show low values for the majority of the global domain, although higher values can be found on specific locations depending on the fields analyzed.

3.2. Evaluation of Single Members and Sub-Ensemble Predictive Skill

3.2.1. Selection of Single Members to Build ENS3

The predictive skill of the individual members usually varies depending on the field and domain, lead time and even 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, we have focused on ⟨ACC⟩ for SST at lead years 2–9 to choose among the single members. Figure 5 shows ⟨ACC⟩ scores for SST at lead years 2–9 over the EUR, SA and NA domains for ENS40, ENS3 and the 10 single members available for DD, with ICDC used as the drift correction method. As stated in Section 2.5.3, the member with the highest predictive skill ("best"), the member with the lowest skill ("worst") and another member with an intermediate behavior ("intermediate") have been chosen to build ENS3 for each domain.

In the EUR domain, ENS3 encompasses member 2 ("best"), member 8 ("worst") and member 7 ("intermediate"). ENS40, ENS3 and single members obtain correlations above 0.85. The highest correlations are depicted for ENS40 and ENS3, with values around 0.92 and 0.91, respectively. Only a difference of nearly 0.01 is observed between ENS3 and member 2 (the "best"). With respect to the SA domain, member 7 has been chosen as the

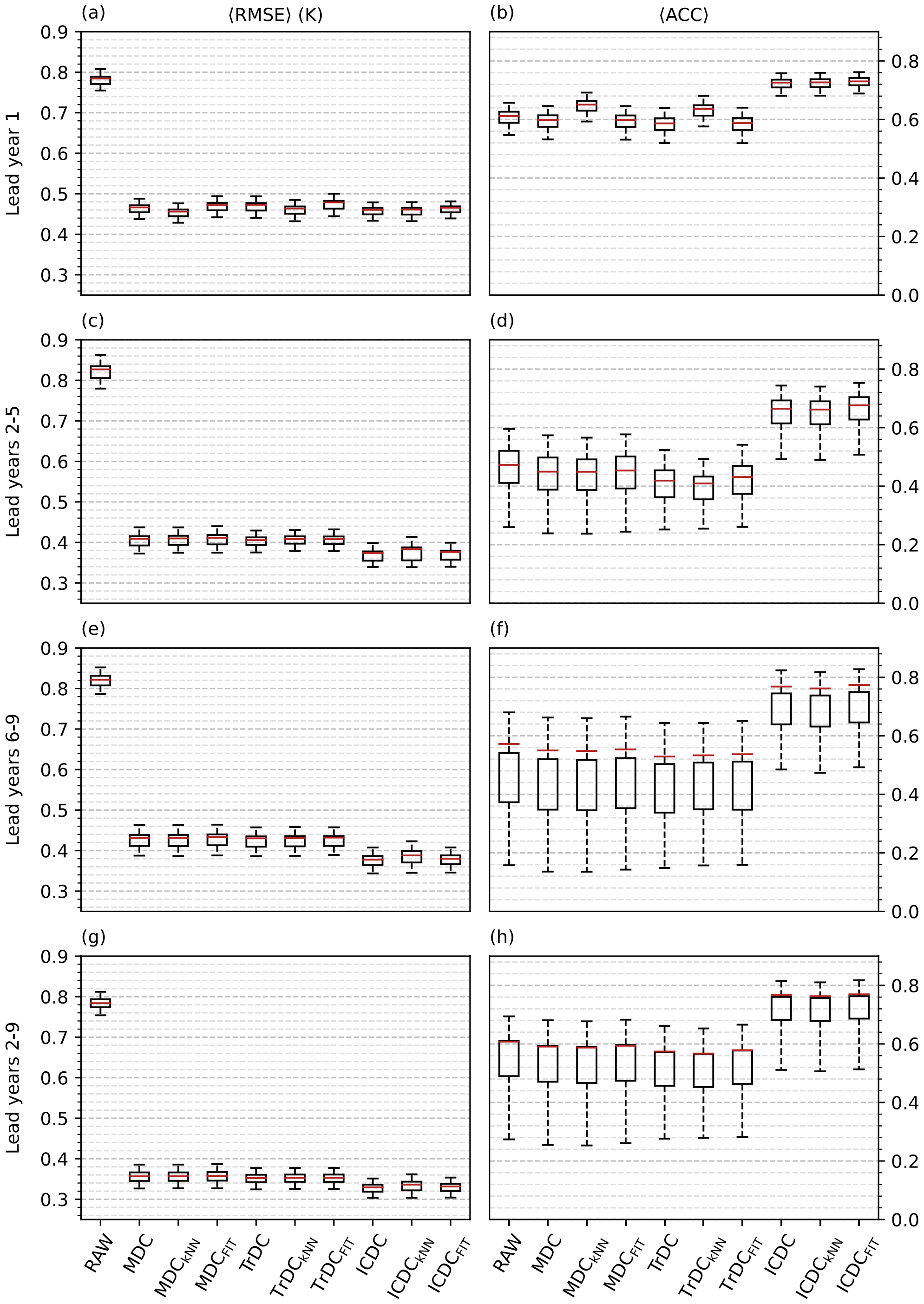

Figure 4. As Figure 1 but for the NA domain.

"best" member, member 2 as the "worst" and member 3 represents an intermediate behavior in the building of ENS3. The differences in terms of predictive skill are larger than for the EUR domain, with averaged correlations ranging from about 0.42 to 0.54, approximately. Member 7 performs similar as ENS3, getting an ⟨ACC⟩ around 0.54, whereas ENS40 shows a value close to 0.58. Finally, member 4 has been chosen as the "best," member 3 as the "worst" and member 10 as a member with intermediate performance in the NA domain. Likewise in the SA domain, slightly larger differences between single members are observed with respect to the EUR domain, with averages scores ranging from 0.61 to 0.74, approximately. In this case, all members are outperformed by

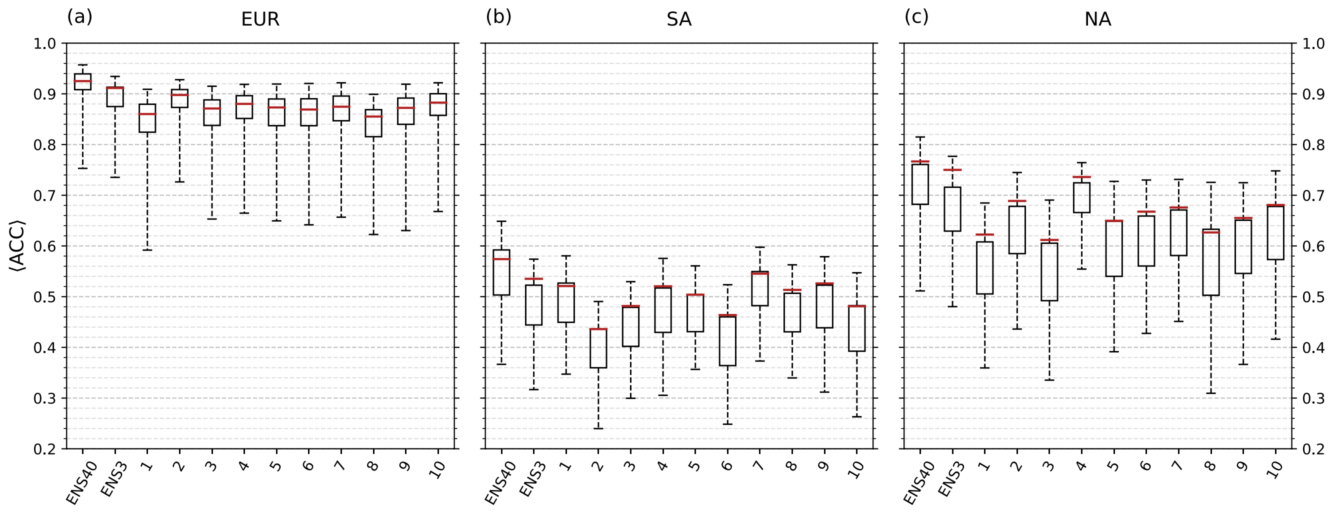

Figure 5. Spatially averaged anomaly correlation coefficient for sea surface temperature in the EUR (a), SA (b), and NA (c) domains at lead years 2–9 with ICDC used as the drift correction method. Results are depicted for ENS40, ENS3, and each single member. Red lines within boxes denote averages, boxes identify the 50% confidence interval of the spatial sample average and whiskers enclose the 95% confidence interval.

ENS3, which shows an ⟨ACC⟩ around 0.75. A gap of near 0.02 is observed between ENS40 and ENS3 averaged correlations, with ENS40 getting the highest predictive skill.

3.2.2. Dependence of SST Predictive Skill on Ensemble Size

The reduction of the ensemble size, as mentioned in Section 1, leads to an inevitably loss of predictive skill with respect to the full ensemble. However, the computing constraints and the huge amount of computing resources needed to dynamically downscale a large ensemble of decadal predictions complicate this task. Thus, some concessions may be made in relation to the number of members considered in hypothetical future DD simulations. In this section, the impact of ensemble size on SST predictive skill has been addressed. Figure 6 depicts how ⟨RMSE⟩ and ⟨ACC⟩ for SST varies with ensemble sizes from 3 to 10 members over the EUR, SA and NA domains at lead years 2–9. Ensemble size 3 with the symbol "*" represents the ENS3 selected in Section 3.2.1, while ensemble sizes without the symbols represent sub-ensembles with members randomly selected (see Section 2.5.3). Results for ENS40, as the potential predictive skill which can be achieved by DPLE, have also been included in the plots for comparison purposes, although only 10 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:

$$N_{sim} = N_y \times N_d \times N_{ens} = 10\ \text{years}/\text{date} \times 50\ \text{dates} \times N_{ens} = 500\ \text{years} \times N_{ens}$$

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.

In general, there is an improvement in prediction skill with increasing ensemble sizes, as expected. There is not any stabilization in the median of the skill scores for an ensemble size of up to 10 members. The median scores in the ensemble size of 40 members, the maximum attainable for DPLE, get always the best results. However, these differences in ⟨RMSE⟩ and ⟨ACC⟩ between different ensemble sizes are very low, and the width of confidence intervals is not reduced 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 focusing on the ensemble averaged scores for ENS3 and ENS10, the differences are even lower. There is a large contrast between the gain in predictive skill by increasing the ensemble size and the increase of the number of years which have to be simulated. For example, while the added value to ⟨ACC⟩ of using ENS10 instead of ENS3 if about 0.02, 0.04, and 0.02 for the EUR, SA and NA domains, respectively, 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 in predictive skill. Similar outcomes were found by Sienz et al. (2016) in correlations for North-Atlantic SST at lead years 2–5. In the cited study, the added value of increasing the ensemble

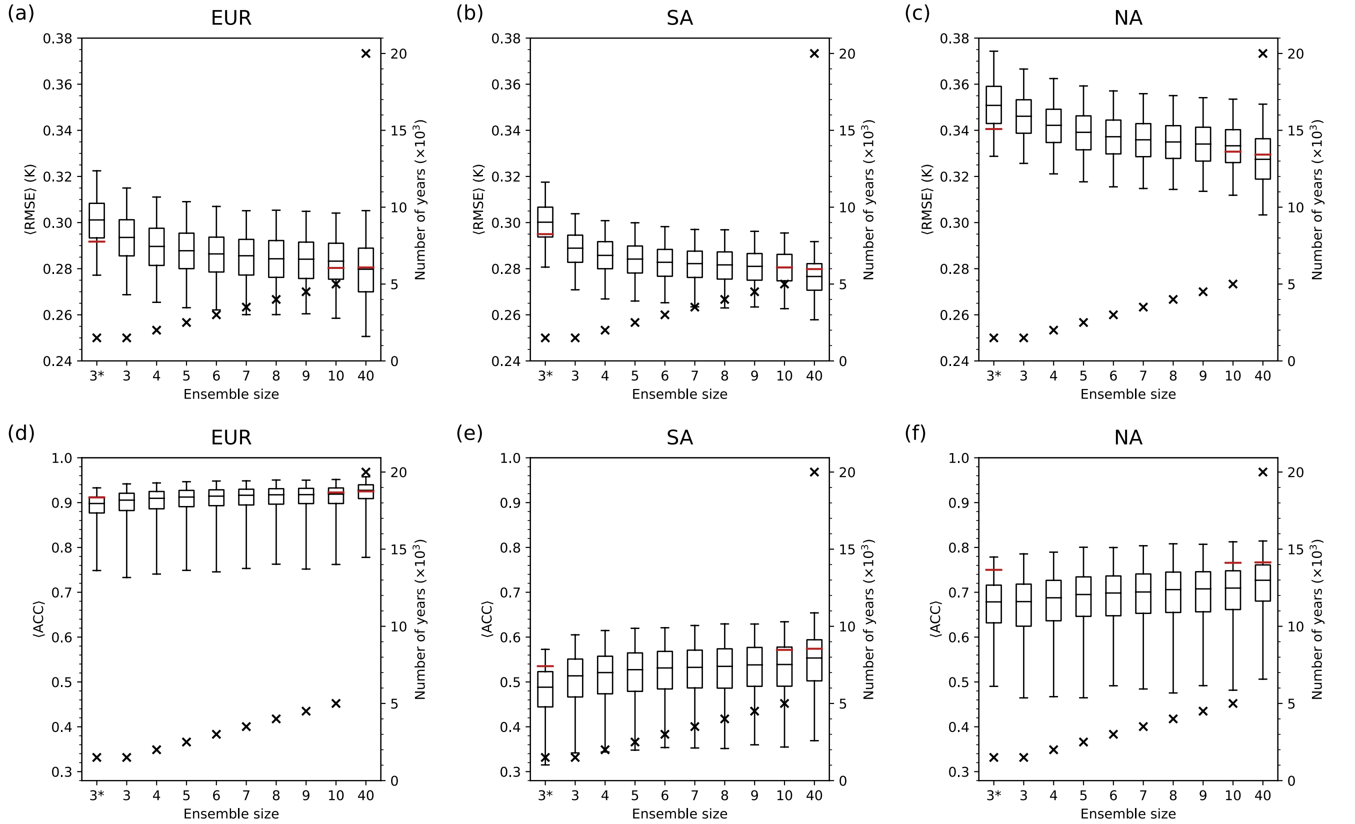

Figure 6. On the left axis, dependence of ⟨RMSE⟩ (a–c) and ⟨ACC⟩ (d–f) for sea surface temperature on ensemble size over the EUR, SA, and NA domains is represented by box plots. ICDC has been used for drift correction. Box plots show the results of a bootstrapping (see Section 2.5.3) where black lines denote the median, and boxes and whiskers the confidence intervals at the 50% and 95% levels, respectively. Red lines show the averages for ENS3, ENS10, and ENS40. On the right axis, number of years to be simulated per ensemble size is represented by crosses. Ensemble size three with the symbol "*" represents the ENS3 selected in Section 3.2.1.

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 prediction skill on ensemble size. They found an 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 at lead years 1–5. Added value with respect to uninitialized simulations (with ensemble size of 10 members) in correlation was observed from a ensemble size of 3 members. However, the improvement in correlation for precipitation by increasing the ensemble size were much larger, with differences of almost 0.4 between 10-member and 3-member ensembles. In addition, an added value with respect to uninitialized simulations was observed in precipitation only for ensemble sizes equal to 7 members and above.

The performance of ENS3 with respect to a 3-member sub-ensemble randomly selected can also be analyzed by examining Figure 6. The median values for ENS3 are above and below those for a random 3-member sub-ensemble 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 sub-ensemble). However, the sub-ensemble averaged scores for ENS3 outperforms almost always the median scores of the random sub-ensemble. In other words, this simple member selection carried out to build ENS3 performs better than at least the 50% of the random 3-member sub-ensembles generated for all domains, except for ⟨RMSE⟩ in the SA domain.

3.2.3. Dependence of Skill Score Coverage on Ensemble Size

The bootstrapping conducted in Section 3.2.2 has been repeated 5,000 times in order to get the coverage percentage of ENS10 and ENS40 skill scores by confidence intervals of sub-ensembles with different sizes. As can be observed in Figure 7, all confidence intervals always cover ENS10 and ENS40 skill scores, excepting the ENS3 in the SA domain and, with very high coverage percentages, the NA domain. This may mainly be due to the fact

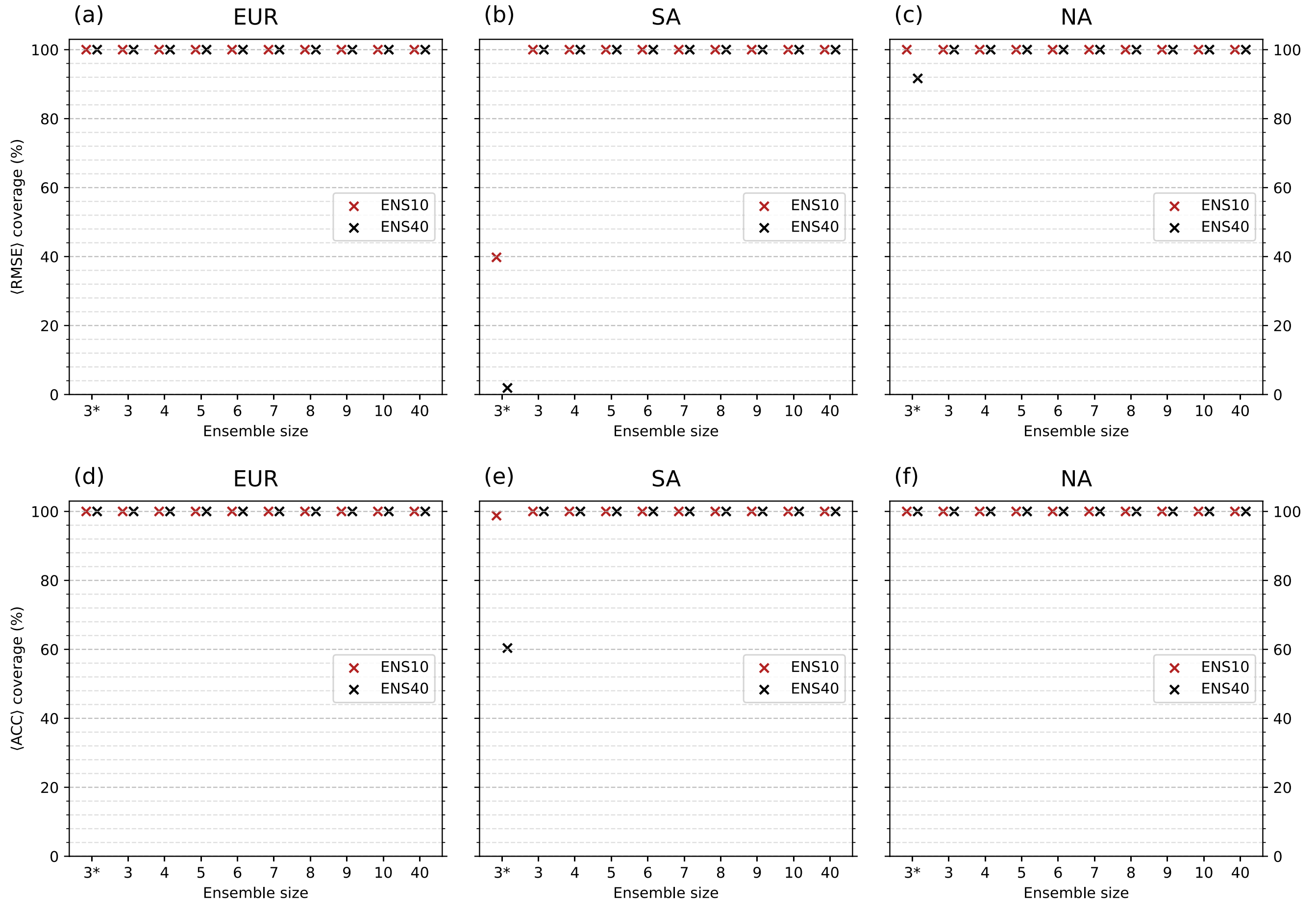

Figure 7. Percentage of skill score coverage by confidence intervals for different ensemble sizes. Skill scores have been calculated for sea surface temperature with initial condition-based drift correction used as the drift correction method. Red crosses correspond to the coverage of ENS10 skill scores, whereas black crosses denotes the results for ENS40. Ensemble size 3 with the symbol "*" represents the ENS3 selected in Section 3.2.1 More information about how the coverage has been implemented in Section 2.5.3.

that the bootstrapping for ENS3 only considers 3 members, as opposed to the 10 members included in the bootstrapping of the randomly selected sub-ensembles. In consequence, there is a slight offset of the ENS3 median and confidence interval limits with respect to those for the other sub-ensembles, as mentioned in Section 3.2.2. When the average scores for ENS10 and ENS40 are very close to the upper boundary of the ENS3 confidence intervals in Figure 6, the probability of that these averages are not covered by the ENS3 confidence intervals in a bootstrap iteration is very large. For the EUR and NA domains, ENS3 confidence intervals get a 100% coverage of ENS10 skill scores, while they get only a 40% for ⟨RMSE⟩ and around 98% for ⟨ACC⟩ in the SA domain. In the case of the coverage of ENS40 scores, results for the SA domain are even more pessimistic. ENS3 confidence intervals show only a 2% coverage of ⟨RMSE⟩ and around a 60% coverage of ⟨ACC⟩. The results for the coverage of ENS40 scores in the NA domain are more promising, since ⟨RMSE⟩ coverage is around 92% and ⟨ACC⟩ coverage is 100%.

3.2.4. Dependence of CRPSS on Ensemble Size

Results for the spatially-averaged CRPSS of SST at lead years 2–9 in the EUR, SA, and NA domains are depicted in Figure 8. The best results are found for the EUR domain, where the closest median values to zero are shown, ranging from −0.092 to −0.088. On the other hand, the most pessimistic values are found for the SA domain, where the median values are enclosed in the interval from −0.12 to −0.116. There is not a pronounced dependence of ⟨CRPSS⟩ on ensemble size in any domain. The ⟨CRPSS⟩ of ENS3 is always above the median of the randomly selected 3-member sub-ensemble. Its value in the EUR domain, about −0.086, is comparable to those of ENS10 and ENS40 in the EUR domain. In the SA domain, the ⟨CRPSS⟩ of ENS3 is lower than that of ENS10

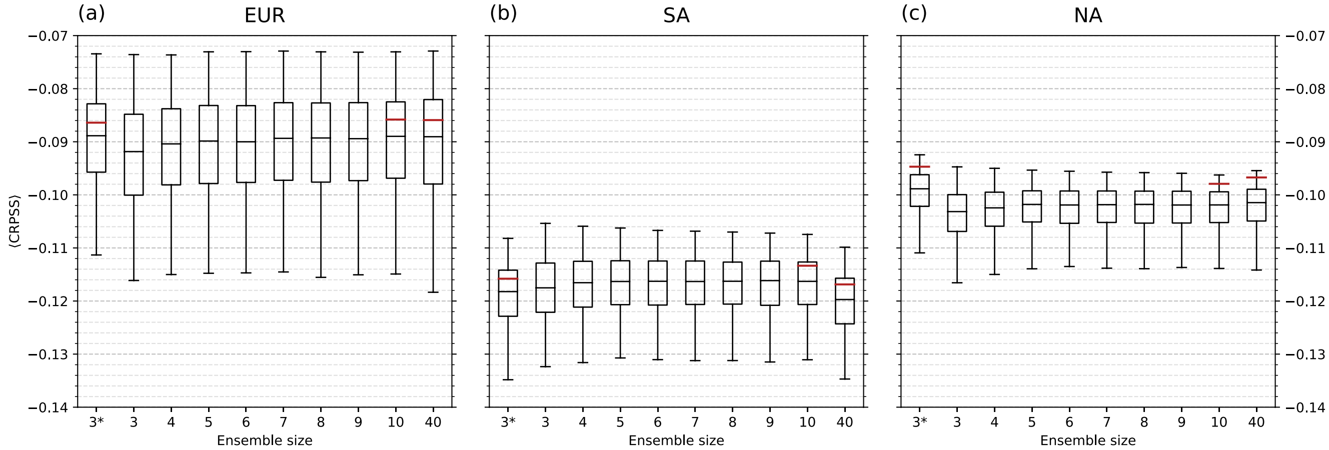

Figure 8. ⟨CRPSS⟩ of sea surface temperature at lead years 2–9 for different ensemble sizes in the EUR (a), the SA (b), and NA (c) domains. ICDC has been used for drift correction. Box plots show the results of a bootstrapping where black lines denote the median, and boxes and whiskers the confidence intervals at the 50% and 95% levels, respectively. Red lines show the averages for ENS3, ENS10, and ENS40. Ensemble size 3 with the symbol "*" represents the ENS3 selected in Section 3.2.1.

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 2.5.3, the desired value for CRPSS is 0. This would be attained for $\sigma_H = \sigma_R$, from Equations 19 and 20. If the standard error of the ensemble mean with respect to the observations over time is equal to the average spread of the ensemble members, 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 do not satisfy that desired condition. Therefore, neither the sub-ensemble nor full ensemble spreads are good representations of the hindcast uncertainty on average for SST over any domain.

4. Conclusions

The DCP is one of the branches of climate science in which the research community has been investing many efforts during the last years. The DCP builds a bridge between observationally initialized forecasting and long-term climate projections by taking advantage of both initialization and boundary forcing information in order to enhance the magnitude of the predictable signal of climate change at this time scale. Since many decadal experiments are needed to properly evaluate the performance of a decadal prediction system, many research groups cannot afford the task of assessing the added value of DCP at regional scale through DD.

In this context, the aim of this study is to select a set of representative sub-ensembles of three members (ENS3) from the DPLE to produce decadal climate information at high resolution by DD in future studies, minimizing the loss of predictive skill and allowing a similar uncertainty representation to that for the full ensemble. In the first part of the study, an analysis to choose the most appropriate method of drift correction to minimize the drift in ENS40 has been done. The MDC, TrDC, and ICDC methods have been examined, as in their conventional as alternative (FIT and kNN) approaches, in the CORDEX EUR, SA, and NA domains. Several variables have been analyzed: SST, NSTA, and SLP. In addition, some climate indices have been considered to enhance the assessment of the added value to predictive skill by the methods. The most qualified methods for each domain have been chosen to build the ENS3 with the "best," "worst," and an "intermediate" members, in terms of their predictive skill. In the second part of the study, the performance of ENS3 in the predictive skill of SST has been evaluated along with ENS40 and all the 10 single members available for DD. Additionally, the dependence of the performance on the ensemble size has been evaluated to assess the loss of skill as consequence of selecting an sub-ensemble composed of 3 members.

In the evaluation of the correction methods, ICDC-like methods are considered as the most skilful techniques in the three domains: EUR, SA, and NA. Since ⟨RMSE⟩ scores are very similar at all lead times, this conclusion is made through analyzing ⟨ACC⟩ for the prediction of the fields and climate indices. The dependence of the model drift on the initial state of the climate system during the initialization may explain the success of ICDC-like methods with respect to the others. Similar performances among ICDC-like methods have been found in the analysis of the climate fields and none of them clearly outperforms the rest. When analyzing the climate indices, only ICDC-like methods have shown a good predictive skill, whereas MDC-like and TrDC-like methods generally does not show statistically-significant correlations. Among ICDC-like methods in this part of the study, FIT have been less skilful than the conventional and kNN versions, which have obtained again similar results. The results of these two methods are very promising in the prediction of NAO and AMV, for example, getting high correlations and multi-year scale along the whole decade and, at decadal scale, higher than those found in previous studies which used larger ensembles and/or other postprocessing techniques (Smith et al., 2019, 2020). ENSO prediction also benefits from the use of these methods as they prevent the strong loss of predictive skill along lead time observed in previous studies (Choi et al., 2016; Gonzalez & Goddard, 2016). Since the computation of the conventional version is more straightforward than for the kNN version, ICDC have been selected as the most appropriate method to correct the model drift for all variables and domains.

The predictive skill of the single members varies depending on the field and domain, lead time and even skill score under analysis. Thus, the selection of the members to build ENS3 is not simple. Since the ocean is the main reservoir of memory in the climate system, we have focused on ⟨ACC⟩ of SST at lead years 2–9 to choose among the single members. The member with the highest (the "best"), the member with the lowest (the "worst") and a member with intermediate predictive skill have been selected to build ENS3. As expected, there is a loss of predictive skill for ⟨RMSE⟩ and ⟨ACC⟩ when the ensemble size is reduced. However, the added value of increasing the number of ensemble member is very low in comparison with the increase of years which have to be simulated. ENS3 always performs better than at least the 50% of randomly generated 3-member ensembles in the deterministic skill scores, except for ⟨RMSE⟩ in the SA domain. Again, ENS3 confidence intervals at the 95% level cover the average skill scores for ENS10 (the maximum size attainable for dynamically-downscaled DPLE) in all domains excepting the SA domain. Finally, although ENS3 shows a performance comparable to those of ENS10 and ENS40 in terms of ⟨CRPSS⟩, results indicate that ensemble spread is not a good measure of prediction uncertainty on average, since statistically-significant negative values have been found in all domains.

Results showed by sub-ensemble performances in terms of SST indicate that huge investments in computing resources are necessary to get small improvements in predictive skill. This behavior is shared with other temperature variables, as shown by Reyers et al. (2019) and Sienz et al. (2016). Although the predictive skill of fields such us precipitation clearly benefits from using larger ensemble sizes (Reyers et al., 2019), the modest ENS3 could be a good alternative to very large ensembles in case of computing resources constraints when the analysis is focused on SST, since it has shown to be a relatively good representation of ENS10 and ENS40 in terms of predictive skill in this study.

Data Availability Statement

The DPLE data sets used for the analysis of hindcasts in the study are available at ESGF, with model name CESM1-1-CAM5-CMIP5, via https://doi.org/10.22033/ESGF/CMIP6.11552 with CC BY 4.0 license (Danabasoglu, 2019). The GISTEMP4 data set used for the analysis of NSTA is available at NASA Goddard Institute for Space Studies via https://data.giss.nasa.gov/gistemp (GISTEMP Team, 2023). The HadSLP2r data set used for the analysis of SLP and NAO index is available at Met Office Hadley Centre via https://www.metoffice.gov.uk/hadobs/hadslp2, subject to Crown copyright protection (Allan & Ansell, 2006). The ERSSTv5 data set used for the analysis of SST along with AMV and ENSO indices is available at NOAA National Centers for Environmental Information via https://doi.org/10.7289/V5T72FNM (Huang et al., 2017b). The ERA5 datasets used for drift correction are available at Climate Data Store via https://doi.org/10.24381/cds.f17050d7, subject to the Licence to use Copernicus Products (Hersbach et al., 2023).

The data analysis has been done by using the Python packages Xarray (Hoyer et al., 2022; Hoyer & Hamman, 2017), NumPy (Harris et al., 2020; NumPy Developers, 2022), and SciPy (Gommers et al., 2022; Virtanen et al., 2020). All figures have been made with the Python packages Matplotlib (Caswell et al., 2022; Hunter, 2007) and Cartopy (Elson et al., 2022; Met Office, 2010). Climate Data Operators (Schulzweida, 2022; Schulzweida et al., 2022) have also been used for spatial interpolation and time averaging. All software used in this work is free and open source.

Acknowledgments

J. J. Rosa-Cánovas acknowledges the Spanish Ministry of Science, Innovation and Universities for the predoctoral fellowship (grant code: PRE2018-083921). This research has been carried out in the framework of the projects PID2021-126401OB-I00, funded by MCIN/AEI/10.13039/501100011033/FEDER Una manera de hacer Europa, CGL2017-89836-R, funded by the Spanish Ministry of Economy and Competitiveness with additional FEDER funds, P20_00035, funded by FEDER/Junta de Andalucía-Consejería de Transformación Económica, Industria, Conocimiento y Universidades, and LifeWatch-2019-10-UGR-01, funded by FEDER/Ministerio de Ciencia e Innovación. We also acknowledge the Working Group on Coupled Modelling of the World Climate Research Programme, which coordinated and promoted CMIP6, and NCAR for providing the DPLE data through ESGF. Finally, we thank the reviewers of this article for their useful comments and suggestions, which really contributed to improve the original manuscript. Funding for open access charge: Universidad de Granada/CBUA.

References

Allan, R., & 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

Boer, G. J., Smith, D. M., Cassou, C., Doblas-Reyes, F., Danabasoglu, G., Kirtman, B., et al. (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

Brönnimann, S., Xoplaki, E., Casty, C., Pauling, A., & Luterbacher, J. (2006). ENSO influence on Europe during the last centuries. *Climate Dynamics*, *28*(2–3), 181–197. https://doi.org/10.1007/s00382-006-0175-z

Cai, W., McPhaden, M. J., Grimm, A. M., Rodrigues, R. R., Taschetto, A. S., Garreaud, R. D., et al. (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

Caswell, T. A., Droettboom, M., Lee, A., De Andrade, E. S., Hoffmann, T., Klymak, J., et al. (2022). matplotlib/matplotlib: REL: v3.5.2 [Software]. Zenodo. (Version 3.5.2). https://doi.org/10.5281/zenodo.6513224

Chen, D., Rojas, M., Samset, B., Cobb, K., Diongue Niang, A., Edwards, P., et al. (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, S., Hamdi, R., Ochege, F. U., Du, H., Chen, X., Yang, W., & Zhang, C. (2019). Added value of a dynamical downscaling approach for simulating precipitation and temperature over Tianshan Mountains area, central Asia. *Journal of Geophysical Research: Atmospheres*, *124*(21), 11051–11069. https://doi.org/10.1029/2019JD031016

Choi, J., Son, S.-W., Ham, Y.-G., Lee, J.-Y., & Kim, H.-M. (2016). Seasonal-to-interannual prediction skills of near-surface air temperature in the CMIP5 decadal hindcast experiments. *Journal of Climate*, *29*(4), 1511–1527. https://doi.org/10.1175/JCLI-D-15-0182.1

Choudhury, D., Sen Gupta, A., Sharma, A., Mehrotra, R., & Sivakumar, B. (2017). An assessment of drift correction alternatives for CMIP5 decadal predictions. *Journal of Geophysical Research: Atmospheres*, *122*(19), 10282. https://doi.org/10.1002/2017JD026900

Choudhury, D., Sharma, A., Sen Gupta, A., Mehrotra, R., & Sivakumar, B. (2016). Sampling biases in CMIP5 decadal forecasts. *Journal of Geophysical Research: Atmospheres*, *121*(7), 3435–3445. https://doi.org/10.1002/2016JD024804

CLIVAR. (2011). *Data and bias correction for decadal climate predictions (CLIVAR Publication Series No. 150)*. International CLIVAR Office Project.

Danabasoglu, G. (2019). NCAR CESM1-1-CAM5-CMIP5 model output prepared for CMIP6 DCPP dcppA-hindcast [Dataset]. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.11552

Danabasoglu, G., Bates, S. C., Briegleb, B. P., Jayne, S. R., Jochum, M., Large, W. G., et al. (2012). The CCSM4 ocean component. *Journal of Climate*, *25*(5), 1361–1389. https://doi.org/10.1175/JCLI-D-11-00091.1

Ehmele, F., Kautz, L.-A., Feldmann, H., & Pinto, J. G. (2020). Long-term variance of heavy precipitation across central Europe using a large ensemble of regional climate model simulations. *Earth System Dynamics*, *11*(2), 469–490. https://doi.org/10.5194/esd-11-469-2020

Elson, P., De Andrade, E. S., Lucas, G., May, R., Hattersley, R., Campbell, E., et al. (2022). SciTools/cartopy: v0.20.2 [Software]. Zenodo. (Version 0.20.2). https://doi.org/10.5281/zenodo.5842769

Feldmann, H., Pinto, J. G., Laube, N., Uhlig, M., Moemken, J., Pasternack, A., et al. (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

Fučkar, N. S., Volpi, D., Guemas, V., & 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

Gangstø, R., Weigel, A., Liniger, M., & 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 Ojeda, M., Gámiz-Fortis, S. R., Castro-Díez, Y., & Esteban-Parra, M. J. (2017). Evaluation of WRF capability to detect dry and wet periods in Spain using drought indices. *Journal of Geophysical Research: Atmospheres*, *122*(3), 1569–1594. https://doi.org/10.1002/2016JD025683

García-Valdecasas Ojeda, M., Rosa-Cánovas, J. J., Romero-Jiménez, E., Yeste, P., Gámiz-Fortis, S. R., Castro-Díez, Y., & 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 Ojeda, M., Yeste, P., Gámiz-Fortis, S. R., Castro-Díez, Y., & 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

Giorgi, F., Jones, C., & Asrar, G. R. (2009). Addressing climate information needs at the regional level: The CORDEX framework. *World Meteorological Organization Bulletin*, *58*(3), 175–183.

Giorgi, F., & Mearns, L. O. (1991). Approaches to the simulation of regional climate change: A review. *Reviews of Geophysics*, *29*(2), 191–216. https://doi.org/10.1029/90RG02636

GISTEMP Team. (2023). GISS surface temperature analysis (GISTEMP) [Dataset]. NASA Goddard Institute for Space Studies. Retrieved from https://data.giss.nasa.gov/gistemp/

Gneiting, T., & 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., Kumar, A., Solomon, A., Smith, D., Boer, G., Gonzalez, P., et al. (2013). A verification framework for interannual-to-decadal predictions experiments. *Climate Dynamics*, *40*(1–2), 245–272. https://doi.org/10.1007/s00382-012-1481-2

Gommers, R., Virtanen, P., Burovski, E., Weckesser, W., Oliphant, T. E., Haberland, M., et al. (2022). scipy/scipy: SciPy 1.8.1 [Software]. Zenodo. (Version 1.8.1). https://doi.org/10.5281/zenodo.6560517

Gonzalez, P. L. M., & Goddard, L. (2016). Long-lead ENSO predictability from CMIP5 decadal hindcasts. *Climate Dynamics*, *46*(9–10), 3127–3147. https://doi.org/10.1007/s00382-015-2757-0

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz-Sabater, J., et al. (2023). ERA5 monthly averaged data on single levels from 1940 to present [Dataset]. Copernicus Climate Change Service (C3S): Climate Data Store (CDS). https://doi.org/10.24381/cds.f17050d7

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Hoyer, S., & 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

Hoyer, S., Roos, M., Joseph, H., Magin, J., Cherian, D., Fitzgerald, C., et al. (2022). xarray [Software]. Zenodo. (Version 2022.03.0). https://doi.org/10.5281/zenodo.6323468

Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., et al. (2017a). 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

Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., et al. (2017b). NOAA extended reconstructed sea surface temperature (ERSST), version 5 [Dataset]. NOAA National Centers for Environmental Information. https://doi.org/10.7289/V5T72FNM

Hunke, E. C., & 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., Gent, P. R., Ghan, S., Kay, J. E., Kushner, P. J., et al. (2013). The community Earth system model: A framework for collaborative research. *Bulletin of the American Meteorological Society*. https://doi.org/10.1175/BAMS-D-12-00121

Hurrell, J. W., Kushnir, Y., Ottersen, G., & Visbeck, M. (2003). An overview of the North Atlantic oscillation. In J. W. Hurrell, Y. Kushnir, G. Ottersen, & M. Visbeck (Eds.), *Geophysical monograph series* (Vol. 134, pp. 1–35). American Geophysical Union. https://doi.org/10.1029/134GM01

Infanti, J. M., & Kirtman, B. P. (2016). North American rainfall and temperature prediction response to the diversity of ENSO. *Climate Dynamics*, *46*(9–10), 3007–3023. https://doi.org/10.1007/s00382-015-2749-0

Jolliffe, I. T. & Stephenson, D. B. (Eds.) (2012). *Forecast verification: A practitioner's guide in atmospheric science* (2nd ed.). Wiley-Blackwell.

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., et al. (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., & Lee, W.-S. (2012). Statistical adjustment of decadal predictions in a changing climate. *Geophysical Research Letters*, *39*(19), GL05267. https://doi.org/10.1029/2012GL052647

Knight, J. R., Folland, C. K., & Scaife, A. A. (2006). Climate impacts of the Atlantic multidecadal oscillation. *Geophysical Research Letters*, *33*(17), L17706. https://doi.org/10.1029/2006GL026242

Kotamarthi, R., Hayhoe, K., Mearns, L., Wuebbles, D., Jacobs, J., & Jurado, J. (2021). *Downscaling techniques for high-resolution climate projections: From global change to local impacts* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781108601269

Kruschke, T., Rust, H. W., Kadow, C., Müller, W. A., Pohlmann, H., Leckebusch, G. C., & 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

Kushnir, Y., Scaife, A. A., Arritt, R., Balsamo, G., Boer, G., Doblas-Reyes, F., et al. (2019). Towards operational predictions of the near-term climate. *Nature Climate Change*, *9*(2), 94–101. https://doi.org/10.1038/s41558-018-0359-7

Lamarque, J.-F., Bond, T. C., Eyring, V., Granier, C., Heil, A., Klimont, Z., et al. (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

Lawrence, D. M., Oleson, K. W., Flanner, M. G., Thornton, P. E., Swenson, S. C., Lawrence, P. J., et al. (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

Lenssen, N. J. L., Schmidt, G. A., Hansen, J. E., Menne, M. J., Persin, A., Ruedy, R., & Zyss, D. (2019). Improvements in the GISTEMP uncertainty model. *Journal of Geophysical Research: Atmospheres*, *124*(12), 6307–6326. https://doi.org/10.1029/2018JD029522

Matei, D., Pohlmann, H., Jungclaus, J., Müller, W., Haak, H., & 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

Meehl, G. A., Goddard, L., Boer, G., Burgman, R., Branstator, G., Cassou, C., et al. (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., et al. (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., et al. (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., et al. (2011). The RCP greenhouse gas concentrations and their extensions from 1765 to 2300. *Climatic Change*, *109*(1–2), 213–241. https://doi.org/10.1007/s10584-011-0156-z

Met Office. (2010). *Cartopy: A cartographic Python library with a Matplotlib interface*. Exeter, Devon. Retrieved from https://scitools.org.uk/cartopy

NumPy Developers. (2022). numpy [Software]. Github (Version 1.22.4). Retrieved from https://github.com/numpy/numpy/releases/tag/v1.22.4

Paeth, H., Li, J., Pollinger, F., Müller, W. A., Pohlmann, H., Feldmann, H., & Panitz, H.-J. (2019). An effective drift correction for dynamical downscaling of decadal global climate predictions. *Climate Dynamics*, *52*(3–4), 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., et al. (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

Posada-Marín, J. A., Rendón, A. M., Salazar, J. F., Mejía, J. F., & Villegas, J. C. (2019). WRF downscaling improves ERA-Interim representation of precipitation around a tropical Andean valley during El Niño: Implications for GCM-scale simulation of precipitation over complex terrain. *Climate Dynamics*, *52*(5–6), 3609–3629. https://doi.org/10.1007/s00382-018-4403-0

Randall, D. A., Bitz, C. M., Danabasoglu, G., Denning, A. S., Gent, P. R., Gettelman, A., et al. (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

Reyers, M., Feldmann, H., Mieruch, S., Pinto, J. G., Uhlig, M., Ahrens, B., et al. (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

Schulzweida, U. (2022). CDO user guide. *Zenodo*. https://doi.org/10.5281/zenodo.5614769(v2.0.5)

Schulzweida, U., Mueller, R., Heidmann, O., Kornblueh, L., Wachsmann, F., Ansorge, C., et al. (2022). Climate data operators [Software]. Max-Planck-Institute for Meteorology. Retrieved from https://code.mpimet.mpg.de/attachments/26823(Version2.0.5)

Sienz, F., Müller, W. A., & 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

Smith, D. M., Eade, R., Scaife, A. A., Caron, L.-P., Danabasoglu, G., DelSole, T. M., et al. (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., et al. (2020). North Atlantic climate far more predictable than models imply. *Nature*, *583*(7818), 796–800. https://doi.org/10.1038/s41586-020-2525-0

Smith, D. M., Scaife, A. A., Hawkins, E., Bilbao, R., Boer, G. J., Caian, M., et al. (2018). Predicted chance that global warming will temporarily exceed 1.5°C. *Geophysical Research Letters*, *45*(21), 11895. https://doi.org/10.1029/2018GL079362

Trenberth, K. E., & 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., & 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

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (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

Yeager, S. G., Danabasoglu, G., Rosenbloom, N. A., Strand, W., Bates, S. C., Meehl, G. A., et al. (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., & 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