Tesis Doctoral

Metodologías de simulación de agentes naturales y desarrollo de sistemas. Modelo de Verificación y Gestión de Terminales Portuarias.

Aplicación al puerto de la Bahía de Cádiz.

Presentada por:
Sebastián Solari

Directores:
Dr. Miguel A. Losada
Dr. Antonio Moñino

Programa Oficial de Posgrado
Dinámica de Flujos Biogeoquímicos y sus Aplicaciones
Universidades de Granada, Córdoba y Málaga

Editor: Editorial de la Universidad de Granada
Autor: Sebastián Solari
D.L.: GR 1157-2012
ISBN: 978-84-695-1165-7

Resumen

En este trabajo se profundiza en el conocimiento y desarrollo de las metodologías y herramientas necesarias para diseñar, verificar y optimizar sistemas costeros y portuarios mediante técnicas probabilistas y cálculo de riesgos.

Para ello se trabaja en dos frentes: por un lado se desarrollan metodologías de simulación de series temporales multivariadas para las variables de estado que caracterizan a los principales agentes forzantes de los sistemas costeros y portuarios; por otro se desarrolla una metodología de verificación y optimización con base en riesgo de las áreas navegables de un puerto, la cual se aplica en un caso de estudio para optimizar la profundidad de dragado de los canales de navegación del Puerto de la Bahía de Cádiz.

El proceso de simulación de series temporales requiere conocer la distribución marginal de las variables aleatorias, su dependencia temporal y la interdependencia entre las distintas variables. Todo esto debe representarse adecuadamente mediante modelos paramétricos que permitan simulaciones eficientes y con capacidad de innovación. Además estos modelos deben contemplar que los valores observados provienen de distintas poblaciones, de modo de representar adecuadamente tanto el comportamiento medio como el extremal, y los ciclos de variación estacionales y plurianual de las variables.

En primer lugar se explora el uso de distribuciones mixtas para modelar la distribución marginal de variables geofísicas, consistentes en una distribución central truncada, representativa del régimen central o el grueso de los datos, y de dos distribuciones Pareto generalizadas para los regímenes de máximos y mínimos, representativas de las colas superior e inferior, respectivamente. Los umbrales que definen los límites entre estos regímenes y el régimen central son parámetros del modelo y se calculan junto con los restantes parámetros del mismo mediante máxima verosimilitud.

El modelo se aplica a cuatro series de datos hidrológicos: dos de caudal medio diario, correspondientes a los ríos Támesis (Reino Unido) y Guadalfeo (España), y dos de precipitación diaria, correspondientes a Fort Collins (CO,EEUU) y Órgiva (España); y a dos series de altura de ola significante: Cádiz y Barcelona, España. En todos los casos se observó que los modelos mixtos mejoran el ajuste de los datos respecto al ajuste obtenido con las funciones de distribución paramétricas comúnmente usadas, y en particular se observó que mejoran el ajuste en la cola superior.

Adicionalmente, se concluye que el modelos mixto propuesto es capaz de representar correctamente todo el rango de valores de algunas variables geofísicas importantes, proporcionando también una herramienta automática y objetiva para la estimación del umbral necesario para aplicar el método de picos sobre el umbral. Para la aplicación de

este último se ha propuesto una metodología simple que permite incluir le incertidumbre del umbral en el cálculo de la incertidumbre de los cuantiles.

Luego, el modelo mixto propuesto se extiende a condiciones no estacionarias, para lo cual sus parámetros se modelan mediante series de Fourier de distintas escalas temporales. Esto permite incluir en el modelo tanto los ciclos estacionales de escala anual como los ciclos plurianual, característicos de las variables geofísicas.

En segundo lugar se estudian distintos métodos de simulación de series temporales uni- y multivariadas. Los métodos más populares para la simulación de series temporales de variables atmosféricas y oceanográficas están basados en el uso de modelos autorregresivos, previa transformación de las variables para hacerlas estacionarias y normalizadas. En general, cuando se utilizan estos modelos, la atención está centrada en la capacidad de los mismos para reproducir la autocorrelación, y correlación cruzada cuando corresponde, de las series originales.

Se propone un modelo de simulación de series temporales univariadas basado en: (i) el modelo mixto no estacionario analizado anteriormente, y (ii) el uso de copulas para modelar la dependencia temporal de la variable. El modelo propuesto se evaluó comparando las series originales con las simuladas en términos de su autocorrelación, los regímenes medios y máximos (anuales y picos sobre el umbral), y los regímenes de persistencias por encima de distintos umbrales. También se compararon los resultados con los obtenidos con modelos autorregresivos de media móvil, concluyéndose que el modelo propuesto proporciona mejores resultados que estos últimos.

Luego se estudia el uso de modelos autorregresivos vectoriales (VAR), y modelos VAR con cambio de régimen, para simular series temporales de altura, dirección y período de ola, y velocidad y dirección de viento. Para normalizar las series y hacerlas estacionarias se ajusta una función de distribución mixta no estacionaria a cada una de las variable. Luego se ajustan los tres modelos VAR (uno estándar y dos con cambio de régimen) a las series normalizadas estacionarias, y finalmente se simulan nuevas series temporales multivariadas. Se realiza un análisis en profundidad de las series temporales simuladas, comparando sus características con las de las series originales. Se encuentra que los modelos VAR logran capturar las características principales de las series originales, pero fallan en reproducir algunos de los regímenes de persistencia, y no son capaces de reproducir todas las características de las distribuciones marginales bivariadas. Por otro lado, aunque los modelos VAR con cambio de régimen mejoran algunos aspectos de las simulaciones, producen resultados inesperados en cuanto a la autocorrelación de las series simuladas.

Por último, se presenta el desarrollo y la aplicación de un modelo de simulación para la verificación y el diseño del calado y el ancho de las áreas navegables de un puerto, atendiendo los modos de fallo toque de fondo y salidas de márgenes, y la parada operativa del canal.

Utilizando como forzante las series temporales simuladas con la metodología antes descrita, el modelo simula los tránsitos de entrada y salida del puerto, y en cada tránsito calcula los tiempos de espera y la probabilidad de que el barco toque fondo. Los resultados de las simulaciones se usan para calcular el riesgo en la vida útil del canal, así como el desempeño del mismo en cuanto a tiempos de espera y operatividad.

El modelo propuesto se aplica en un caso de estudio para la optimización del calado del canal de acceso de la futura terminal de contenedores del Puerto de la Bahía de Cádiz, España.

Abstract

This work deepens the knowledge and the development of methodologies and tools to design, verify and optimize coastal and port systems by probabilistic techniques and risk calculation.

To achieve this it works on two fronts: firstly develops simulation methodologies for multivariate time series of the state variables that characterize the predominant forcing agents of coasts and ports, secondly it develops a methodology for the verification and risk-based optimization of the navigable areas of a port, which is applied in a case study to optimize the depth of the entrance channel of the Port of the Bay of Cadiz.

The time series simulation process has to take into account the marginal distribution of the variables, their time dependence, and the interdependence between the different variables. All this must be adequately represented by parametric models that allow efficient simulations, and that are capable of innovation. Furthermore, these models must consider that the observed values come from different populations, in order to adequately represent both the central and the extreme behavior, and the seasonal and pluriannual cycles of the variables.

First it is explored the use of a mixture model for determining the marginal distribution of geophysical variables, consisting of a truncated central distribution that is representative of the central or main-mass regime, and of two generalized Pareto distributions for the maximum and minimum regimes, representatives of the upper and lower tails respectively. The thresholds defining the limits between these regimes and the central regime are parameters of the model and are calculated together with the remaining parameters by maximum likelihood.

The model was applied to four hydrological data series: two of mean daily flow, the Thames at Kingston (United Kingdom) and the Guadalfeo River at Orgiva (Spain), and two of daily precipitation, Fort Collins (CO, USA) and Orgiva (Spain); and to two significant wave height data series: Cádiz and Barcelona, Spain. For all cases, it was observed that the mixture model improved the fit of the data series with respect to the fit obtained with commonly used parametric distributions and, in particular, provided a good fit for the upper tail.

Moreover, it can be concluded that the proposed mixture model is able to accommodate the entire range of values of some significant geophysical variables, yielding an automatic and objective identification of the threshold required for the application of the peaks over threshold method. For this purpose, a simple methodology has been devised for including threshold uncertainties in quantile calculations.

Then, the proposed mixture model is extended to non-stationary conditions, for which the parameters are modeled using Fourier series of different scales. This allows the model to include both the annual scale seasonal cycles as well as pluriannual cycles characteristic of geophysical variables.

Secondly, an study on univariate and multivariate time series simulation methodologies is performed. The most popular methods of simulating time series of atmospheric and oceanographic variables are based on the use of autoregressive models and the transformation of variables to make them normal and stationary. Generally, when these models are used, attention is centred on their capacity to represent the autocorrelation of the series.

A simulation model is proposed for univariate series that is based on the following: (i) the non-stationary parametric mixture model previously analyzed, and (ii) the use of copulas to model the time dependency of the variable. The model has been evaluated by comparing the original series and the simulated series in terms of the autocorrelation function, the central, the annual maxima and peaks-over-threshold regimes, and the persistences regime. It has also been compared to an ARMA model and found to yield more satisfactory results.

Afterwards, the use of vector autorregresive (VAR) and Regime Switching VAR models for the simulation of wave height, period and direction, and wind speed and direction, is studied. In order to normalize and stationarize the series, non-stationary mixture uni-variate distributions are fitted to the above five variables. Then the three different VAR models (standard model and two regime switching models) are fitted and new time series are simulated. An in depth analysis of the long term simulations is performed, in order to study its ability to reproduce the behavior of the original series. It is found that VAR models are able to capture main features of the original series, but they fail in reproducing some of the persistence regimes and some aspects of the bi-variate distributions. On the other hand, although Regime Switching VAR models improve some aspects of the simulations, they produce some unexpected behavior in the correlation of the simulated series.

Finally, it is presented the development and the implementation of a simulation model for verification and design of the depth and width of the navigable areas of a port, that takes into account the failure modes of the ships in transit as well as the operational stoppage of the channel.

Using as input the time series simulated with the previously described methodologies, the model simulate entrance and exit transits in the port, and for each transit it calculates waiting time, and the probability of the ship touching the bottom. The outputs of the simulation are used to calculate the total risk in the useful life of the channel, as well as its performance in regards with waiting times and operability.

The proposed model is applied to a case study for the optimization of the depth of the entrance channel of the new container terminal of the Port of the Bay of Cádiz.

Contenido

Lista de figuras

Lista de tablas

Introducción

Motivación

Puertos y costas son sistemas complejos, en los que las acciones de los agentes climáticos, tanto oceánicos como atmosféricos, son predominantes. Estos sistemas por tanto están sujetos a variabilidad de diversas escalas temporales y espaciales, así como a un alto grado de incertidumbre, ambas características inherentes de las variables geofísicas.

En un contexto de demanda creciente sobre los recursos de la costa –agua, suelo y energía–, de contracción de la inversión por parte de los organismos públicos y privados y de crecientes restricciones ambientales, proyectistas y gestores necesitan metodologías y herramientas que les permitan optimizar el diseño y la explotación de los sistemas costeros y portuarios desde el punto de vista económico, social y ambiental, asegurando que los mismos cumplirán a lo largo de su vida útil con los requisitos de seguridad, operatividad y conservación del medio ambiente y el patrimonio histórico y cultural que impone la normativa vigente. Para esto, proyectistas y gestores deben ser capaces de acotar la incertidumbre propia de estos sistemas, teniendo en cuenta las distintas escalas de variabilidad temporal y espacial características de los mismos, así como las tendencias que a mediano plazo impone el cambio global.

Los métodos de verificación y diseño probabilistas, conocidos como métodos de Nivel II y III, así como los métodos basados en el cálculo del riesgo y/o el beneficio, dan un marco teórico de trabajo para realizar esta tarea. Sin embargo la aplicación de los mismos no siempre es inmediata, ya que se requiere conocer, entender, modelar y predecir el comportamiento de: (a) los agentes externos que ejercen acciones sobre el sistema, en particular los agentes climáticos, y (b) la respuesta de cada uno de los componentes del sistema y del sistema en su conjunto frente a estas acciones.

Respecto a los agentes climáticos, y en particular los marinos y atmosféricos, cabe decir que en la última década se han realizado esfuerzos importantes en la generación de bases de datos de larga duración y alta densidad espacial. Por lo tanto, los desafíos que enfrenta hoy la investigación científica en el área de la ingeniería de costas no se refieren tanto a la obtención de nueva información sino al diseño de metodologías y herramientas para el cabal aprovechamiento de la existente. Una de las líneas de investigación abiertas y en la que esta tesis profundiza es el modelado y simulación de series temporales multivariadas no estacionarias de variables geofísicas. Disponer de estas metodologías es una condición necesaria para poder aplicar métodos de verificación y optmización de Nivel III basados en simulaciones de Monte Carlo, en donde la calidad

de las series simuladas determina la calidad de los resultados obtenidos. Por lo tanto una mejora en los métodos de simulación de series temporales implica una mejora en los procesos de diseño y optimización de los sistemas costeros y portuarios.

En lo que se refiere a los sistemas, existe un mayor conocimiento de la respuesta de cada uno de sus componentes por separado del que existe de la respuesta del sistema en su conjunto. La respuesta de los componentes tiene en general una parte aleatoria, inherente a cada componente, y por lo tanto la misma está dada en términos de probabilidad. A su vez estas respuestas están dispersas en el tiempo y en el espacio. Cómo integrar la respuesta de los componentes individuales de forma correcta desde un punto de vista teórico, tanto físico como matemático, y a la vez funcional, de forma tal que permita el cálculo del riesgo y de la incertidumbre, es una línea de investigación abierta.

Uno de los sistemas que estudia la ingeniería portuaria es el compuesto por las áreas de navegación y maniobras de un puerto. A la complejidad consustancial de todos los sistemas portuarios, este suma la particularidad que el elemento sujeto a fallo –el buque– está en movimiento dentro del sistema. En este sentido, el desarrollo de un marco teórico para la verificación y optimización de canales de navegación mediante técnicas probabilistas, que sea coherente con la metodología de trabajo de las Recomendaciones para Obras Marítimas de Puertos del Estado, en particular con la ROM 0.0–Procedimiento general y bases de cálculo en el proyecto de obras marítimas y portuarias, es un problema aún no resuelto en la ingeniería portuaria, cuya resolución puede tener importantes repercusiones económicas y ambientales.

Objetivos

El objetivo de este trabajo es profundizar en el conocimiento y desarrollo de las metodologías y herramientas necesarias para diseñar, verificar y optimizar sistemas costeros y portuarios mediante técnicas probabilistas y cálculo de riesgos.

Para ello se plantean dos objetivos generales. Por un lado desarrollar metodologías de simulación de series temporales multivariadas de las variables de estado que caracterizan a los principales agentes forzantes de los sistemas costeros y portuarios. Por otro lado desarrollar, e implementar para un caso de estudio, una metodología de verificación y optimización de canales de navegación.

El primer objetivo es por tanto investigar metodologías de simulación de series temporales de variables geofísicas, particularmente variables de estado oceanográficas y atmosféricas como ser la altura de ola significante y la velocidad de viento media a $10\,m$. Estas metodologías deben permitir generar nuevas series aleatorias que tengan las mismas características que las series originales. En términos generales se busca que las series simuladas compartan con las originales sus distribuciones de probabilidad marginales, tanto univariadas como bivariadas, su estructura de dependencia temporal y los ciclos de variación intra-anual –estaciones– y pluri-anual, y que reproduzcan la variabilidad interanual característica de los sistemas naturales.

El segundo objetivo de esta tesis es definir, e implementar para un caso de estudio, una metodología de verificación y optimización mediante técnicas de Nivel III y cálculo

de riesgo de un sistema formado por las áreas navegables de un puerto. Para esto es necesario utilizar la metodología de simulación de series temporales discutida anteriormente, ya que estas series son el principal forzante del sistema. Sin embargo para la verificación del mismo también es necesario definir un marco teórico y una metodología de trabajo que permitan calcular la respuesta del sistema frente a estos forzantes y expresar esta respuesta en términos probabilistas y de riesgos.

El caso de estudio seleccionado para aplicar las metodologías desarrolladas es la ampliación del puerto de la Bahía de Cádiz, España.

Objetivos específicos

La consecución de los objetivos generales de esta tesis se logra una vez alcanzados los siguientes objetivos específicos.

- Desarrollar funciones de distribución de probabilidad que modelen adecuadamente todo el rango de valores de las variables aleatorias geofísicas, e implementar metodologías de estimación de los parámetros de las mismas.
- Implementar una metodología para incluir en las funciones desarrolladas las variaciones temporales de distintas escalas, características de las variables geofísicas.
- Investigar modelos de dependencia temporal (auto-correlación) para simular nuevas series temporales univariadas.
- Investigar modelos de dependencia temporal (auto-correlación) y dependencia entre variables (correlación cruzada) para simular series temporales multivariadas.
- Definir formalmente una metodología basada en simulaciones de Monte Carlo para el cálculo del riesgo y el desempeño de las areas navegables de un puerto.
- Implementar esta metodología en una herramienta informática y aplicarla para la optimización de la profundidad de dragado del canal de navegación y el área de maniobras del puerto de la Bahía de Cádiz.

Organización de la tesis

La tesis se compone de 5 capítulos en los que se abordan los objetivos específicos planteados en la sección anterior. Cada uno de estos capítulos corresponde a un artículo que ha sido publicado o se encuentra en proceso de publicación. Estos artículos se reproducen íntegramente tal cual fueron enviados a publicación excepto por algunos cambios en el formato

Los siguientes capítulos reproducen el camino transitado para resolver de forma novedosa un problema característico de la ingeniería portuaria, a saber: la optimización de la profundidad de los canales de navegación. Partiendo de una serie de datos se alcanza un procedimiento de simulación de series temporales multivariadas. Luego se

define una metodología de trabajo que, utilizando las series simuladas, calcula el riesgo en la vida útil de las áreas nevegables de un puerto.

Capítulo 1: A unified statistical model for hydrological variables including the selection of threshold for the POT method[1].

Se analiza el uso de modelos de distribución mixtos para modelar todo el rango de valores de una variable geofísica. Éste es el primer paso para lograr un proceso de simulación aleatoria de series temporales que conserven las mismas características, desde el punto de vista estadístico, que las series originales.

El modelo y la metodología de trabajo propuestos se aplican a cuatro series de precipitación y caudales de diversas características.

Capítulo 2: Unified distribution models for met-ocean variables: application to series of significant wave height[2].

En este capítulo se extiende el uso de los modelos propuestos en el capítulo 1 para series de altura de ola significante. A su vez se propone una distribución de probabilidad alternativa a la propuesta en el capítulo 1, la cual, en ciertas condiciones, presenta un mejor ajuste de los datos.

Los modelos propuestos se aplican a dos series de alturas de ola significante provenientes de retroanálisis, una correspondiente a la costa andaluza, en el Golfo de Cádiz, y otra correspondiente a la costa catalana, en el Mar Mediterráneo.

Capítulo 3: Non-stationary wave height climate modeling and simulation[3].

Partiendo del modelo de distribución mixto presentado en los capítulos 1 y 2, se procede a extender el mismo a condiciones no estacionarias. Luego, utilizando este modelo para la normalización de la serie de datos, se propone una metodología de simulación de series temporales basada en el uso de copulas.

La metodología propuesta se aplica a la serie alturas de ola significantes del Golfo de Cádiz, y se compara su desempeño con el obtenido con otros modelos disponibles en la bibliografía –modelos AR y ARMA–, concluyéndose que las series temporales simuladas con el modelo propuesto reproducen mejor el comportamiento de las series originales que las series simuladas con los otros modelos.

Capítulo 4: On the use of Vector Autoregressive (VAR) and Regime Switching VAR models for the simulation of sea and wind state parameters[4].

Se extiende el uso de distribuciones mixtas no estacionarias, incluyendo el modelo propuesto en los capítulos 1 y 2, para modelar la distribución de probabilidad de cinco

[1] En revisión en Water Resource Research.

[2] En revisión en Coastal Engineering.

[3] Solari, S., & Losada, M. A. (2011). Non-stationary wave height climate modeling and simulation. *Journal of Geophysical Research*, 116(C09032), 1-18. doi:10.1029/2011JC007101.

[4] Aceptado para su publicación en *CENTEC Anniversary Book*, editado por el Prof. Guedes Soares del Instituto Superior Técnico de Lisboa.

variables aleatorias: altura de ola significante, período de pico y dirección media del oleaje y velocidad y dirección de viento.

Normalizadas las cinco variables mediante los modelos anteriores, se investiga el uso de tres modelos autorregresivos vectoriales (VAR), a saber: modelo VAR estándar, modelo VAR autoexitante con cambio de régimen y modelo VAR con cambio de régimen según cadena de Markov.

Los tres modelos analizados se utilizan para simular nuevas series temporales multivariadas no estacionarias y se compara el desempeño de los mismos en base a la similitud existente entre las series originales y las series simuladas.

Capítulo 5: Diseño de la profundidad de un canal de navegación con base en riesgo y uso y explotación[5].

En este capítulo se desarrolla una metodología de trabajo para calcular el riesgo y el desempeño a lo largo de la vida útil de los canales de navegación y las áreas de maniobra de un puerto. La metodología desarrollada se basa en técnicas de simulación de Monte Carlo, para lo cuál se utilizan los modelos desarrollados en los capítulos 1 a 4.

La metodología propuesta da un marco teórico de trabajo aplicable a cualquier puerto. El mismo es aplicado en un caso de estudio en el Puerto de la Bahía de Cádiz, para definir la profundidad de dragado de los canales de navegación que minimiza el riesgo en la vida útil de los mismos.

Por último se incluye un capítulo de **Conclusiones** en el cual se hace un resumen de las conclusiones alcanzadas en cada uno de los capítulos anteriores y se presentan las conclusiones globales que se desprenden de esta tesis, a la vez que se señalan las líneas de trabajo que quedan abiertas.

[5] A ser enviado al Journal of Waterway, Port, Coastal, and Ocean Engineering. Resultados parciales publicados en:
Solari, Moñino, Baquerizo & Losada (2010) Simulation Model for Harbor Verification and Managment. In J. McKee Smith & P. Lynett (Eds.), *Proceedings of 32nd Conference on Coastal Engineering, Shanghai, China, 2010.* Coastal Engineering Research Council, ASCE.
http://journals.tdl.org/ICCE/article/view/1294

Chapter 1

A unified statistical model for hydrological variables including the selection of threshold for the POT method

1.1 Abstract

This paper explores the use of a mixture model (LNGPD) for determining the marginal distribution of hydrological variables, consisting of a truncated central distribution that is representative of the central or main-mass regime, which for the cases studied is a log normal distribution (LN), and of two generalized Pareto distributions (GPD) for the maximum and minimum regimes, representatives of the upper and lower tails respectively. The thresholds defining the limits between these regimes and the central regime are parameters of the model and are calculated together with the remaining parameters by maximum likelihood (ML). The model was applied to four hydrological data series: two of mean daily flow, the Thames at Kingston (United Kingdom) and the Guadalfeo River at Orgiva (Spain), and two of daily precipitation, Fort Collins (CO, USA) and Orgiva (Spain). For all four cases, it was observed that the LNGPD mixture model improved the fit of the data series with respect to the fit obtained with the LN and, in particular, provided a good fit for the upper tail. When the fit of the parameters for the minimum regime gives a threshold value lower than the minimum value of the series, the GPD of the minima did not improve the fit of the lower tail. Moreover, it can be concluded that the LNGPD mixture model is able to accommodate the entire range of values of some significant hydrological variables, yielding an automatic and objective identification of the threshold required for the application of the peaks over threshold (POT) method. For this purpose, a simple methodology has been devised for including threshold uncertainties in quantile calculations.

1.2 Introduction

In the last several years, the development and use of mixture distribution models has increased due to their flexibility (*Evin et al.*, 2011), in that they allow the user to model series of data from different populations. These models have been used with success for various geophysical applications, e.g., statistical downscaling of precipitation (*Vrac and Naveau*, 2007), simulating extreme rainfall events (*Furrer and Katz*, 2008), flow analysis (*Evin et al.*, 2011) and time series simulations of sea state parameters (*Solari and Losada*, 2011).

In general, the mixture models used have been composed of a central distribution and one (or two) distribution(s) of the tail(s). For these models, either the transition (threshold) values between the central and tail distributions are left undefined (e.g., *Vrac and Naveau*, 2007; *Hundecha et al.*, 2009, use dynamic models imposing the threshold equal to zero), or these are defined a priori using a different method from that used to estimate the other parameters of the model (*Furrer and Katz*, 2008). Recently, *Carreau et al.* (2009) expressed threshold values as a function of the other parameters of the model. Therefore, it should be investigated whether it is possible to include the determination of threshold values in the estimation method used for the distribution parameters. If so, it seems reasonable to explore whether the threshold value of the upper tail is a good choice as the estimate of the threshold value required to apply the peaks over threshold (POT) method. This would provide another way to estimate threshold values that is complementary to that described in *Coles* (2001), which unfortunately cannot be automated and requires user intervention in the process. The aim of this paper is to analyze the potential for applying a mixture model to parametrically model the entire range of values of hydrological variables and use the values of the upper thresholds that are obtained to define the series of peaks in the POT method.

As an alternative to previously used models, we propose the use of a mixture model that is composed of a truncated central distribution, representative of the central regime, and two generalized Pareto distributions (GPD) for the upper and lower tails, representing the maximum and minimum regimes, respectively. The transition thresholds between the three distributions are parameters of the model and are calculated by maximum likelihood (ML) simultaneously with the other parameters in the model.

This paper is organized as follows:

In Section 4.3, the background, different models and methodologies that have been used are presented, and their advantages and disadvantages are discussed. In Section 1.4, an alternative mixture model and a working methodology are introduced to solve the main problems identified in the previous section. In Section 1.5, the results from applying the model to four data sets are presented, analyzing its capacity to fit the entire ranges of the values of the variables, particularly in the tails, and the potential for the use of the upper threshold when applying the POT method. Particular attention is paid to the following: (a) the practical aspects of applying the method and (b) the relationship between the physical characteristics of the system from which the variables originated and the results obtained with the model; in particular, with respect to obtaining or not obtaining heavy tails for the maxima regime (*Vrac and Naveau*, 2007). Finally, Section

1.6 summarizes the conclusions, and in the appendices, specific aspects concerning the estimation methodology of the model parameters are discussed.

1.3 Background

1.3.1 Central regime

When modeling the bulk of the data for hydrologic variables such as stream flow and precipitation, it is common practice to use biparametric distributions such as the gamma, log-normal (LN) or biparametric Weibull for minima (WB) distributions (*Chow*, 1988). Usually, these models provide a good fit with the data in the central area around the mean and the mode, but not at the tails (see, e.g., *Furrer and Katz*, 2008).

There is no theoretical justification for choosing one specific model for the distribution of hydrological variables. However, given the amount of data recorded for the central regime, empirical distribution functions usually provide a sufficiently good estimate of the main-mass of the data. Therefore, adequate parametric modeling is not essential for modeling the central regime.

1.3.2 Extreme climate

Parametric modeling for extreme conditions is required when attempting to infer unrecorded conditions from available data.

Extreme value theory states that according to certain hypotheses, the distribution of the maxima or minima of an independent and identically distributed (i.i.d.) series of n elements tends to have one of the three forms of the generalized extreme value distribution (GEVD). According to these hypotheses, the distribution of the values that exceed a given threshold of a series of i.i.d. data tends to have a generalized Pareto distribution (GPD) when the threshold tends towards the upper bound of the variable (see, e.g., *Coles*, 2001; *Castillo et al.*, 2005; *Kottegoda and Rosso*, 2008).

The cumulative distribution function (CDF) for the GEVD of maxima is given by

$$F_{GEVD}(x) = \begin{cases} \exp\left\{-\left[1+\xi(x-\mu)/\psi\right]^{-1/\xi}\right\} & \xi \neq 0 \\ \exp\left\{-\exp\left\{1+\xi(x-\mu)/\psi\right\}\right\} & \xi = 0 \end{cases} \tag{1.1}$$

where $1+\xi(x-\mu)/\psi > 0$; $-\infty < \mu < \infty$ is the location parameter, $\psi > 0$ is the scale parameter and ξ is the shape parameter. When $\xi = 0$, the GEVD reduces to a Gumbel distribution, and the variable is unbounded.

The CDF for the GPD of the values that exceed a threshold is given by

$$F_{GPD}(x) = \begin{cases} 1-\left[1+\xi(x-u)/\sigma_u\right]^{-1/\xi} & \xi \neq 0 \\ 1-\exp\left\{-(x-u)/\sigma_u\right\} & \xi = 0 \end{cases} \tag{1.2}$$

where u is the position parameter, such that $x > u$, ξ is the shape parameter and $\sigma_u > 0$ is the scale parameter. When $\xi = 0$, the GPD reduces to an exponential distribution. If $\xi > 0$, the distribution has no upper limit, whereas it has an upper bound of $u - \sigma_u/\xi$ if $\xi < 0$.

Both distributions share the same shape parameter, ξ, while the scale parameters are related by

$$\sigma_u = \psi + \xi(u - \mu) \tag{1.3}$$

These results establish the theoretical foundation for the two most widely accepted methods for modeling the extremes of several geophysical variables: the annual maxima (AM) method and the POT method. When the entire time series of data is available, the use of the AM method is associated with a significant loss of information concerning extreme events. The POT method, on the other hand, uses the data more efficiently by considering more than one sample per year. In this sense, the POT method is preferred over the AM method when an entire time series is available.

1.3.3 The POT method

Given a series of i.i.d. data, the values that exceed a sufficiently high threshold follow a GPD (*Pickands*, 1975). However, when these values show a tendency to form clusters (i.e., for storm events), as is the case for several geophysical variables, there are two ways of dealing with this problem: (i) by declustering the data and (ii) by accounting for the dependence of the series. The first framework is the most widely adopted (*Coles*, 2001) and includes the POT method. In contrast, the second framework may require the use of more sophisticated statistical models and will not be discussed here.

The POT method is considered to be the declustering method most commonly used by hydrologists and coastal engineers. References to it can be found in *Davison and Smith* (1990), although, as pointed out by *Coles* (2001), the general idea is much older. Given a certain threshold, the exceedance values that are separated by less than a given minimum time span are assumed to form a cluster. Each cluster is assumed to be generated by the same extreme event. For every cluster defined in this way, the maximum recorded value is taken. This leads to the construction of a POT series of independent observations. It is clear that the application of the POT method requires a previously defined threshold as well as a minimum time between threshold exceedance events that ensures the independence of the POT series. The study of alternative declustering methods is beyond the scope of this paper. Such information can be found in *Ferro and Segers* (2003) and references therein. This paper only deals with the application of the most common POT declustering method and provides an alternative model for threshold selection and uncertainty quantification. The minimum time between extreme events is assumed to be a given parameter; therefore, no analysis is performed for it.

1.3.4 Threshold identification

A threshold is an essential parameter for the GPD to be used, whether or not the POT method is used. The threshold is, in fact, a distribution parameter (the position parameter of the GPD), and the threshold can be estimated once a data series is defined. However, the problem is that a pre-defined threshold is required for a data series. This implies that threshold estimation is not straightforward, and generally, this estimation cannot be accomplished with the same methods used for estimating the rest of the parameters of a GPD.

Despite the importance of the threshold in the analysis of extreme events, the existing methods for threshold identification are based totally or partially on subjective judgment. Moreover, most methods do not account for any uncertainty in the threshold estimation. As a consequence, the confidence bounds of extreme values do not include the uncertainty associated with the threshold.

Two common ways of choosing a threshold are based on expert judgment. One way is to select a fixed quantile corresponding to a sufficiently high non-exceedance probability, usually 95%, 99% or 99.5% (see, e.g., *Luceño et al.* (2006) or *Smith* (1987)). The other way is to impose a minimum on the mean number of clusters per year. Neither of these methods allows for the estimation of the uncertainty associated with the threshold. Thus, a sensitivity analysis is frequently performed.

There are other methods that provide some guidance for threshold identification and limit the subjectivity of its selection: the graphical method (GM) and the optimal bias-robust estimation (OBRE) method. The GM is based on the stability of the GPD parameters. The OBRE method is based on the procedure used for parameter estimation. These methods are discussed below. Recently, *Thompson et al.* (2009) proposed a new method that is also based on the stability of the GPD parameters. This method will not be discussed in this work.

Graphical method (GM)

The graphical method is based on the stability of the shape and scale parameters of a GPD (see *Coles*, 2001). Provided that all of the exceedance values of the threshold u_0 follow a GPD with parameters ξ and σ_{u_0}, the exceedance values of any other threshold u, such that $u > u_0$, must follow a GPD with parameters ξ and σ_u that satisfy $\sigma_u = \sigma_{u_0} + \xi(u - u_0)$. Then, the expectation of exceedance, given by $E(X - u|X > u) = \sigma_u/(1+\xi)$, must be a linear function of u for all $u > u_0$.

Consequently, there are two ways of applying the GM. The simplest is to construct the mean residual life plot (MRLP). Given a series of thresholds, the MRLP is the locus of points given by

$$\left\{ \left(u, \frac{1}{n_u} \sum_{i=1}^{n_u} (x_i - u) \right) : u < x_{max} \right\} \tag{1.4}$$

For $u > u_0$, being u_0 the threshold at which the GPD is a good approximation of the data, the MRLP should be approximately linear with u. The other option is to

estimate ξ and σ for several thresholds and plot (u, ξ_u) and $(u, \sigma_u - \xi_u u)$. In this case, the plots should be constant for $u > u_0$.

A major advantage of the GM is that its implementation is straightforward and it reduces the subjectivity associated with threshold selection. However, the GM has a remaining subjective component that requires human judgment and cannot be automated, and it is unable to provide the uncertainty of the threshold estimation.

The OBRE method

Dupuis (1998) proposed performing parameter estimation and selecting the threshold by introducing the optimal bias-robust estimator (OBRE), which is an M-estimator (see *Zea Bermudez and Kotz*, 2010a). This method attributes weights (equal to or less than one) to the observations used in parameter estimations. Observations with a very low weight do not fit the GPD model. *Dupuis* (1998) suggests using this weight as a guide for threshold selection and increasing the threshold until the weight assigned to the observation is sufficiently close to one. The OBRE method also does not provide the uncertainty of the threshold estimation. A disadvantage of this method is that a practicing engineer is required to decide on the weight, a variable with which the engineer may not be familiar.

1.3.5 Mixture models

Mixture models, such as the one used in this work, incorporate both the central and extreme populations into a single model, and thresholds are included as parameters. Different versions of this approach have been recently used by *Frigessi et al.* (2002) for modeling the Danish Fire Loss Data Set, by *Vaz de Melo Mendes and Freitas Lopes* (2004) for economic and environmental indices, by *Behrens et al.* (2004) for economic indices, by *Tancredi et al.* (2006) for daily maximum flow rates, by *Vrac and Naveau* (2007), *Furrer and Katz* (2008) and *Hundecha et al.* (2009) for precipitation and by *Cai et al.* (2007, 2008) for significant wave heights. Some authors refer to the mixture models composed of truncated distributions that do not overlap as hybrid models, while in this paper, we use the term mixture model indiscriminately.

Frigessi et al. (2002) proposed a dynamically weighted mixture model composed by a GPD and another light-tailed distribution and applied it to the Danish Fire Loss Data Set. This model was used by *Vrac and Naveau* (2007) and *Hundecha et al.* (2009) for modeling precipitation data sets while imposing zero as the threshold of the GPD.

Vaz de Melo Mendes and Freitas Lopes (2004) proposed a mixture model that is composed of a normal distribution for the central regime and two GPDs for the minima and maxima regimes, corresponding to the upper and lower tails, respectively. The limits between the central distribution and the tails are defined by the upper and lower thresholds, which are model parameters. For parameter estimations, sets of thresholds are selected. For each threshold, the normal distribution parameters are estimated by using the ML method, and the GPD parameters are estimated by using L-moments statistics. Then, the likelihood function is calculated using the obtained parameters.

Model parameters are specified such as those for which the likelihood function is maximized. The application of the model requires the normalization of the data. This model has been applied to the Mexico index series and to a storm surge series included in *Coles* (2001). In the latter case, the thresholds that were obtained were in concurrence with those obtained with the GM. This methodology does not include a procedure for estimating threshold uncertainties.

Behrens et al. (2004) used a mixture model composed of a gamma distribution for the central regime and a GPD for the maxima regime. This model was applied to the Nasdaq 100 index series. Parameters are estimated by means of Bayesian techniques with Markov chain Monte Carlo methods. For this methodology, the parameter distribution is obtained beforehand. Therefore, the uncertainty of the parameters can be estimated.

Tancredi et al. (2006) proposed a mixture model composed of a series of uniform distributions for the central regime and a GPD for the maxima regime. Parameters and their uncertainties, including thresholds, are estimated by means of Bayesian techniques with Markov chain Monte Carlo methods. This approach was applied to the River Nidd flow data set. The resulting threshold was similar to that obtained by *Davison and Smith* (1990) with the GM. This model is flexible because it uses series of uniform distributions, but its implementation may be difficult.

Cai et al. (2007) used a mixture model composed of a biparametric Weibull distribution for the minima and central regimes and a GPD for the maxima regime. *Cai et al.* (2008) used a combination of an empirical distribution for the minima and central regimes and a GPD for the maxima regime. However, continuity was not assured in the probability density functions (PDF) of these models, and no information was given in regards to their threshold calculations.

Furrer and Katz (2008) used a mixture model composed of a truncated gamma distribution for the central regime and a GPD for the upper tail or maxima regime, similar to that used by *Behrens et al.* (2004). The threshold that defines the limit between the central and tail distributions is defined a priori. To ensure continuity in the PDF, they first calculated the parameters of the GPD using only the data greater than the threshold, and they then adjusted the scale parameter of the GPD to ensure $f_{gamma}(u) = f_{GPD}(u)$, where u is the threshold.

1.4 Model description

With the exception of the mixture models, all the methods reviewed analyze the central and maxima regimes separately. More specifically, the study of the central regime is based on all available observations, and the study of the maxima regime is based on a subsample of data representative of the extreme conditions. In line with *Vaz de Melo Mendes and Freitas Lopes* (2004); *Behrens et al.* (2004); *Tancredi et al.* (2006); *Cai et al.* (2007); *Furrer and Katz* (2008), here, the use of a parametric mixture model that is valid for the entire range of values of the variable is proposed. It also differentiates the three populations: (i) a central regime for the bulk of the data; (ii) a minima regime

for the lower tail and (iii) a maxima regime for the upper tail. The thresholds that define the limits between these three populations are parameters of the model and are thus estimated in the same way as the other parameters. The model proposed is that of (3.1), where F_c is the distribution function assumed for the central regime, F_m is the distribution function for the minima regime and F_M is the distribution function for the maxima regime.

$$f(x) = \begin{cases} f_m(x)F_c(u_1) & x < u_1 \\ f_c(x) & u_1 \leq x \leq u_2 \\ f_M(x)\left(1 - F_c(u_2)\right) & x > u_2 \end{cases} \tag{1.5a}$$

$$F(x) = \begin{cases} F_m(x)F_c(u_1) & x < u_1 \\ F_c(x) & u_1 \leq x \leq u_2 \\ F_c(u_2) + F_M(x)\left(1 - F_c(u_2)\right) & x > u_2 \end{cases} \tag{1.5b}$$

This model does not consider the time dependency of the data. It is a model for the marginal distribution of the random variable and, as such, cannot be directly used to calculate the quantiles of high-return periods, unless the series is i.i.d. The main advantage of the model is that it calculates the thresholds that indicates the limit between the minima, central and maxima regimes, as well as their uncertainties. One advantage of this model is that this is done automatically and objectively. For the central regime, an LN distribution is used:

$$f_c(x) = \frac{1}{x\sigma_{LN}\sqrt{2\pi}} \exp\left\{-\frac{[\ln(x) - \mu_{LN}]^2}{2\sigma_{LN}^2}\right\} \tag{1.6}$$

where μ_{LN} and $\sigma_{LN} > 0$ are position and scale parameters, respectively. Similarly, a gamma, WB or any other distribution could be used. Both of these, or any other distribution function, are equally valid for the procedure described in the following paragraphs. The minima and maxima GPDs are used for describing the minima and maxima regimes:

$$f_m(x|x < u_1) = \frac{1}{\sigma_1}\left(1 - \frac{\xi_1}{\sigma_1}(x - u_1)\right)^{-\frac{1}{\xi_1}-1} \tag{1.7}$$

$$f_M(x|x > u_2) = \frac{1}{\sigma_2}\left(1 + \frac{\xi_2}{\sigma_2}(x - u_2)\right)^{-\frac{1}{\xi_2}-1} \tag{1.8}$$

where u_1 and u_2 are the upper and lower thresholds of the central regime; σ and ξ ($\xi_1 \neq 0$ and $\xi_2 \neq 0$) are the scale and shape parameters, respectively, such that $\sigma > 0$ and $-\infty < \xi < \infty$. The case of $\xi = 0$, in which the GPD reduces to an exponential distribution with the parameter $\sigma > 0$, is not considered in this analysis. Moreover, for the minima GPD, $u_1 + \sigma_1/\xi_1 \leq x \leq u_1$ if $\xi_1 < 0$, and $x \leq u_1$ if $\xi_1 > 0$. Conversely, for the

maxima GPD, $u_2 \leq x \leq u_2 - \sigma_2/\xi_2$ if $\xi_2 < 0$, and $x \geq u_2$ if $\xi_2 > 0$. The model proposed (3.1) has eight parameters, which are estimated using the ML method. Accordingly, the negative log-likelihood function (NLLF) is minimized. The log-likelihood function (LLF) $l(x|\theta)$ is given by:

$$\begin{aligned} l(x|\theta) &= N_1 \left[\ln\left(F_c(u_1)\right) - \ln\left(\sigma_1\right)\right] \\ &\quad - \left(\frac{1}{\xi_1} + 1\right) \sum_{N_1} \ln\left(1 - \frac{\xi_1}{\sigma_1}(x - u_1)\right) \\ &\quad - N_c \ln\left(\sigma_{LN}\sqrt{2\pi}\right) - N_c \ln\left(x\right) \\ &\quad - \frac{1}{2} \sum_{N_c} \left(\frac{\ln(x) - \mu_{LN}}{\sigma_{LN}}\right)^2 \\ &\quad + N_2 \left[\ln\left(1 - F_c(u_2)\right) - \ln\left(\sigma_2\right)\right] \\ &\quad - \left(\frac{1}{\xi_2} + 1\right) \sum_{N2} \ln\left(1 + \frac{\sigma_2}{\xi_2}(x - u_2)\right) \end{aligned} \tag{1.9}$$

where θ is the vector of parameters to be estimated; N_1 is the number of data records such that $x < u_1$; N_c is the number of data records such that $u_1 \leq x \leq u_2$ and N_2 is number of data records such that $x > u_2$.

1.4.1 Continuity and physical bounds

Model (3.1) as well as the models proposed by *Vaz de Melo Mendes and Freitas Lopes* (2004), *Behrens et al.* (2004), *Tancredi et al.* (2006) and *Cai et al.* (2007) have discontinuities in the PDF at the threshold values. However, experience does not indicate that such discontinuities exist for geophysical variables. Accordingly, continuity is imposed on the PDF of the model. Continuity in (3.1) allows for the expression of the scale parameters of their GPDs as a function of the corresponding location parameters. The number of parameters of the mixture distribution is thus reduced in two. As the variables studied in the following section (mean daily flow and daily precipitation) must be positive, the condition $x \geq 0$ is imposed on the model. This permits the expression of the scale parameter of the minima GPD as a function of the shape and position (threshold) parameters: $\sigma_1 = -\xi_1 u_1$.

Using both conditions, the following relationships are obtained:

$$\xi_1 = -\frac{F_c(u_1)}{u_1 f_c(u_1)} \quad ; \quad \sigma_2 = \frac{1 - F_c(u_2)}{f_c(u_2)} \tag{1.10}$$

As a result, the number of parameters of the model is reduced to five (u_1, u_2, μ_{LN}, σ_{LN}, ξ_2). This simplifies the model, and as previously specified physical information has been used, its uncertainty is also reduced.

1.4.2 Parameter estimation

There are several methods that estimate the distribution parameters based on the observed data. One of the most commonly applied, given its flexibility and properties, is the ML method (see *Coles*, 2001, chap. 2). This is the method used for this research. For an in-depth discussion of methods for GPD parameter estimation, see *Zea Bermudez and Kotz* (2010b,a). A discussion of the methods used to minimize the NLLF and the precautions that should be taken in this process can be found in the appendices.

1.4.3 Confidence intervals

The confidence intervals of the parameters are estimated by using the covariance matrix $\widehat{Cov}(\widehat{\theta})$. This is calculated as the inverse of the information matrix, which is obtained numerically. The diagonal elements of the covariance matrix represent the variance of the parameters. When a normal distribution is assumed, these variances can be used to obtain the confidence intervals of the model parameters at the $(1-\alpha)$ level: $\theta_j \in (\widehat{\theta}_j \pm z_{\alpha/2}\widehat{\sigma}_{\widehat{\theta}_j})$ for $j = 1, \ldots, 5$ (see *Coles*, 2001; *Castillo et al.*, 2005).

Quantile confidence intervals are calculated using the delta method. Given the quantile x_P, corresponding to the non-exceedance probability P, the gradient of x_P is numerically obtained as $\nabla_{\theta_i} x_P = \partial x_P / \partial \theta_i$. The variance of the quantile is estimated by $\widehat{\sigma}^2_{\widehat{x}_P} = \nabla^T_{\theta} x_P \widehat{Cov}(\widehat{\theta}) \nabla_{\theta} x_P$. Then, assuming a normal distribution, the confidence limits at the $(1-\alpha)$ level are: $x_P \in (\widehat{x}_P \pm z_{\alpha/2}\widehat{\sigma}_{\widehat{x}_P})$.

The expression for x_P is:

$$x_P(P) = \begin{cases} x_m(P) & P < F_c(u_1) \\ x_c(P) & F_c(u_1) \leq P \leq F_c(u_2) \\ x_M(P) & P > F_c(u_2) \end{cases} \tag{1.11}$$

where

$$\begin{aligned} x_m(P) &= u_1 \left(\frac{P}{F_c(u_1)} \right)^{\frac{F_c(u_1)}{u_1 f_c(u_1)}} \\ x_c(P) &= \exp\left[\mu_{LN} + \sigma_{LN}\sqrt{2} erf^{-1}(2P-1) \right] \\ x_M(P) &= u_2 \\ &\quad - \frac{1-F_c(u_2)}{\xi_2 f_c(u_2)} \left[1 - \left(\frac{1-F_c(u_2)}{1-P} \right)^{\xi_2} \right] \end{aligned}$$

1.4.4 Estimations of extreme events using the proposed model

When series are i.i.d., the quantiles for any return period can be obtained directly from the model (3.1). However, geophysical variables are rarely i.i.d. because most of the time, high values tend to group together to form clusters (due to storm events in the

case of daily flow and precipitation data). The quantiles for high-return periods for non i.i.d. series can be estimated using the POT method. This section explains the application of the POT method using the threshold obtained from the model (3.1).

Assuming that the minimum time between threshold exceedance events that assures the independence of POT observations is a known parameter, the POT method can be applied by using the upper threshold obtained with the model (3.1). In this way, an i.i.d. POT series is obtained. A GPD model can be fit to this series by using the ML method, and the covariance matrix $\widehat{Cov}_{\xi\sigma}$ can be obtained for the model parameters ξ and σ. When applying the POT method, the mean number of clusters (storms) per years ν is also obtained. Estimates of ν, its variance and its covariance relative to ξ and σ can be determined assuming that storm occurrences is a Poisson process and including the parameter ν in the likelihood function (see, e.g., *Méndez et al.*, 2006, eq. 5).

$$L(\nu, \xi, \sigma) = \frac{(\nu T)^N e^{-\nu T}}{N!} \prod_{i=1}^{N} f_{GPD}(x_i - u|\xi, \sigma) \tag{1.12}$$

However, the same results can be obtained by estimating ν as $\nu = N/T$, where N is the number of observed clusters and T is the number of recorded years, and by estimating the variance of ν by calculating the variance of the distribution as $\widehat{\sigma}^2_\nu = \widehat{\sigma}^2_{Poisson}/N = \nu/T = \nu^2/N$. From (1.12), it can be deduced that $\partial l/\partial\nu\partial\lambda = \partial l/\partial\nu\partial\kappa = 0$. It can thus be concluded that the covariances for (ξ, ν) and (σ, ν) are zero. Then, the covariance matrix for (ν, ξ, σ) can be expressed as

$$\widehat{Cov}(\nu, \xi, \sigma) = \begin{bmatrix} \widehat{\sigma}^2_\nu & 0 & 0 \\ 0 & \widehat{Cov}_{\xi\sigma}(1,1) & \widehat{Cov}_{\xi\sigma}(1,2) \\ 0 & \widehat{Cov}_{\xi\sigma}(2,1) & \widehat{Cov}_{\xi\sigma}(2,2) \end{bmatrix} \tag{1.13}$$

Based on the above expression, in order to include the uncertainty associated with the threshold (which is another model parameter) in the calculation of quantiles of a high-return period, the covariance matrix can be approximated as follows:

$$\widehat{Cov}(\widehat{\theta}) = \begin{bmatrix} \widehat{\sigma}^2_u & 0 & 0 & 0 \\ 0 & \widehat{\sigma}^2_\nu & 0 & 0 \\ 0 & 0 & \widehat{Cov}_{\xi\sigma}(1,1) & \widehat{Cov}_{\xi\sigma}(1,2) \\ 0 & 0 & \widehat{Cov}_{\xi\sigma}(2,1) & \widehat{Cov}_{\xi\sigma}(2,2) \end{bmatrix} \tag{1.14}$$

where $\widehat{\sigma}^2_u$ is obtained from the model (3.1) and it is assumed that the covariances in reference to u are all zero. It should be highlighted that according to the hypothesis of a Poisson distribution for ν, the lower the upper threshold is, the higher is the possible number of clusters per year and, consequently, the higher is the variance estimated for ν. This variance can only be reduced by increasing the number of years measured.

Using the GPD adjusted for the POT data series and the covariance matrix (2.8), the quantiles for high-return periods (T_r) are estimated by

$$x_{T_r} = u + \frac{\sigma}{\xi}\left((T_r\nu)^{\xi} - 1\right) \tag{1.15}$$

with the following confidence intervals for the $(1-\alpha)$ level

$$x_{T_r} \in (\widehat{x}_{T_r} \pm z_{\alpha/2}\widehat{\sigma}_{\widehat{x}_{T_r}}) \tag{1.16}$$

where $\widehat{\sigma}^2_{\widehat{x}_{T_r}} = \nabla^T_{\theta} x_{T_r} \widehat{Cov}(\widehat{\theta}) \nabla_{\theta} x_{T_r}$, being

$$\begin{aligned} \nabla^T_{\theta} x_{T_r} &= \left[\frac{\partial x_{T_r}}{\partial u}, \frac{\partial x_{T_r}}{\partial \nu}, \frac{\partial x_{T_r}}{\partial \xi}, \frac{\partial x_{T_r}}{\partial \sigma}\right] \\ &= \left[1, \frac{\sigma}{\nu}(T_r\nu)^{\xi}, \frac{\sigma}{\xi^2}(1-(T_r\nu)^{\xi})\right. \\ &\quad \left. + \frac{\sigma}{\xi}(T_r\nu)^{\xi}\ln(T_r\nu), \frac{1}{\xi}((T_r\nu)^{\xi}-1)\right] \end{aligned} \tag{1.17}$$

1.5 Application

For this section, model (3.1) (hereafter called LNGPD) was used to model four sets of data. For each case, the results obtained by applying the model and by comparing the upper threshold u_2 obtained from the model with that obtained by using the GM (in particular using the MRLP) were analyzed. In addition, we examined the results obtained by using the u_2 threshold for the POT method.

A daily precipitation data series from Fort Collins, USA, a mean daily flow data set from the Thames in Kingston, UK and daily precipitation and mean daily flow data sets from Orgiva, Spain were used. These data series were selected to encompass different situations:

- The Fort Collins and the Thames River at Kingston data series are of very long duration, while the data series of flows in Orgiva is of short duration.
- For the precipitation series (Fort Collins and Orgiva), as for the flow series in Orgiva, there is no physical basis to justify imposing an upper bound on the variables, and therefore, it is expected that the distributions of the maximum values of these (both from the entire series as well as the peaks over the threshold) are heavy-tailed distributions. The Thames River, on the other hand, is regulated to avoid flooding, and for this series, the existence of an upper limit for the maximum values of the variable is justified, except in exceptional cases in which the flood control system has been exceeded.

1.5.1 Daily precipitation at Fort Collins

The daily precipitation series at the station in Fort Collins, CO, obtained from the Colorado Climate Center ($http://ccc.atmos.colostate.edu/dly\_form.html$) was analyzed. This same data set has been used previously in the work of *Katz et al.* (2002) and *Furrer and Katz* (2008).

Table 1.1: Parameters of the LNGPD model fitted to the Fort Collins non-zero daily precipitation data, imposing $u_1 = 0$.

u_2	μ_{LN}	σ_{LN}	ξ_2
9.8	0.68	1.39	0.196
0.15	0.22×10^{-3}	0.12×10^{-3}	0.81×10^{-3}

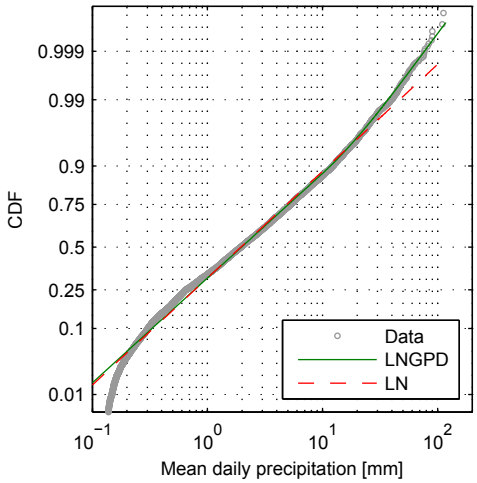


Figure 1.1: Empirical (gray), fitted LN (red) and fitted LNGPD (green) CDFs graphed on a log-normal probability scale. Fort Collins non-zero daily precipitation (mm) data.

The series used covers the period from 1900 to 2010. Of the 40543 available data, only 9036 correspond to non-zero data, equivalent to 22.3% of the total data.

First, the WB, LN and gamma distributions were adjusted to the non-zero data. From these, the distribution with the best fit with respect to the values of the likelihood function and the powers of the χ^2 and Kolmogorov-Smirnov tests (both were rejected for the three distributions tested) was selected. The best fit distribution according to these criteria is the LN; however, the visual evaluation of the LN fit indicates that it has a poor fit in the tails (see Figure 1.1).

To improve the fit, the LNGPD mixture model was used, estimating the parameters by using ML method. First, a value for the u_1 parameter that is less than the minimum value of the non-zero data ($u_1 < \min(x) = 0.1\,mm$) was obtained. Therefore, the parameters are again estimated with $u_1 = 0$, thus discarding the GPD of the minimums. The estimated parameters and their variances are listed in Table 1.1.

Figure 1.1 shows the empirical cumulative distribution function (CDF) of the non-zero data and those obtained with the LN and LNGPD models. The LNGPD model fits the data better than the LN, particularly in the upper tail. This is also seen in the Q-Q plot that is presented in Figure 1.2.

The u_2 upper threshold obtained by adjusting the LNGPD model was $9.8\,mm$, with a confidence interval of 90% ($9.2\,mm$, $10.5\,mm$). Expressed in inches, the threshold is $0.39\,in$, which is almost equal to the threshold used by *Katz et al.* (2002) for the POT

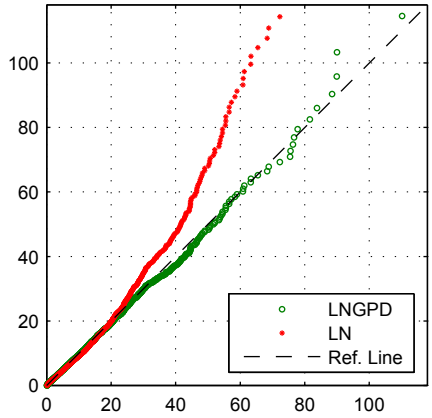


Figure 1.2: QQ plot. Fitted LN (red dots) and fitted LNGPD (green dots) models. Fort Collins non-zero daily precipitation (mm) data.

method ($0.40\,in$).

The MRLP (figure 1.3) shows that there are two trends (indicated as A and B) that correspond to the $u_2 \simeq 10\,mm$ threshold (equivalent to that obtained by adjusting the LNGPD model) and the $u_2 \simeq 40\,mm$ threshold. These trends slope upwards, indicating that the GPD shape parameter adjusted with the two thresholds must be positive.

To verify this, a GPD distribution was adjusted to the data greater than $40\,mm$. The shape parameter obtained was 0.114, with a confidence interval of 90% (-0.10, 0.33), while that obtained with the LNGPD model was 0.196, with a confidence interval of 90% (0.15, 0.24). Although the shape parameter obtained with the $40\,mm$ threshold has the same sign as that obtained with the $9.8\,mm$ threshold (both correspond to a distribution with a heavy tail), the confidence interval of the first is considerably broader than that of the second and does not exclude the possibility that the shape parameter can take on negative values, corresponding to a light tail.

Next, the feasibility of using the upper threshold u_2 obtained by adjusting the LNGPD model for the POT method was studied. To differentiate the analysis of the POT data from that of the entire dataset, the threshold used to construct the POT series is called u, and the shape parameter of the GPD of the maxima adjusted using the POT series is called ξ.

Figure 1.4 presents the MRLP of the POT data. As with all non-zero data, there are two parallel trends, for which there are two possible thresholds for applying the POT method: $u \simeq 10\,mm$ and $u \simeq 40\,mm$.

In principle, based only on the MRLP, one would tend to select the higher threshold ($40\,mm$); nevertheless, with this selection, less than one storm was obtained each year (on average), with which the use of the POT method instead of the AM method is not justified. In order to achieve a number of annual storms greater than one, the threshold obtained by adjusting the LNGPD to all of the data was selected ($9.8\,mm$).

It is interesting to compare the behavior of the ξ parameters of the GPD adjusted

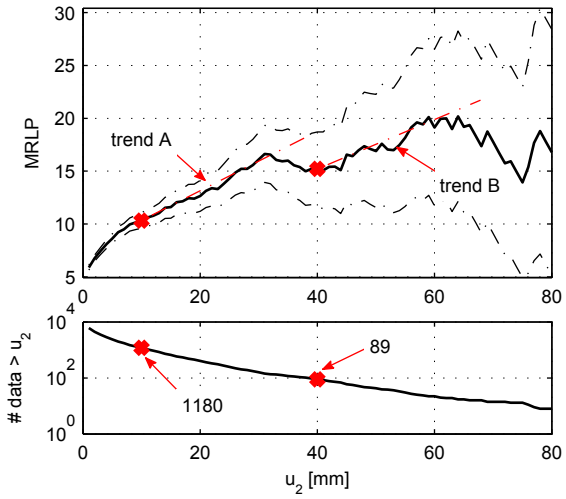


Figure 1.3: Mean Residual Life Plot for Fort Collins daily precipitation (mm).

to the POT data obtained with the different thresholds. When $u = u_2 = 9.8\,mm$, the shape parameter is 0.17, with a confidence interval of 90% (0.11, 0.24), i.e., the GPD has a heavy tail. When the threshold is set to $u = 40\,mm$, the shape parameter is 0.12, with a confidence interval of 90% $(-0.11, 0.35)$, i.e., a heavy tail is also obtained with $u = 40\,mm$, but the variance is such that the 90% confidence interval does not rule out a light tail.

Finally, the effect of including the variance of the threshold when calculating the precipitation amounts of distinct return periods was analyzed. Figure 1.5 presents the POT data series obtained with the threshold of $u = 9.8\,mm$, in addition to the 90% confidence interval obtained using the covariance matrices (1.13) and (2.8), once adjusted the GPD of the maxima with ML. The results show that including the variance of the threshold has an insignificant effect on the confidence intervals, which only increases in significance for values of low-return periods, for which a widening of the confidence intervals is seen.

We conclude that in the case of the non-zero daily rainfall data series for Fort Collins, CO, the LNGPD distribution adequately models the full range of values of rainfall and provides a u_2 upper threshold value that is adequate for applying the POT method.

1.5.2 Mean daily flow at Thames at Kingston

We used the mean daily flow values (gauged daily flow) from the River Thames station Thames at Kingston. Data were obtained from the UK Centre for Ecology & Hydrology ($http://www.ceh.ac.uk/data/nrfa/data/time\_series.html?39001$). The series used covered the period from $1883 - 2009$, with a total of 46386 data. This same series was used by *Eastoe and Tawn* (2010) to study alternatives to the Poisson model for modeling occurrence of extreme events.

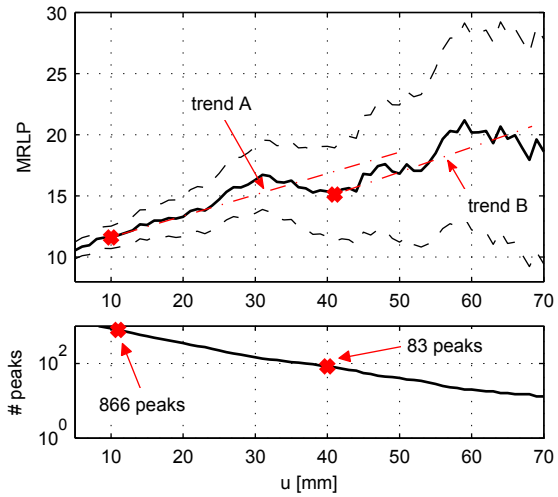


Figure 1.4: Mean Residual Life Plot for Fort Collins daily precipitation (mm) POT series.

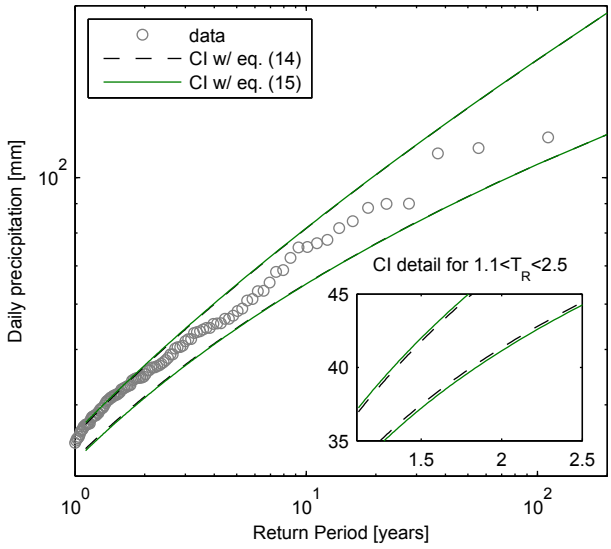


Figure 1.5: Fort Collins daily precipitation (mm) POT series and 90% confidence intervals (CIs) of the GPD fitted using $u = 9.8\,mm$.

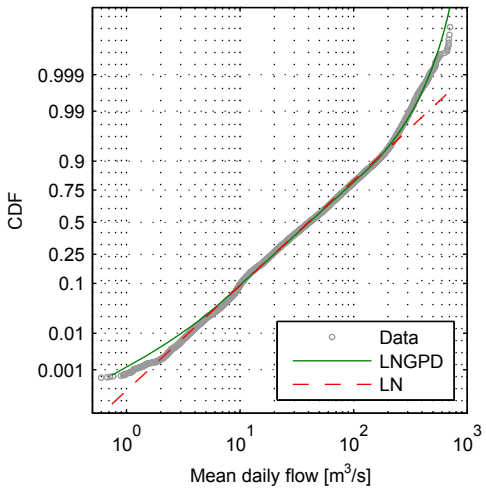


Figure 1.6: Empirical (gray), fitted LN (red) and fitted LNGPD (green) CDFs graphed on a log-normal probability scale. Thames at Kingston mean daily flow (m^3/s).

Table 1.2: Parameters of the LNGPD model and their variances fitted to the Thames at Kingston mean daily flow data.

u_1	u_2	μ_{LN}	σ_{LN}	ξ_2
7.2	124.5	3.71	1.07	-0.092
0.014	1.8	2.5×10^{-5}	1.4×10^{-5}	4.4×10^{-5}

As with the data series from Fort Collins, WB, LN and gamma distributions were fitted and evaluated, finding that the LN distribution provided the best fit, although it is a poor fit in the tails. Then, the LNGPD mixture model was used, leading to a significant improvement in the fit with respect to that obtained with the LN model (see figure 1.6). Table 1.2 lists the estimated values and the variances of the parameters of the LNGPD model.

Figure 1.7 shows the MRLP of the series. The threshold of the GPD maxima that is identified by the MRLP is about $300\,m^3/s$. On the other hand, the upper threshold u_2 of the LNGPD model is $124\,m^3/s$, with a 90% confidence interval of (122, 127). These thresholds correspond to the two different trends indicated in the MRLP (figure 1.7): the A trend corresponds to the $124\,m^3/s$ threshold and is associated with a light tail, while the B trend corresponds to the $300\,m^3/s$ threshold and is associated with a heavy tail. Although *Katz et al.* (2002) point out that, in general, hydrological variables have heavy tails, it is possible that this is not true in a basin subject to a flood control system such as this one. Figure 1.6 indicates that the LNGPD (that in this case has a light upper tail) provides a satisfactory fit to the data of the upper tail.

Assuming that this is a closely regulated river, for which the level of regulation has varied over time, an analysis on the extremes of the flow series is not recommended for engineering applications (it would be interesting to determine if the over-dispersion of the distribution of extreme floods identified by *Eastoe and Tawn* (2010) has its origins in

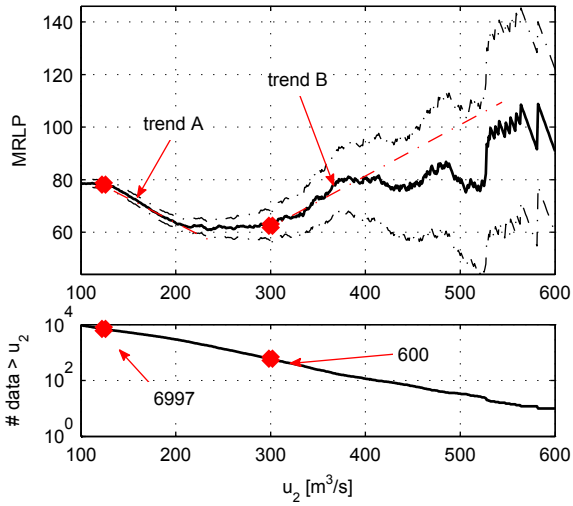


Figure 1.7: Mean Residual Life Plot for the Thames at Kingston mean daily flow (m^3/s).

the inhomogeneities of the series that result from the time varying flood control system; this, however, is outside of the objectives of this work). Despite this, we proceeded to study the behavior of the POT series because it is instructive to explore the capabilities of the LNGPD model, and it complements the analysis on whether the upper tail of this series is heavy or not.

Figure 1.8 presents the MRLP of the POT series. According to this MRLP, the upper threshold appropriate for applying the POT method is $u = 230\,m^3/s$, higher than that obtained from the LNGPD mixture model, $u = u_2 = 124\,m^3/s$. It is also clear from the graph that the adjusted GPD with threshold $u = 230\,m^3/s$ will have a heavy tail, while the adjusted GPD with $u = 124\,m^3/s$ will have a light tail.

A GPD was adjusted to each of the POT series built with the two identified thresholds ($124\,m^3/s$ and $230\,m^3/s$). Figure 1.9 shows the 90% confidence intervals for the quantiles that were obtained with these GPDs as well as the respective POT series. The fit obtained using the $124\,m^3/s$ threshold is good, even for the exceptional values of the mean daily flow: for high-return periods, the GPD tracks the data trend well, and only the data for the floods of 1894 and 1947 fall outside of the confidence interval. In contrast, the confidence interval of the GPD adjusted with the $230\,m^3/s$ threshold includes the values that correspond to the floods of 1894 and 1947. Interestingly, the lower limits of the confidence interval coincide for both thresholds.

To complement the analysis above, we studied the behavior of the MRLP and the GPD when the two years of exceptional floods (1894 and 1947) were removed from the data series, under the assumption that these floods are not from the same population as the rest of the data because they correspond to situations where the flood control system failed in its function. Figure 1.10 shows the MRLP obtained using all of the data and that obtained when removing these two years. The MRLP obtained for all of the years is included within the 95% confidence interval of the MRLP obtained when

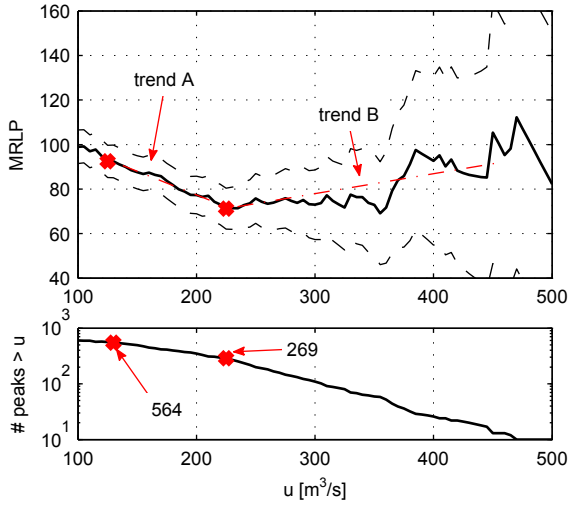


Figure 1.8: Mean Residual Life Plot for the Thames at Kingston mean daily flow (m^3/s) POT series.

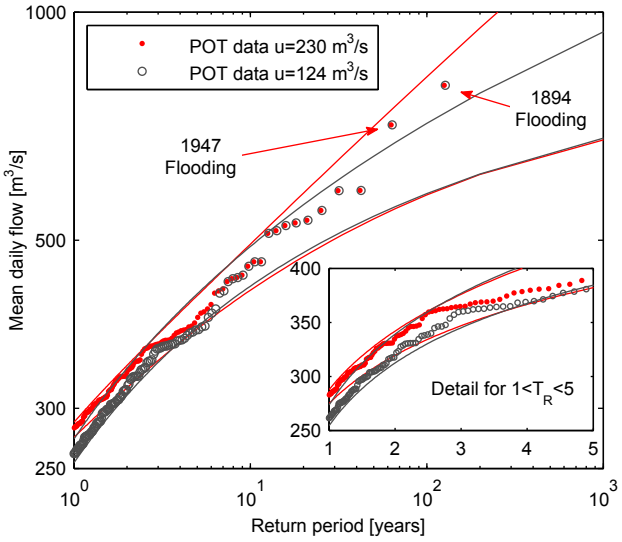


Figure 1.9: Thames at Kingston mean daily flow (m^3/s) POT series and 90% confidence intervals (CIs) estimated using thresholds of $u = u_2 = 124\,m^3/s$ (gray) and $u = 230\,m^3/s$.

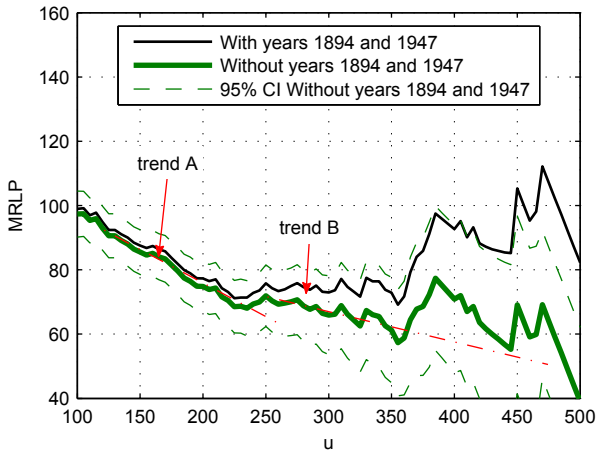


Figure 1.10: Mean Residual Life Plot for the Thames at Kingston mean daily flow (m^3/s) POT series.

excluding the years of 1984 and 1947. The new MRLP still has two trends, but in this case, both are qualitatively similar and both correspond to a GPD with a light tail. This analysis indicates that with regards to the selection of the threshold required to apply the POT method, the MRLP is more sensitive to the presence of outliers than the LNGPD model.

In summary, it appears that the LNGPD mixture model improves the data fit for the whole range of values for the variable of mean daily flow recorded at Thames at Kingston. At the same time, the model allows for the identification of the threshold necessary in order to apply the POT method, being less sensitive to outliers than the GM. In this case, the GPD obtained for the POT series has a light tail, which is consistent with the degree of regulation to which the river is subject, and correctly represents the more extreme data, except for the two exceptional floods of 1894 and 1947.

1.5.3 Orgiva streamflow and precipitation series

Lastly, we analyzed two shorter duration data sets corresponding to a Mediterranean basin located on the Iberian Peninsula. Daily precipitation and mean daily flow data series from Orgiva, Spain (coordinates $36° 54' N$, $3° 25' W$) were used. A description of the characteristics of this basin can be found in *Herrero et al.* (2009), *Millares et al.* (2009) and *Mans et al.* (2011).

(a) Precipitation

A series of 16948 data points of daily precipitation from $1961 - 2008$, albeit with some timeframes where data were missing, were used.

Rainfall in Orgiva has a bimodal distribution (see figure 1.11), with a first peak corresponding to very low-intensity daily precipitation values ($0.1\,mm$). An approximation for this type of distribution proposed by *Carreau et al.* (2009) consists of using a "Hybrid Pareto" mixture model, composed of several Normal-Pareto mixture distributions

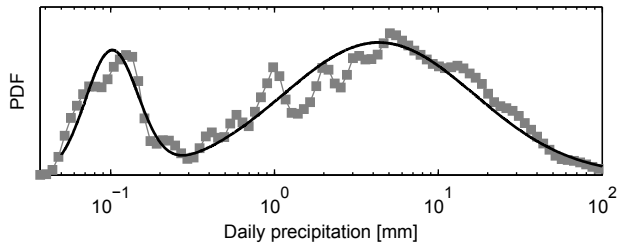


Figure 1.11: PDF of precipitation data at Órgivade: Empirical (squares) and smoothed empirical (line).

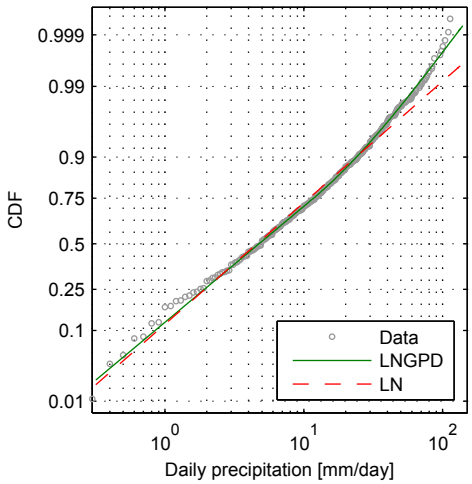


Figure 1.12: Empirical (gray), fitted LN (red) and fitted LNGPD (green) CDFs on a log-normal probability scale. Órgiva daily precipitation (mm/day).

similar to model (3.1). However, here, we attempted to demonstrate the ability of the LNGPD mixture model to explore and model this series of hydrological variables while avoiding the use of more complicated model such as that presented by *Carreau et al.* (2009). As a result, records below $0.3\,mm/day$ were discarded for this analysis.

Among the WB, LN and gamma distributions adjusted to the data greater than $0.3\,mm/day$, the LN distribution provides the best fit. Nevertheless, this distribution does not have a good fit in the upper tail for values greater than $30\,mm/day$ (see figure 1.12). Therefore, the LNGPD distribution was adjusted to obtain a significant improvement in fit (see figures 1.12 and 1.13). The parameters of the adjusted LNGPD distribution are presented in Table 1.3. The estimated value of the upper threshold is $u_2 = 11.6\,mm/day$, while that obtained using the MRLP (not shown) is approximately $10\,mm/day$.

The results from using a threshold of $u = u_2 = 11.6\,mm/day$ to apply the POT method were analyzed. Figure 1.14 shows the MRLP for the POT series, from which a threshold of $u = 26\,mm/day$ was selected. The shape parameter for the GPD obtained by using $u = 26\,mm/day$ is 0.2, while that obtained by using $u = u_2 = 11.6\,mm/day$ is 0.09. Both are positive, i.e., with heavy tails. However, the use of the lower threshold

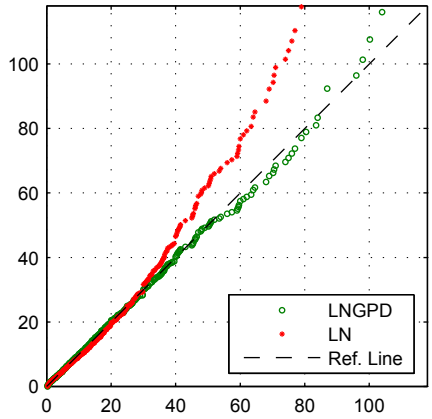


Figure 1.13: QQ plot of daily precipitation values (mm/day) in Órgiva. LN (red) and LNGPD (green).

Table 1.3: Parameters of the LNGPD model fitted to Órgiva daily precipitation values imposing $u_1 = 0$ and their standard deviations.

u_2	μ_{LN}	σ_{LN}	ξ_2
11.57	1.58	1.32	0.18
0.58	0.028	0.023	0.041

yields a larger number of events, which again, as in the previous cases, reduces the uncertainty in the extrapolation when high-return period values are calculated (see figure 1.15).

(b) Streamflow

For the mean daily streamflow data series, 5546 data points were used, corresponding to the period from 1991-2009, with several periods of missing data. In total, there were about 15 years of data.

There are two flow populations in the Orgiva basin: one from snowmelt and the other from surface runoff. However, the two populations overlap, generally following a LN distribution, although this distribution fits the tails poorly (see Figure 1.16).

By adjusting the LNGPD model to the data, the lower threshold is determined to be zero. The parameter values of the LNGPD model estimated by ML are listed in Table 1.4. Figures 1.16 and 1.17 present the CDF and QQ graphs. The fit obtained with the LNGPD model is significantly improved compared to that obtained with the LN distribution. As with the precipitation data, the upper threshold obtained with the LNGPD model roughly coincides with the one that would be selected from the MRLP (see Figure1.18).

As for applying the POT method, the MRLP (figure 1.19) indicates that the value of the threshold appropriate for applying the GPD distribution to the POT series is about $20\,m^3/s$, resulting in a light tail. However, this is hardly consistent with the precipitation data, for which both the MRLP as well as the LNGPD model yielded thresholds that resulted in heavy tails. On the other hand, the MRLP of the POT data

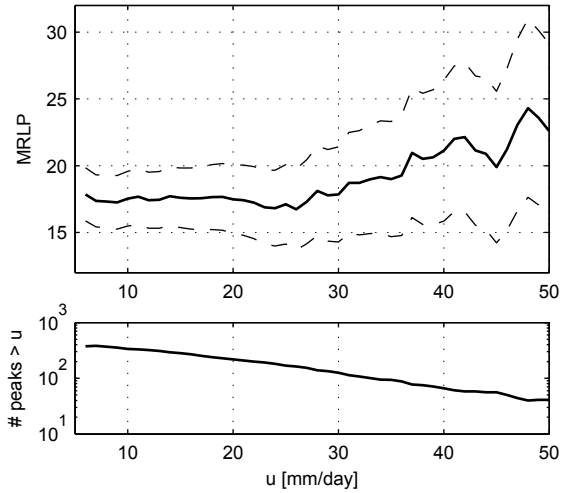


Figure 1.14: MRLP of the daily precipitation POT series for Órgiva.

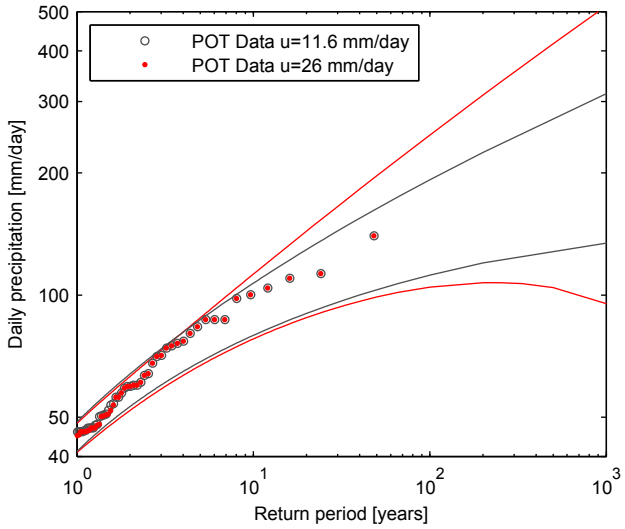


Figure 1.15: Órgiva mean daily flow (m^3/s) POT series (dots) and 90% confidence intervals (CIs) of the GPDs fitted using $u = u_2 = 11.6\,mm/day$ (gray lines) and $u = 26\,mm/day$ (red lines) thresholds.

Table 1.4: Parameters of the LNGPD model fitted to the Orgiva mean daily streamflow (m^3/s) data when imposing $u_1 = 0$ and their standard deviations.

u_2	μ_{LN}	σ_{LN}	ξ_2
7.94	0.591	1.124	0.075
0.38	0.015	0.011	0.036

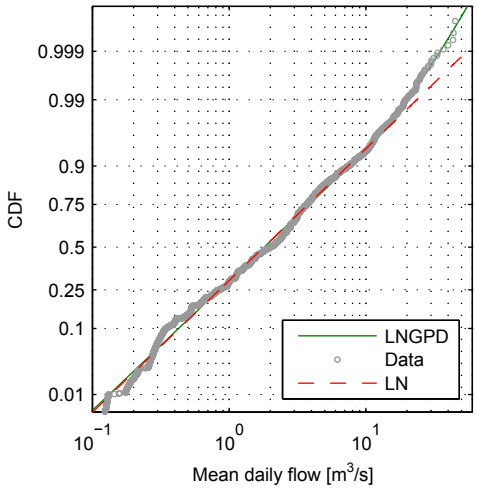


Figure 1.16: Empirical (gray), fitted LN (red) and fitted LNGPD (green) CDFs on a log-normal probability scale. Órgiva mean daily streamflow (m^3/s).

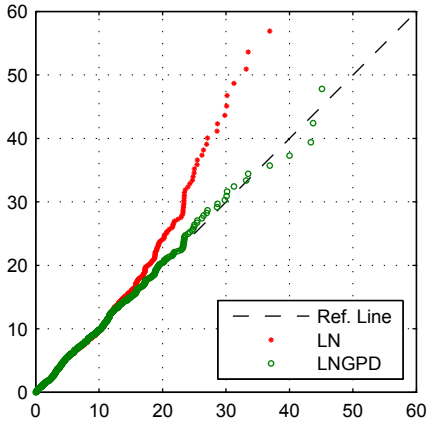


Figure 1.17: QQ plot of the mean daily streamflow (m^3/s) in Órgiva. LN (red) and LNGPD (green).

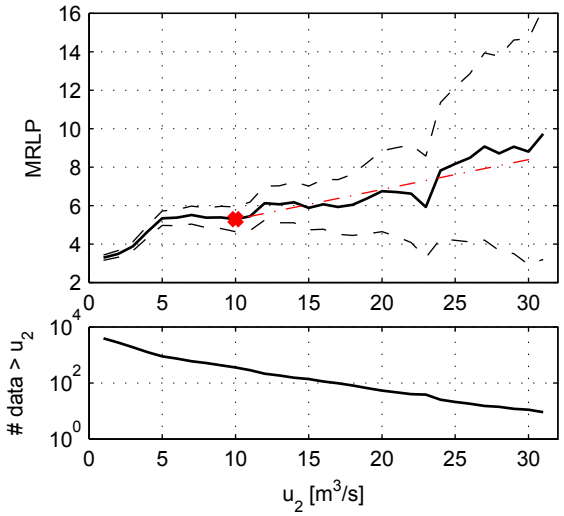


Figure 1.18: MRLP for Órgiva mean daily streamflow (m^3/s) data series.

indicates that the threshold of $u = u_2 = 7.94\,m^3/s$ obtained using the LNGPD model is also valid given that the A trend (figure 1.19) remains within the confidence intervals of the MRLP. With this threshold, a GPD with a heavy tail is obtained, consistent with that observed in the precipitation series.

In summary, when applying the LNGPD model to the series of daily precipitation and mean daily streamflow data recorded in Orgiva, one finds that it provides a good fit for the full range of values of both variables, except for values less than 0.3, mm/day in the precipitation series, and that in both cases a suitable upper threshold to apply the POT method can be identified.

1.6 Conclusions

This paper explored the use of a mixture model (LNGPD) for the marginal distribution of hydrological variables. This distribution comprises a truncated central distribution that is representative of the central regime, which was the LN distribution for the cases analyzed, and two GPDs for the upper and lower tails, to represent the maxima and minima regimes respectively.

The LNGPD model is able to work over the entire range of values of some significant hydrological variables, such as precipitation and streamflow, regarding the data records as coming from three different populations. The thresholds are model parameters and are estimated by ML. Consequently, the threshold calculation is automatic and objective, does not require the predefinition of any parameter, and yields the minima, central and maxima regimes of the hydrological variables by determining the thresholds u_1 and u_2, which indicate the transition points between regimes.

A significant advantage of this model and methodology over existing ones is that the calculated u_2 is a convenient threshold value that is required for the calculation

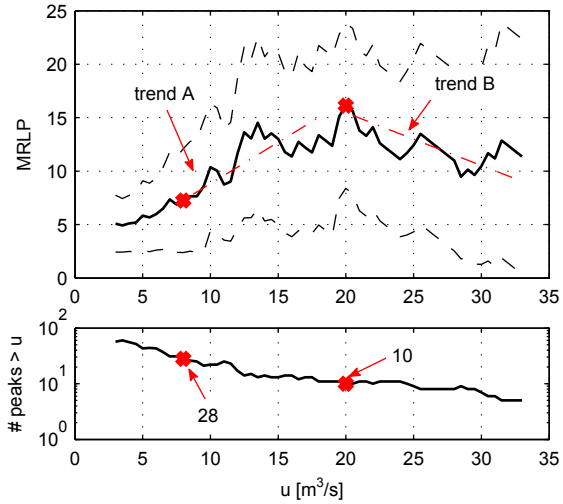


Figure 1.19: MRLP of the mean daily streamflow POT series in Órgiva.

of high-return period quantiles by applying the POT method. For this purpose, a simple methodology has been devised for including threshold uncertainties in quantile calculations.

The same conclusions apply to the minima regime. However, when the u_1 threshold value is less than the lowest value in the series, including the GPD of the minima does not improve the fit of the lower tail.

The proposed mixture model was tested against four series of hydrological data: two of mean daily flow, the Thames at Kingston (UK) and the Guadalfeo River at Orgiva (Spain), and two of daily precipitation, Fort Collins (CO, USA) and Orgiva (Spain). In all cases, the LNGPD mixture model improved the fit of the data series relative to the fit obtained with the LN distribution; in particular, it provided a good fit in the upper tail.

In the four cases studied, the u_2 thresholds obtained from the LNGPD models were suitable to apply the POT method. Therefore, when using the u_2 threshold value for fiiting the GPD to the POT series, a heavy-tailed GPD was obtained for the precipitation series and for the flow series from Orgiva, as expected, and a light tail was obtained for the flow series for the Thames River at Kingston.

In all four cases, a u_2 value was obtained that was less than the threshold value obtained by using the MRLP. From this, we conclude that the two significant relevant differences are as follows:

(a) The confidence intervals of the quantiles of high-return periods calculated by applying u_2 are tighter than those obtained by using the thresholds given by MRLPs.

(b) u_2 threshold of the LNGPD model results in more than one peak per year. On the other hand, in some of the analyzed cases, the threshold obtained by the MRLP results in less than one peak per year.

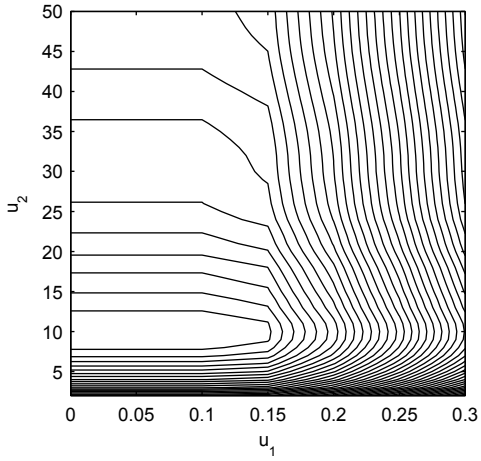


Figure 1.20: Log-likelihood function (LLF) as a function of the u_1 and u_2 thresholds for the Fort Collins data series.

In addition, it was determined that including the uncertainty of the threshold in the calculation of the confidence intervals of the high-return period quantiles does not have a significant impact on these confidence intervals. The only observable effect is a small broadening of the confidence intervals for low return periods.

1.A Appendixes

1.A.1 Uniqueness of the solution

The proposed LNGPD model has 5 parameters. This is a greater number of parameters than that which is commonly estimated by means of ML during the application of a parametric model (e.g., log-normal and gamma distributions have 2 parameters, and the generalized extreme value distribution has 3 parameters).

Here, the parameter-fitting method is analyzed to find whether it effectively maximizes the likelihood function or not. This is done as follows:

(a) a set of u_1 and u_2 thresholds ($u_1 < u_2$) that covers the entire range of values that the variable assumes is defined.

(b) for each pair of values (u_1, u_2), the other 3 parameters (μ_{LN}, σ_{LN}, ξ_2) of the model are estimated by ML.

(c) the iso-probability curves are built in the $u_1 - u_2$ plane, and the(u_1, u_2) point of ML is identified.

Figures 1.20 and 1.21 present the iso-probability curves for the Fort Collins and the Thames at Kingston data series, respectively. Note that both surfaces have a unique maximum that corresponds to the values of the thresholds identified in section 1.5.

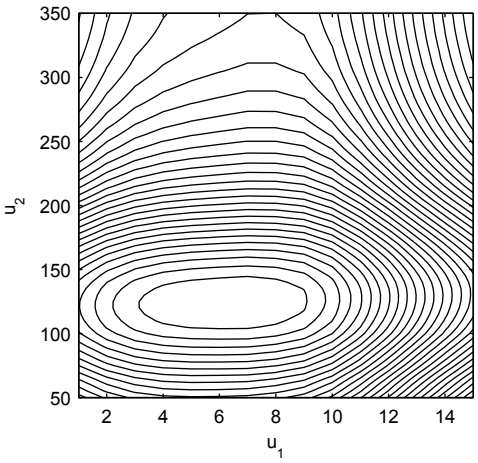


Figure 1.21: Log-likelihood function (LLF) as a function of the u_1 and u_2 thresholds for the Thames at Kingston data series.

1.A.2 Optimization method

The NLLF is minimized using the BFGS method (quasi-Newton method, see e.g. *Nocedal and Wright* (2006) chap. 6) implemented in the MATLAB ® optimization toolbox.

Commonly measured or hindcast data are truncated with a given precision. This produces data that clusters at specific values. It was observed that this led to relative minimums and maximums in (1.9), which may lead to problems when using optimization algorithms for minimizing the NLLF. There are two alternatives for solving this problem. The first is the application of more complex optimization procedures (e.g., global optimization procedures). The second is the uniform distribution of data in the corresponding intervals. However, the second solution is preferable for two reasons: (a) it is easier to implement and (b) because confidence intervals depend on the curvature of the LLF at the optimum point through the information matrix, it is more realistic to use uniformly distributed data, as explained below.

Figure 1.22 shows the LLF for the Fort Collins daily precipitation series. The LLF was estimated for $u_1 = 0$ and for u_2 varying between $9\,mm$ and $11.5\,mm$. The original values of the variable were truncated with a precision of $0.1\,in$. Figure 1.23 shows the LLF for the Thames at Kingston data as a function of u_1 (left) and u_2 (right), obtained with the original data (top) and with the distributed data (bottom). As shown, the use of distributed data results in an important smoothing of the likelihood function. This avoids the singular values observed when using the original data. Furthermore, it is clear that the values obtained when using the original data produce a fictitious curvature that may influence the calculation of the confidence intervals.

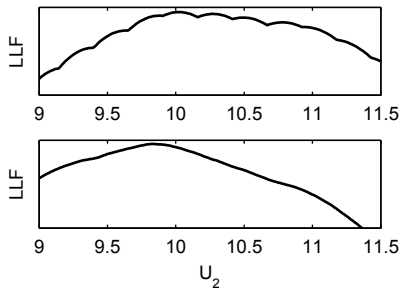


Figure 1.22: Log-likelihood function (LLF) as a function of the u_2 threshold, for the original data (top) and for the uniformly distributed data (left). Fort Collins daily precipitation data.

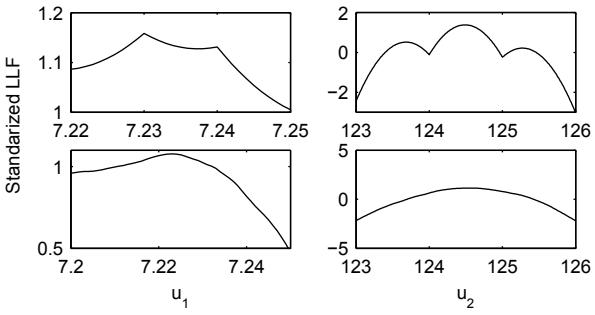


Figure 1.23: Log-likelihood function (LLF) as a function of the u_1 and u_2 thresholds, for the original data (top) and for the uniformly distributed data (left). Thames at Kingston mean daily flow data.

Acknowledgments

This research was funded by the Spanish Ministry of Education through its postgraduate fellowship program, grant AP2009-03235. Partial funding was also received from the Spanish Ministry of Science and Innovation (research project CTM2009-10520), the Spanish Ministry of Public Works (projects CIT-460000-2009-21 and 53/08 –orden FOM/3864/2008–) and the Andalusian Regional Government (research project P09-TEP-4630).

References

Behrens, C. N., H. F. Lopes, and D. Gamerman (2004), Bayesian analysis of extreme events with threshold estimation, *Statistical Modelling*, *4*, 227–244.

Cai, Y., B. Gouldby, P. Dunning, and P. Hawkes (2007), A simulation method for flood risk variables, in *2nd Institute of Mathematics and its Applications International Conference on Flood Risk Assessment, 4th September 2007 , University of Plymouth, England.*

Cai, Y., B. Gouldby, P. Hawkes, and P. Dunning (2008), Statistical simulation of flood variables: incorporating short-term sequencing, *Journal of Flood Risk Management*, *1*, 3–12.

Carreau, J., P. Naveau, and E. Sauquet (2009), A statistical rainfall-runoff mixture model with heavy-tailed components, *Water Resources Research*, *45*(10), 1–11, doi: 10.1029/2009WR007880.

Castillo, E., A. S. Hadi, N. Balakrishnan, and J. M. Sarabia (2005), *Extreme Value and Related Models with Applications in Engineering and Science*, 362 pp., Wiley Seires in Probability and Statistics, John Wiley & Sons, Inc.

Chow, V. T. (1988), *Applied Hydrology*, 585 pp., McGraw-Hill Publishing Company.

Coles, S. (2001), *An Introduction to Statistical Modeling of Extreme Values*, 208 pp., Springer Series in Statistics, Springer, Berlin.

Davison, A., and R. Smith (1990), Models for exceedances over high thresholds, *Journal of the Royal Society Series B*, *52*(3), 393–442.

Dupuis, D. (1998), Exceedances over high thresholds: A guide to threshold selection, *Extremes*, *1*(3), 251–261.

Eastoe, E. F., and J. A. Tawn (2010), Statistical models for overdispersion in the frequency of peaks over threshold data for flow aeries, *Water Resources Research*, *46*, doi:10.1029/2009WR007757.

Evin, G., J. Merleau, and L. Perreault (2011), Two-component mixtures of normal , gamma , and Gumbel distributions for hydrological applications, *Water Resources Research*, *47*, doi:10.1029/2010WR010266.

Ferro, C. A. T., and J. Segers (2003), Inference for clusters of extreme values, *Journal of the Royal Statistical Society: Series B*, *65*(2), 545–556.

Frigessi, A., O. Haug, and H. Rue (2002), A dynamic mixture model for unsupervised tail estimation without threshold selection, *Extremes*, *5*, 219–235.

Furrer, E. M., and R. W. Katz (2008), Improving the simulation of extreme precipitation events by stochastic weather generators, *Water Resources Research*, *44*(12), 1–13, doi:10.1029/2008WR007316.

Herrero, J., M. Polo, A. Moñino, and M. A. Losada (2009), An energy balance snowmelt model in a Mediterranean site, *Journal of Hydrology*, *371*(1-4), 98–107, doi:10.1016/j.jhydrol.2009.03.021.

Hundecha, Y., M. Pahlow, and A. Schumann (2009), Modeling of daily precipitation at multiple locations using a mixture of distributions to characterize the extremes, *Water Resources Research*, *45*(12), doi:10.1029/2008WR007453.

Katz, R. W., M. B. Parlange, and P. Naveau (2002), Statistics of extremes in hydrology, *Advances in Water Resources*, *25*(8-12), 1287–1304, doi:10.1016/S0309-1708(02)00056-8.

Kottegoda, N. T., and R. Rosso (2008), *Applied statistics for civil and environmental engineers*, 718 pp., 2nd ed., Blackwell Publishing Ltd.

Luceño, A., M. Menéndez, and F. Méndez (2006), The effect of temporal dependence on the estimation of the frequency of extreme ocean climate events, *Proccedings of the Royal Society A*, *462*, 1638–1697.

Mans, C., S. Bramato, A. Baquerizo, and M. A. Losada (2011), Surface Seiche Formation on a Shallow Reservoir in Complex Terrain, *Journal of Hydraulic Engineering*, *137*(5), 517–529, doi:10.1061/(ASCE)HY.1943-7900.0000328.

Millares, A., M. Polo, and M. A. Losada (2009), The hydrological response of baseflow in fractured mountain areas, *Hydrology and Earth System Sciences Discussions*, *6*(2), 3359–3384, doi:10.5194/hessd-6-3359-2009.

Méndez, F. J., M. Menéndez, A. Luceño, and I. J. Losada (2006), Estimation of the long-term variability of extreme significant wave height using a time-dependent peak over threshold (POT) model, *Journal of Geophysical Research*, *111 (C07024)*, 1–13.

Nocedal, J., and S. J. Wright (2006), *Numerical Optimization*, 2nd ed., 686 pp., Springer.

Pickands, J. I. (1975), Statistical inference using extreme order statistics, *The Annals of Statistics*, *3*(1), 119–131.

Smith, J. A. (1987), Estimating the upper tail of flood frequency distributions, *Water Resources Research*, *23*(8), 1657–1666, doi:10.1029/WR023i008p01657.

Solari, S., and M. A. Losada (2011), Non-Stationary Wave Height Climate Modeling and Simulation, *Journal of Geophysical Research*, *116* (C09032), 1–18, doi: 10.1029/2011JC007101.

Tancredi, A., C. Anderson, and A. O'Hagan (2006), Accounting for threshold uncertainty in extreme value estimation, *Extremes*, *9*, 87–106.

Thompson, P., Y. Cai, D. Reeve, and J. Stander (2009), Automated threshold selection methods for extreme wave analysis, *Coastal Engineering*, *56*, 1013–1021.

Vaz de Melo Mendes, B., and H. Freitas Lopes (2004), Data driven estimates for mixtures, *Computational Statistics and Data Analysis*, *47*, 583–598.

Vrac, M., and P. Naveau (2007), Stochastic downscaling of precipitation: From dry events to heavy rainfalls, *Water Resources Research*, *43*(7), 1–13, doi: 10.1029/2006WR005308.

Zea Bermudez, P., and S. Kotz (2010b), Parameter estimation of the generalized pareto distribution - part I, *Journal of Statistical Planning and Inference*, *140*(6), 1353–1373, doi:10.1016/j.jspi.2008.11.019.

Zea Bermudez, P., and S. Kotz (2010a), Parameter estimation of the generalized pareto distribution - part II, *Journal of Statistical Planning and Inference*, *140*(6), 1374–1388, doi: 10.1016/j.jspi.2008.11.020.

Chapter 2

Unified distribution models for met-ocean variables: application to series of significant wave height.

2.1 Abstract

The design of maritime works requires statistical models for several met-ocean variables, such as significant wave height H_S, that adequately represent the probability of occurrence and the uncertainty for the entire range of values of the variables. In general, the mean climate of H_S is modeled empirically or by log-normal or two-parameter Weibull distributions using all available data. Extremal climate studies are conducted separately, usually by means of the peaks-over-threshold (POT) method. The methods used to define the threshold tend to be subjective and generally do not allow for calculation of the associated uncertainty.

This paper proposes a mixture model for the marginal distribution of H_S that includes thresholds between the central regime and minimum and maximum regimes as parameters of the model. The parameters of the model are estimated by maximum likelihood, for which specific recommendations are given. The distribution is able to parametrically model the mean climate of the variable. For calculating extreme values, a simple methodology is described that accounts for the uncertainty stemming from the estimation of the threshold.

The implementation of this model using two H_S series shows that it provides a better fit for the data than that obtained with parametric distributions such as log-normal or Weibull. Furthermore, this model automatically and objectively determines the threshold necessary to apply the POT method and the uncertainty associated with the threshold.

2.2 Introduction

Today, a probabilistic design method is often used to quantify the failure probability and service performance of structures, as well as their uncertainty. Design standards and recommendations require designers to analyze the performance of structures from a global perspective, calculating not only their reliability but also their serviceability and operationality during their entire useful life. This is achieved through the use of ultimate limit states (ULS), serviceability limit states (SLS) and operational limit states (OLS) (e.g.: Losada, 2002). These states cover the entire range of climate conditions: minima or calms, central or normal and maximum or severe conditions (central conditions are the range of values the variable normally takes and are those of relatively high probability, i.e., the bulk of the data or the mean climate). Examples of such conditions are provided in the performance-based designs discussed by Takahashi et al. (2001) for caisson breakwaters and by Sato et al. (2001) for beaches.

Within this context, it is necessary to have probability models for met-ocean variables that quantify as accurately as possible their frequency of occurrence and their uncertainty. Such models would preferably cover the entire range of values of the variables and would thus model both the central distribution and the tails. This aspect is particularly important when the system response depends not only on storm conditions (maximum regime) but also on central and calm conditions (central and minimum regimes, respectively), as in the case of beaches.

The met-ocean variables that are of greatest interest in coastal engineering are wave, wind and water level. In this study, attention is focused on significant wave height (H_S or H_{m0}). However, the procedure followed as well as the conclusions derived are equally valid if the analysis is applied to any of the other variables.

Coastal engineers should be able to answer the following questions when analyzing a series of H_S:

(a) What is the marginal distribution of the variable? Is it possible to obtain a good fit with a parametric function?

(b) After which value is the variable in the extreme regime? In this case, what is the distribution followed by the variable?

(c) What are the values of the high-return period quantiles, and what is their uncertainty?

Figure 2.1 shows the steps conventionally followed to respond to these questions. The study of the central regime (i.e., the bulk of the data) is differentiated from that of the maximum and minimum regimes. For the central regime, either certain standard distributions are tested or an empirical distribution is used. These approximations necessarily limit the validity of the results to the central regime because standard distributions do not usually provide a good fit for the tails (Ochi, 1998), whereas extrapolation is not possible with an empirical distribution. The study of maximum regimes only centers

on the calculation of high-return period (Tr) quantiles. For this purpose, the peaks-over-threshold (POT) method is now commonly used instead of the annual maximum method (Goda, 2000; Coles, 2001; Thompson et al., 2009; Mazas and Hamm, 2011). The definition of the threshold, which is a necessary part of this method, is usually done by a graphical process or is based on good practice rules. However, these approaches have the drawback of being subjective. As an alternative, a semi-automatic method has recently been proposed (see Thompson et al., 2009). High-return period quantiles are calculated on the basis of these estimated thresholds; nevertheless, the calculation of confidence intervals generally does not take into account threshold uncertainty.

This paper proposes a new procedure, which is graphically represented in figure 2.2. The first step is to build a mixture distribution that provides a good fit for the entire range of variable values and thus is valid for the central regime as well as for the maximum and minimum regimes. This mixture distribution includes the thresholds between regimes as parameters of the model. For this reason, their estimation is objective. Finally, the uncertainty in the estimation of thresholds is used in the calculation of confidence intervals of high-return period quantiles.

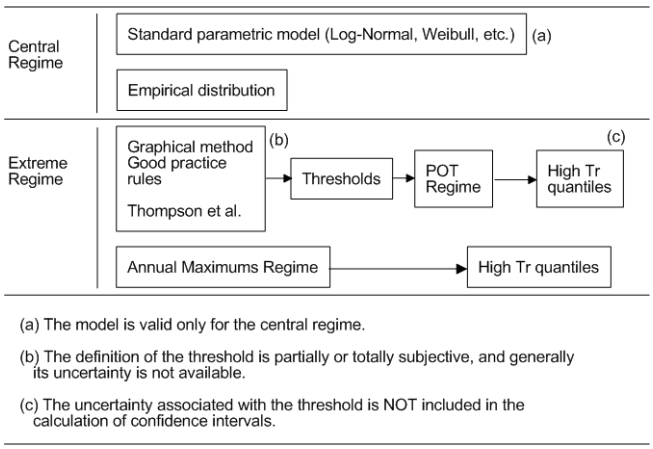


Figure 2.1: Flow chart of the procedure generally followed for the study of a random met-ocean variable (e.g., significant wave height H_S). In the figure, Tr denotes the return period and POT indicates peaks over threshold.

The article is organized as follows. In section 2.3, the proposed mixture distribution models are introduced. The use of the proposed model as a tool for threshold identification for the calculation of extreme values is explained in section 2.4, along with existing methodologies for choosing the threshold. Section 2.5 revisits the methodology for the calculation of high-return period quantiles and proposes a methodology to include the threshold uncertainty in the calculation of the confidence intervals. Section 3.4 is devoted to the application of the proposed model. The wave climate data used throughout this work are described in section 2.6.1. In section 2.6.2, the models proposed are fit, and their performance is compared with the commonly used parametric distributions. Section 2.6.3 analyzes the calculation of extremes with the POT method, in which the threshold identification is discussed. Section 4.11 discusses the obtained results, and

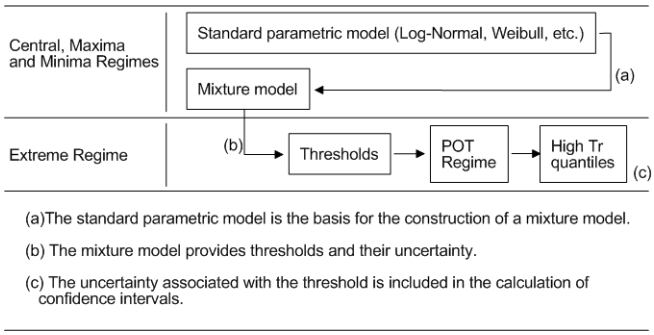


Figure 2.2: Flow chart of the proposed procedure, based on the use of mixture models. In the figure, Tr denotes the return period and POT indicates peaks over threshold.

section 2.8 summarizes the main conclusions derived from this study.

2.3 Proposed models

The proposed model consists of a central distribution, representative of the bulk of the data, and two distributions for the tails of the distribution, which are representative of the minimum and maximum behavior. For the central distribution, a log-normal distribution (LN) was used, although this methodology is valid for any other distribution. Generalized Pareto distributions (GPDs) were used for the tails. The proposed model will henceforth be referred to as LNGPD.

Two versions of the model are proposed. In the first, called LNGPD(A), the transition between the central distribution and those of the tails is abrupt, and it is determined by the values of the location parameters of the GPDs of the tails. In the second model, called LNGPD(F), the transition between the distributions is smoothed through the use of two scaling functions. Both versions are described below.

2.3.1 LNGPD(A)

The LNGPD(A) distribution is given by (3.1), where f_c is the distribution function assumed for the central regime, and f_m and f_M are the distribution functions assumed for the tails.

$$f(x) = \begin{cases} f_m(x)F_c(u_1) & x < u_1 \\ f_c(x) & u_1 \leq x \leq u_2 \\ f_M(x)\left(1 - F_c(u_2)\right) & x > u_2 \end{cases} \tag{2.1}$$

For the distribution of the central regime, an LN is used:

$$f_c(x) = \frac{1}{x\sigma_{LN}\sqrt{2\pi}} \exp\left\{-\frac{[\ln(x) - \mu_{LN}]^2}{2\sigma_{LN}^2}\right\} \tag{2.2}$$

where μ_{LN} and σ_{LN} are the location and scale parameters, respectively.

Minimum and maximum GPDs are used for the tails

$$f_m(x|x < u_1) = \frac{1}{\sigma_1}\left(1 - \frac{\xi_1}{\sigma_1}(x - u_1)\right)^{-\frac{1}{\xi_1}-1} \quad \xi_1 \neq 0 \tag{2.3}$$

$$f_M(x|x > u_2) = \frac{1}{\sigma_2}\left(1 + \frac{\xi_2}{\sigma_2}(x - u_2)\right)^{-\frac{1}{\xi_2}-1} \quad \xi_2 \neq 0 \tag{2.4}$$

where u_1 and u_2 are the location parameters, taken as the lower and upper thresholds of the central regime; σ and ξ are the scale and shape parameters, respectively, such that $\sigma > 0$ and $-\infty < \xi < \infty$. The case with $\xi = 0$, for which the GPD reduces to an exponential distribution with parameter $\sigma > 0$, is excluded. Furthermore, for minimum GPD, $u_1 + \sigma_1/\xi_1 \leq x \leq u_1$ if $\xi_1 < 0$ and $x \leq u_1$ if $\xi_1 > 0$, whereas for maximum GPD, $u_2 \leq x \leq u_2 - \sigma_2/\xi_2$ if $\xi_2 < 0$ and $x \geq u_2$ if $\xi_2 > 0$.

The density function is assumed to be continuous and to always be $x \geq 0$, which leads to the following relationships between parameters:

$$\sigma_1 = -\xi_1 u_1 \qquad \xi_1 = -\frac{F_c(u_1)}{u_1 f_c(u_1)} \qquad \sigma_2 = \frac{1 - F_c(u_2)}{f_c(u_2)} \tag{2.5}$$

The LNGPD(A) model has 5 parameters (u_1, u_2, μ_{LN}, σ_{LN}, ξ_2), which are estimated by maximum likelihood (see appendix for details).

2.3.2 LNGPD(F)

The LNGPD(F) model is given by

$$f(x) = \begin{cases} f_c(x)q(x) + G(1 - q(x))f_m(x) & x < u_1 \\ f_c(x) & u_1 \leq x \leq u_2 \\ f_c(x)p(x) + E(1 - p(x))f_M(x) & x > u_2 \end{cases} \tag{2.6}$$

In this model, functions $q(x)$ and $p(x)$ produce a smooth transition from the central regime (f_c) to the tails (f_m and f_M). In the model, E and G are constant values, given by (2.7), that scale the tail models and ensure that total probability is one.

$$G = \frac{\int_0^{u_1}(1 - q(x))f_c(x)dx}{\int_0^{u_1}(1 - q(x))f_m(x)dx} \tag{2.7a}$$

$$E = \frac{\int_{u_2}^{\infty}(1 - p(x))f_c(x)dx}{\int_{u_2}^{\infty}(1 - p(x))f_M(x)dx} \tag{2.7b}$$

In model (2.6), an LN distribution (2.2) is used for the central regime f_c, whereas GPDs (3.2a) and (3.2b) are used for minimum and maximum regimes (f_m and f_M,

respectively). For $p(x)$ and $q(x)$, decreasing exponential functions are used: $q(x) = \exp(-A_1(u_1 - x))$ and $p(x) = \exp(-A_2(x - u_2))$. Given an observation x of the variable below (above) the threshold u_1 (u_2), functions $q(x)$ ($p(x)$) represent the probability of that observation coming from the central regime f_c.

The use of $p(x)$ and $q(x)$ produces a smooth transition between the central regime and the tails. As a consequence, no particular conditions are required to ensure the continuity of the probability density function of the LNGPD(F) distribution. Nevertheless, once again, the condition $x \geq 0$ is imposed, which implies $\sigma_1 = -\xi_1 u_1$.

The LNGPD(F) has 9 parameters (u_1, u_2, A_1, A_2, μ_{LN}, σ_{LN}, ξ_1, ξ_2 and σ_2), which are estimated by maximum likelihood (see appendix for details).

2.4 Peak over threshold (POT) method

In recent years, the GPD (3.2b) has been used with increasing frequency to model the tails of a distribution (see, e.g., Coles, 2001; Castillo et al., 2005; Holthuijsen, 2007; Thompson et al., 2009; Mazas and Hamm, 2011). The use of the GPD is justified by Pickands (1975) and requires the definition of a threshold over which the data can be approximated by this distribution

When exceedances over the threshold show a tendency to form clusters (i.e., exceedances are produced by a storm event that may last for several days, resulting in observations of H_S that are not independent), it is necessary to decluster the data to obtain a series of independent observations. The POT method is a declustering methodology (see e.g., Davison and Smith, 1990; Coles, 2001): given a certain threshold, exceedances that are separated by less than a given lag are assumed to be generated by the same extreme event (i.e., the same storm); then, for every cluster defined in this way, the maximum recorded value is taken. A comprehensive treatment of the POT method and its application in maritime engineering can be found in Goda (2000).

Prior to the application of the POT method, it is necessary to choose a threshold and a minimum time lag between threshold exceedances that ensures the independence of the POT series. Here, the minimum lag between storms is assumed to be a given parameter, and thus, no analysis of it is performed. With regard to the threshold, various methodologies are available to estimate this value. A few of these methodologies are described below.

In this study, it is shown that parameter u_2 of the LNGPD(A) model is a good approximation of the threshold necessary to apply the POT method; therefore providing an automatable methodology to identify this threshold (see section 2.6.3), which would be an alternative to the methodologies discussed below.

2.4.1 Graphical methods

The Mean Residual Life Plot (MRLP) designates the graph of the series $\{u, 1/n_u \sum_{i=1}^{n_u}(x_i - u)\}$, where $\{x_i\}$ is the series of data such that $u < x_i < x_{max}$ and n_u are the number of elements of $\{x_i\}$. The MRLP should be linear above the threshold u_0 from which point the GPD provides a good approximation of the distribution of the data. Similarly, the

estimates of ξ and $\sigma^* = \sigma - \xi u$, where ξ and σ are the shape and scale parameters of the GPD, should be constant above this threshold. The graphical method consists of creating such graphs, based on which threshold u_0 is then visually estimated (for more details see Coles (2001)).

2.4.2 Method proposed by Thompson et al. (2009)

The method proposed by Thompson et al. (2009) consists of calculating parameter σ^* for a series of thresholds. Above u_0 after which the GPD provides a good approximation to the data, the series $\{\sigma^*_{u_i} - \sigma^*_{u_{i-1}} \,|\, u_i > u_0\}$ should have a zero-mean normal distribution. After the application of the chi-square normality test to the series $\{\sigma^*_{u_i} - \sigma^*_{u_{i-1}}\}$, u_0 is selected as the first value of the series $\{u_i\}$ for which the hypothesis test does not reject the null hypothesis (for more details see Thompson et al. (2009)).

The method requires the previous definition of four parameters: the level of significance used for the hypothesis test and the threshold series $\{u_i\}$, defined by a minimum threshold, a maximum threshold and the number of intermediate thresholds. Thompson et al. (2009) report that, in the resolution of their problem, they obtain good results using a series of 100 thresholds between the quantile corresponding to 50% of the sample and the minimum between the quantile corresponding to 98% and the value that is exceeded only by 100 data.

2.4.3 Method proposed by Mazas and Hamm (2011)

Mazas and Hamm (2011) recommend the use of the graphical method, based on the stability of the parameters ξ y σ^*, in conjunction with an analysis of the mean number of peaks obtained per year. Based on their experience, these authors recommend selecting a threshold that, meeting the stability condition of the parameters ξ y σ^*, ranges from 2 to 5 peaks per year.

2.5 Quantiles and confidence intervals estimation

Assuming that the minimum lag between threshold exceedance that ensures the independence of POT observations is a known parameter, the POT method can be applied using the upper threshold obtained with the LNGPD(A) model (3.1) or with any of the methods discussed previously. Thus, a POT series of independent observations is obtained. A GPD model (3.2b) can be fit to this series by means of maximum likelihood, and covariance matrix $\widehat{Cov}_{\xi\sigma}$ can be obtained for the GPD parameters ξ and σ.

By applying the POT method, the mean number of storms (clusters) per years ν is also obtained. Under the hypothesis that storm occurrences follow a Poisson model, the maximum likelihood estimation of ν and its variances are: $\nu = N/T$ and $\widehat{\sigma}^2_\nu = \nu^2/N$, where N is the number of observed storms (clusters) and T is the number of recorded years.

Then, the covariance matrix for (ν, ξ, σ) can be expressed as (Coles, 2001):

$$\widehat{Cov}(\nu,\xi,\sigma) = \begin{bmatrix} \widehat{\sigma}_\nu^2 & 0 & 0 \\ 0 & \widehat{Cov}_{\xi\sigma}(1,1) & \widehat{Cov}_{\xi\sigma}(1,2) \\ 0 & \widehat{Cov}_{\xi\sigma}(2,1) & \widehat{Cov}_{\xi\sigma}(2,2) \end{bmatrix}$$

Based on the above expression, to include the uncertainty associated with the threshold (which is another parameter of the GPD) in the calculation of the quantiles of a high-return period, it is proposed to approximate the covariance matrix as follows:

$$\widehat{Cov}(\widehat{\theta}) = \begin{bmatrix} \widehat{\sigma}_u^2 & 0 & 0 & 0 \\ 0 & \widehat{\sigma}_\nu^2 & 0 & 0 \\ 0 & 0 & \widehat{Cov}_{\xi\sigma}(1,1) & \widehat{Cov}_{\xi\sigma}(1,2) \\ 0 & 0 & \widehat{Cov}_{\xi\sigma}(2,1) & \widehat{Cov}_{\xi\sigma}(2,2) \end{bmatrix} \tag{2.8}$$

where it is assumed that covariances in reference to u are all zero. In this study, $\widehat{\sigma}_u^2$ is obtained from model (3.1).

With the GPD fit to the POT data series and with covariance matrix (2.8), the quantiles for a high-return period (T_R) are estimated by

$$x_{T_r} = u + \frac{\sigma}{\xi}\left((T_r\nu)^\xi - 1\right) \tag{2.9}$$

with the following confidence intervals for level $(1-\alpha)$:

$$x_{T_r} \in (\widehat{x}_{T_r} \pm z_{\alpha/2}\widehat{\sigma}_{\widehat{x}_{T_r}}) \tag{2.10}$$

where $\widehat{\sigma}^2_{\widehat{x}_{T_r}} = \nabla_\theta^T x_{T_r} \widehat{Cov}(\widehat{\theta}) \nabla_\theta x_{T_r}$, being

$$\nabla_\theta^T x_{T_r} = \left[\frac{\delta x_{T_r}}{\delta u}, \frac{\delta x_{T_r}}{\delta \nu}, \frac{\delta x_{T_r}}{\delta \xi}, \frac{\delta x_{T_r}}{\delta \sigma}\right] =$$
$$\left[1, \frac{\sigma}{\nu}(T_r\nu)^\xi, \frac{\sigma}{\xi^2}(1-(T_r\nu)^\xi) + \frac{\sigma}{\xi}(T_r\nu)^\xi \ln(T_r\nu), \frac{1}{\xi}((T_r\nu)^\xi - 1)\right]$$

2.6 Application

2.6.1 Data series

This research used two series of hindcast spectral significant wave height, provided by Puertos del Estado, Spain (www.puertos.es), one for Cádiz (36.5°N, 6.5°W) and one for Barcelona (41.38°N, 2.38°E). The locations of the two points are shown in figure 2.3.

The Cádiz series comprises 13 years and 3 months of sea states with a duration of 3 hours, although with some gaps in the record (36,496 data). The Barcelona series comprises 14 years and 7 month of sea states with a duration of 3 hours (42,549 data).

The usual distributions for modeling the central regime are the LN, Weibull, Gamma, etc., distributions, although the one that provides the best fit for the data series in this case is the LN, justifying the adoption of the LN for the central regime in the LNGPD

models. Furthermore, in this study, the LN distribution was used as a reference to evaluate the performance of the models proposed.

The original hindcast data were truncated with a precision of $0.1\,m$. For this reason, the data were uniformly distributed beforehand in intervals of $\pm 0.05\,m$. This avoided relative minima and maxima in the likelihood function and facilitated the estimation of the parameters (see appendix).

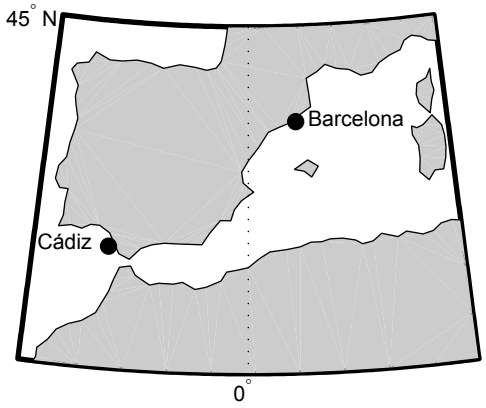


Figure 2.3: Locations of the Cádiz and Barcelona points.

2.6.2 Parametric probability distribution model

This section describes how the LNGPD models were applied to the series of significant wave height presented in the previous section and compares the results to those obtained with an LN distribution. The estimated values of the LNGPD(A) and LNGPD(F) parameters are listed in tables 2.1 and 2.2, respectively, along with their standard deviations.

Figures 2.4 through 2.7 present the empirical probability density function (PDF) and cumulative distribution function (CDF), along with the functions obtained with the adjusted models (LN and LNGPD).

Figure 2.4 (Cádiz) shows that the LNGPD(F) distribution significantly improves the fit of the mode of the data with respect to the fit obtained with the LN distribution. In the case of LNGPD(A), an improvement of the fit in the central zone was observed around the mode, but this improvement was not as notable as in the previous case. In figure 2.5, it can be observed that the two LNGPD distributions improve the fit in the tails with respect to the fit obtained with the LN distribution. The fit obtained with the LNGPD(A) distribution is the one that best follows the trend of the data in the upper tail.

For Barcelona, it can also be observed that both LNGPD distributions significantly improve the fit of the data in the central zone, near the mode (see figure 2.6). In the upper tail, LN and LNGPD(A) display a similar fit (both distributions follow the trend of the data), whereas LNGPD(F) diverges from this trend (see figure 2.7).

Table 2.1: LNGPD(A) parameters with their corresponding standard deviations in brackets.

Location	u_1	u_2	μ_{LN}	σ_{LN}	ξ_2
Cádiz	0.41	3.7	-0.103	0.616	-0.113
	(3.6×10^{-3})	(0.081)	(3.2×10^{-3})	(2.3×10^{-3})	(0.028)
Barcelona	0.27	0.52	-0.572	0.553	0.088
	(3.6×10^{-3})	(4.4×10^{-3})	(2.9×10^{-3})	(2.9×10^{-3})	(6.4×10^{-3})

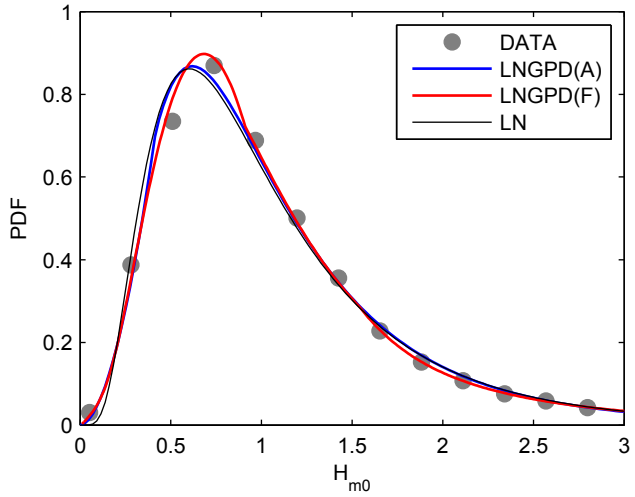


Figure 2.4: Empirical (gray dots), LN (black line), LNGPD(A) (blue line) and LNGPD(F) (red line) PDF for the Cádiz data series.

2.6.3 Extreme values and confidence intervals

This section analyzes the use of the upper threshold u_2, which is obtained from the LNGPD(A) distribution, in the application of the POT method. This requires two steps:

(a) The upper threshold necessary to apply the POT method is estimated by using other available methodologies, namely, the following: the graphical methods presented in Coles (2001) and the methods that have recently been presented by Thompson et al. (2009) and by Mazas and Hamm (2011).

(b) The extreme value distribution is obtained for each of the thresholds, comparing the quantiles of the high-return period and their confidence intervals.

In all of the cases, to define the storms, a minimum time of 2 days is imposed between the end of an excess over the threshold and the beginning of another.

En algunos casos esto resulta en que usando umbrales menores se obtenienen menos picos por año que usando umbrales mayores. In some cases this results in that fewer peaks per year are obtained using lower thresholds than using higher thresholds. An-

Table 2.2: LNGPD(F) parameters with their corresponding standard deviations in brackets.

Location	u_1	u_2	A_1	A_2	μ_{LN}	σ_{LN}	ξ_1	ξ_2	σ_2
Cádiz	0.91	1.54	1.00	1.19	-0.110	0.607	0.444	0.009	0.857
	(9.3×10^{-3})	(0.011)	(0.053)	(0.128)	(3.6×10^{-3})	(3.9×10^{-3})	(9.5×10^{-3})	(3.6×10^{-3})	(0.02)
Barcelona	0.48	0.48	0.78	1.80	-0.574	0.554	0.440	-0.031	0.475
	(0.0)	(0.0)	(0.190)	(0.328)	(3.4×10^{-3})	(4.2×10^{-3})	(0.032)	(0.010)	(0.013)

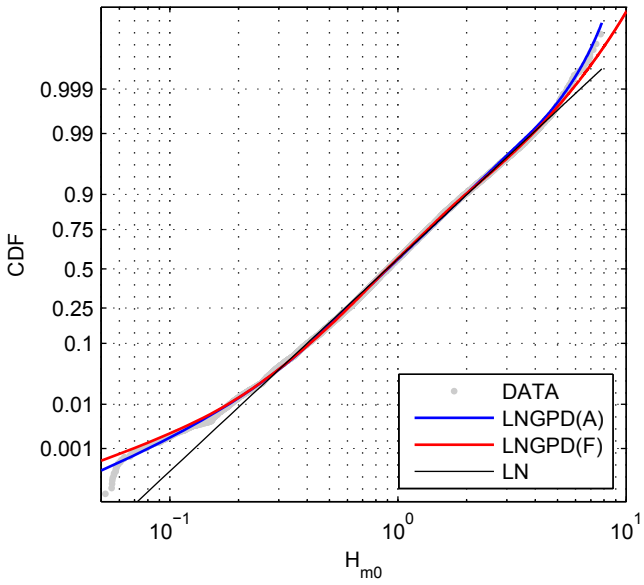


Figure 2.5: Empirical (gray dots), LN (black line), LNGPD(A) (blue line) and LNGPD(F) (red line) CDF for the Cádiz data series.

other way of defining storms could result in different behavior, but the study of the sensitivity of the POT method to this parameter is beyond the scope of this article.

Threshold selection

The thresholds obtained from the LNGPD(A) model are $u = 3.7\,m$ for Cádiz and $u = 0.52\,m$ for Barcelona (see table 2.1).

Figures 2.8 and 2.9 present the MRLP and graphs ξ and σ^* for the data series for Cádiz, whereas figures 2.10 and 2.11 present the same graphs for the data series for Barcelona. In the case of Cádiz, the threshold value that is selected is $u = 3.5\,m$ (indicated in figures 2.8 and 2.9 by a red circle). In the case of Barcelona, there are two possible thresholds. Limiting the analysis to the MRLP (figure 2.10), the selected threshold is $u = 1.9\,m$. However, upon analyzing graphs ξ and σ^* (figure2.11), the selected threshold is found to be $u = 0.8\,m$. Given that the MRLP does not discard the use of the latter threshold, it is assumed that both thresholds are valid.

Table 2.3 summarizes the result of applying the method proposed by Thompson et al. (2009) using different values of significance α for the hypothesis test and a series of thresholds constructed with different numbers of elements N and upper thresholds U_{max}. To select the threshold, the criterion was selected of using the value obtained using a significance of $\alpha = 0.2$ for the chi-square test and a series of $N = 100$ thresholds with the upper threshold U_{max} corresponding to the 98% percentile, as recommended by Thompson et al. (2009). With this criterion, $u = 2.5\,m$ is obtained for for Cádiz and $u = 0.9\,m$ for Barcelona.

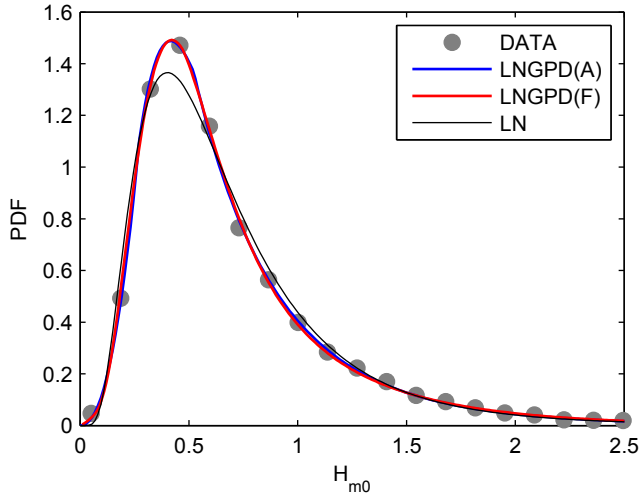


Figure 2.6: Empirical (gray dots), LN (black line), LNGPD(A) (blue line) and LNGPD(F) (redline) PDF for the Barcelona data series.

In Cádiz (12 complete meteorological years), the threshold obtained with the LNGPD(A) model ($u = 3.7\,m$) results in 3.4 peaks per year, whereas the threshold selected using the graphical method ($u = 3.5\,m$) results in 3.6 peaks per year, and the one identified by the method of Thompson et al. (2009) ($u = 2.5\,m$) results in 10.8 peaks per year. For Barcelona (14 complete meteorological years), the threshold of $0.52\,m$ obtained with LNGPD(A) results in 21 peaks per year. The thresholds selected through the graphical method ($u = 0.8\,m$ and $u = 1.9\,m$) result in 37 and 12 peaks per year, respectively, whereas the threshold identified by the method of Thompson et al. (2009) ($u = 0.9\,m$) results in 35 peaks per year.

In the case of Cádiz, the thresholds selected with the graphical method and with LNGPD(A) result in a number of peaks per year that is within the range recommended

Table 2.3: Upper thresholds obtained by applying the methodology proposed by Thompson et al. (2009) using different numbers of elements (N) and upper thresholds (U_{max}) for the definition of the thresholds series and different significance for the chi-square test (α).

Location	U_{max}	$\alpha = 0.2$		$\alpha = 0.05$	
		$N = 100$	$N = 500$	$N = 100$	$N = 500$
Cádiz	98% (3.3 m)	2.5	3.1	1.3	3.1
	98.5% (3.6 m)	2.6	3.4	2.2	3.4
	99% (3.9 m)	2.3	3.8	1.7	3.8
	99.4% (4.4 m)	0.9	4.1	0.9	4.1
Barcelona	98% (2.1 m)	0.9	1.9	0.6	1.9
	98.5% (2.2 m)	1.7	2.1	1.5	2.1
	99% (2.5 m)	1.7	2.3	1.3	2.3
	99.7% (3.2 m)	0.8	3.0	0.6	3.0

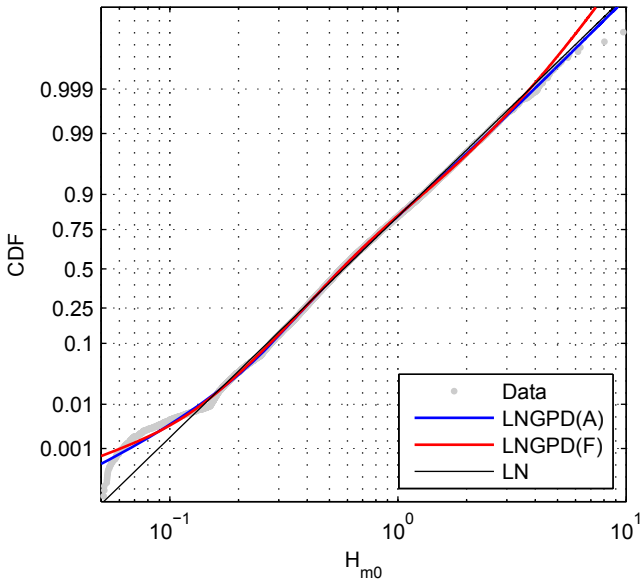


Figure 2.7: Empirical (gray dots), LN (black line), LNGPD(A) (blue line) and LNGPD(F) (red line) CDF for the Barcelona data series.

by Mazas and Hamm (2011), whereas the threshold obtained using the method of Thompson et al. (2009) exceeds this range. In Barcelona, however, all of the methods result in a number of peaks per year that is greater than that recommended by Mazas and Hamm (2011). In this case, it is interesting to note the analysis of Mendoza et al. (2011), who studied storms in the Catalan Mediterranean, where 286 storms were detected in 21 meteorological years, i.e., approximately 13 storms per year.

High return period quantiles

This section compares the extreme values obtained when a GPD was fit to the POT data series constructed using the different thresholds defined in the previous section and listed in table 2.4.

The quantiles estimated for various return periods as well as their 90% confidence intervals are shown in figures 2.12 and 2.13 for Cádiz and Barcelona, respectively. The figures include the empirical quantiles obtained with the different POT data series. Table 2.5 presents the values of H_S from a 100-year return period obtained with the different thresholds, along with the corresponding confidence intervals.

In all of the cases, the confidence intervals were estimated without including the uncertainty associated with the determination of the threshold.

The effect of the uncertainty of the threshold

In this section, the effect of the uncertainty of the threshold in the estimation of the confidence intervals of the quantiles is analyzed. For this, the covariance matrix (2.8) is

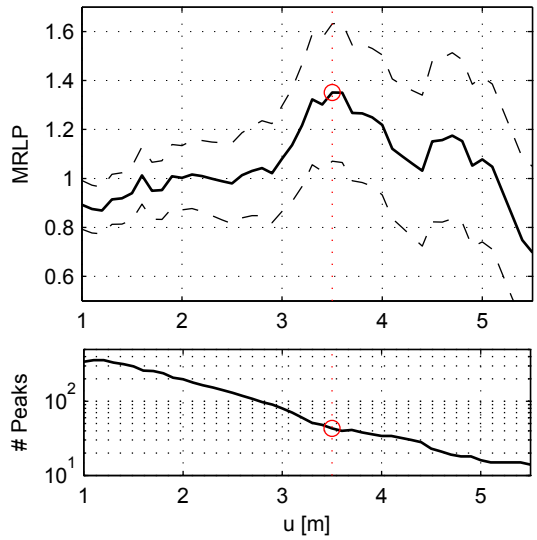


Figure 2.8: MRLP (top) and number of peaks (bottom) for the Cádiz data series.

Table 2.4: Upper thresholds obtained using different methods.

Location	Method	Threshold u [m]
Cádiz	LNGPD(A)	3.7
	Graphical Method	3.5
	Thompson et al. (2009)	2.5
Barcelona	LNGPD(A)	0.52
	Graphical Method	0.8 and 1.9
	Thompson et al. (2009)	0.9

used, taking the threshold (u) and variance (σ_u^2) obtained by adjusting the parameters of the LNGPD(A) model (parameter u_2 in table 2.1).

Figure 2.14 displays the 90% confidence intervals obtained for Cádiz and Barcelona. For Barcelona, it can be observed that the effect of including the variance of the threshold in the calculation of the confidence intervals is negligible. In Cádiz, a small expansion of the confidence intervals is observed, which is limited to the low return periods ($Tr < 10$ years) but is negligible for practical purposes.

2.7 Discussion

2.7.1 Parametric distribution LNGPD models

Both proposed LNGPD models achieved a significantly better fit of the data than the LN distribution. However, the models increase the number of parameters from 2 with the LN to 5 with LNGPD(A) and to 9 with LNGPD(F). Consequently, it was necessary to verify that the improvement of the fit was significant to justify this increase in the

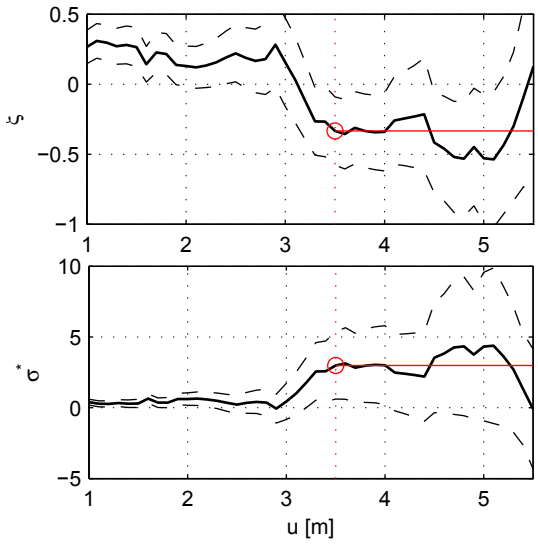


Figure 2.9: Plots of the evolution of ξ and σ^* for different thresholds. Cádiz data series.

Table 2.5: 100-year return period significant wave height ($H_{S,\,Tr=100}$) obtained with the different thresholds and their corresponding 90% confidence intervals (CIs).

Location	Method (u [m])	$H_{S,\,Tr=100}$ [m] (90% C.I.)
Cádiz	LNGPD(A) (3.7)	8.2 (7.0 – 9.5)
	Graphical Method (3.5)	8.2 (7.0 – 9.3)
	Thompson et al. (2009) (2.5)	15.3 (5.9 – 24.7)
Barcelona	LNGPD(A) (0.52)	10.3 (6.7 – 13.8)
	Graphical Method (0.8)	11.7 (7.5 – 15.9)
	Graphical Method (1.9)	12.6 (5.8 – 19.5)
	Thompson et al. (2009) (0.9)	10.6 (7.1 – 14.0)

number of parameters. For that purpose, the Akaike Information Criterion and the Bayesian Information Criterion (see, e.g., Mazas and Hamm, 2011) are used: $AIC = -2\log(L) + 2p$ and $BIC = -2\log(L) + \log(N)p$, where L is the likelihood function evaluated with the estimated parameters, p is the number of parameters of the model, and N is the number of data in the sample. These indexes give a relative measurement of the quality of the fit by penalizing the increase in the likelihood function, depending on the increased number of parameters. Insofar as the AIC and the BIC decrease, the improvement obtained by the increase in the number of parameters is significant.

These indexes were calculated for the three distributions, and the results obtained are shown in table 2.6. It has been observed that in both Cádiz and Barcelona, the LNGPD models achieve a better fit for the data than that obtained with the LN model. In the case of Cádiz, the LNGPD(F) model was the one that provided the best results. This is probably because by allowing a smooth transition between the central and extreme regimes, a better representation of the natural processes is obtained. In the

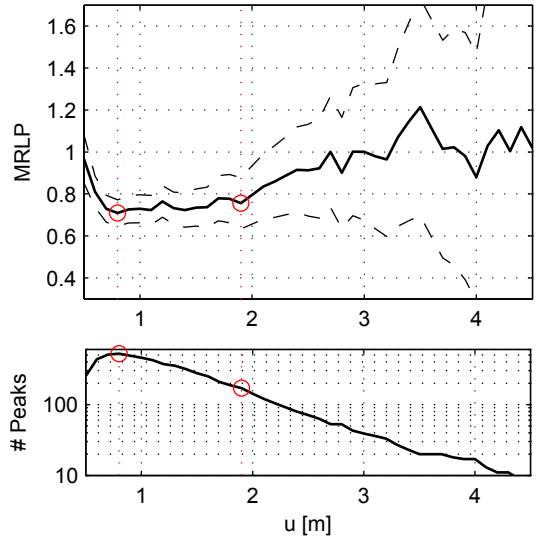


Figure 2.10: MRLP (top) and number of peaks (bottom) for the Barcelona data series.

case of Barcelona, in contrast, the model with the best fit is the LNGPD(A) model, probably because the upper tail follows a GPD based on a relatively low threshold ($0.52\,m$), therefore allowing the central distribution (LN) to fit around the mode, which does not justify the incorporation of the 4 additional parameters that LNGPD(F) has with respect to LNGPD(A).

Table 2.6: *Akaike information Criterion* (AIC) and *Bayesian Information Criterion* (BIC) obtained for the each model and for each data series.

Location	Index	Distribution		
		LN	LNGPD(A)	LNGPD(F)
Cádiz	AIC	61865	61574	61470
	BIC	61882	61616	61547
Barcelona	AIC	31667	31206	31255
	BIC	31684	31249	31333

2.7.2 Threshold selection and high-return period quantiles

Four methods were analyzed to select the threshold needed for applying the POT method, consisting of the graphical methods discussed in Coles (2001), recently proposed methods by Thompson et al. (2009) and Mazas and Hamm (2011), and the use of parameter u_2 obtained by fitting LNGPD(A). The capacity of these methods to identify the threshold was evaluated by adjusting a GPD distribution to the POT series calculated with each threshold and by using this GPD to calculate the quantiles of the high-return period and its confidence intervals.

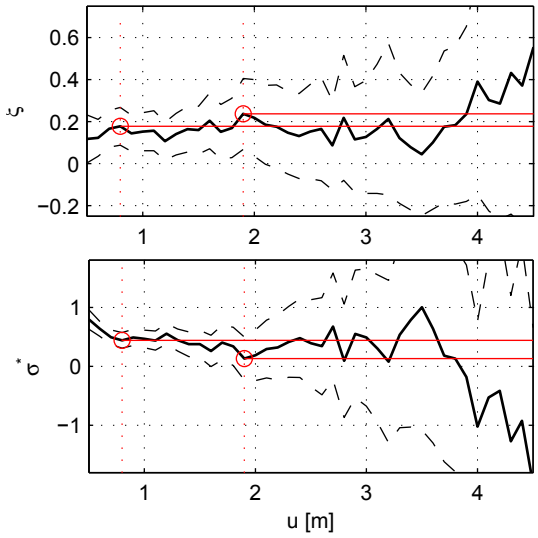


Figure 2.11: Plots of the evolution of ξ and σ^* for different thresholds. Barcelona data series.

It can be observed that for the two data series analyzed, the threshold estimated by the LNGPD(A) model is appropriate to apply the POT method and to calculate high-return period quantiles for the variable. In the case of Barcelona, it may appear that the threshold estimated by LNGPD(A) is low ($u = 0.52\,m$); however, it should be kept in mind that these data correspond to a sheltered location (see figure 2.3).

The LNGPD(A) provided the upper threshold as a parameter of the model, which meant that it could be used as an objective and automatic method of threshold selection. However, one disadvantage of using the LNGPD(A) distribution is that it requires the likelihood function to be rewritten, based on the central distribution f_c to be used, whereas the other methods can be programmed independently of the data.

The graphical method provided a tool that allowed thresholds to be selected that are appropriate for the application of the POT method, reducing the level of subjectivity in the selection. However, because it has some remaining subjective component, the method could not be automatized. On applying this method to the series for Cádiz and Barcelona, adequate thresholds were obtained, although in the case of Barcelona, the GPD obtained with this threshold displayed wider confidence intervals that those obtained with the other thresholds.

The method proposed by Thompson et al. (2009) requires the previous definition of four parameters, and the results obtained with this method were found to be sensitive to these parameters (see table 2.3). This limited the objectivity of the method as well as the possibility of automating it. For Cádiz, thresholds are obtained that are in the range of $0.9 - 4.1\,m$, which includes both the threshold selected using the graphical method $u = 3.5\,m$ and the value of $u_2 = 3.7\,m$ obtained by fitting the LNGPD(A) distribution. For Barcelona, the range of thresholds obtained is $0.6 - 3.0\,m$. This range includes the two possible thresholds selected with the graphical method, $u = 0.8\,m$ and $u = 1.9\,m$,

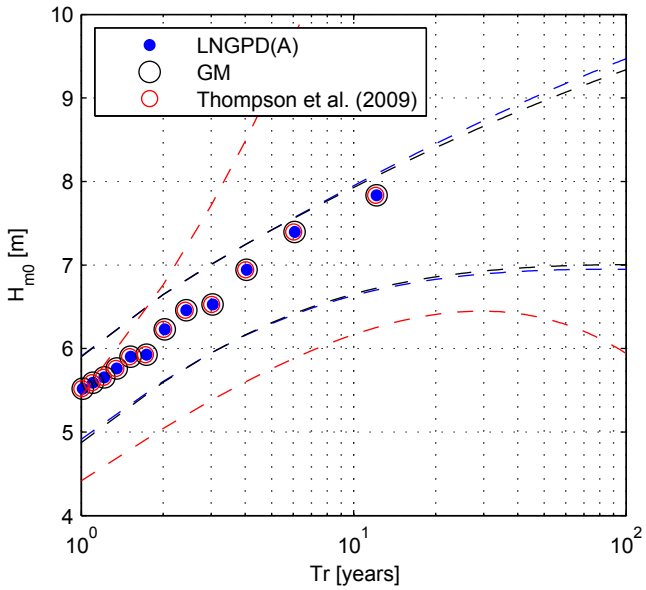


Figure 2.12: Peak over threshold series and 90% confidence intervals for the quantiles obtained with the GPDs fit using thresholds $3.7\,m$ (blue), $3.5\,m$ (black) and $2.5\,m$ (red). Cádiz data series. GM stands for graphical method.

and excludes the value $u_2 = 0.52\,m$ obtained by fitting the LNGPD(A) distribution.

The method proposed by Mazas and Hamm (2011) is based on the use of the graphical method in combination with advice on keeping the mean number of peaks per year between 2 and 5. In Cádiz, the two thresholds that fulfill this recommendation (those obtained with LNGPD(A) and the graphical method) are found to be appropriate for modeling the extreme value behavior of the variable, whereas the threshold that does not meet this recommendation is not appropriate (i.e., that obtained with the method of Thompson et al. (2009), as is explained bellow). In Barcelona, however, all of the thresholds obtained are such that the number of storms per year is greater than 5. In this case, therefore, the recommendation with regard to the range of storms per year is not practical. However, Mazas and Hamm (2011) clarify that the range of 2 to 5 storms per year is not an absolute criterion, and they recommend analyzing the sensitivity of the high-return period quantiles to the selection of the threshold.

With respect to the confidence intervals of the high-return period quantiles, those obtained for Cádiz using the thresholds estimated with LNGPD(A) and with the graphic method are nearly identical. However, the confidence intervals calculated using the threshold obtained with the method of Thompson et al. (2009) are notably larger than the previous ones. They also display a different trend, indicating that this threshold is not appropriate for the calculation of extremes. For Barcelona, in contrast, all of the thresholds used result in GPD distributions that follow the same trend, although some display narrower confidence intervals than others, with the narrowest intervals calculated with the thresholds obtained with LNGPD(A) and the method of Thompson

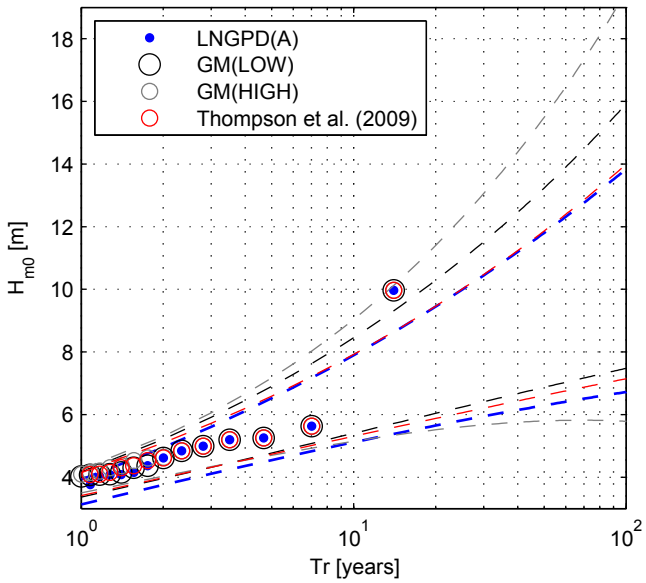


Figure 2.13: Peak over threshold series and 90% confidence intervals for the quantiles obtained with the GPDs fit using thresholds $0.52\,m$ (blue), $0.8\,m$ (black), $1.9\,m$ (grey) and $0.9\,m$ (red). Barcelona data series. GM stands for graphical method.

et al. (2009). Here, it should be kept in mind that the peak of $H_S = 10\,m$ corresponds to an exceptional situation that does not correspond to the return period of 14 years, which is empirically obtained (see Ponce de Leon and Guedes Soares (2008), who analyze this storm because it is the most severe that was detected in the zone in a series of 44 years). Therefore, it is correct that this datum lies outside the confidence intervals.

Taking the values for of 100 years return period listed in table 2.5 as a reference, in the case of Cádiz it can be observed that similar results are obtained with both the threshold identified through LNGPD(A) and the threshold identified with the graphical method, whereas the value obtained with the method of Thompson et al. (2009) is significantly higher and has a larger confidence interval. For Barcelona, in contrast, similar results are obtained with the thresholds identified with LNGPD(A) and the method of Thompson et al. (2009), whereas the thresholds selected using the graphical method are found to have higher values with larger confidence intervals.

Finally, both in Cádiz and in Barcelona, it has been observed that the effect of including the variance of the threshold in the calculation of the confidence intervals has a negligible effect for practical purposes.

2.8 Conclusions

The introduction of this article formulated three questions that coastal engineers should consider when they analyze a series of significant wave heights. This study proposed and applied a model and a methodology to answer these questions with the following

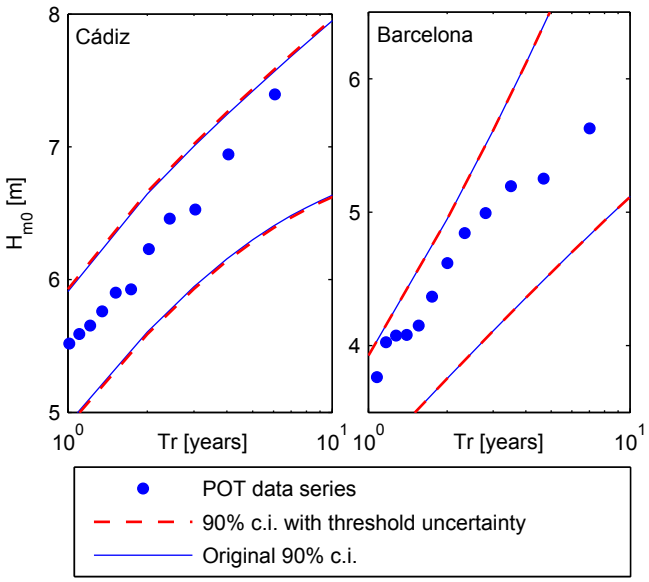


Figure 2.14: Comparison of 90% confidence intervals (c.i.), calculated with (red dashed line) and without (blue line) the threshold variance, for Cádiz (left) and Barcelona (right), respectively.

results:

(a) It was found that LNGPD models provided a better fit for the marginal distribution of the data than that obtained with conventional parametric models. Although this is especially true for the central regime of the distribution, it also applies for the tails.

(b) The threshold obtained with the LNGPD(A) can be used to apply the POT method. High return period quantiles calculated using this threshold are coherent with quantiles calculated with thresholds obtained by other means. However, once again, the LNGPD(A) has the advantage of being objective and susceptible to automatization. Moreover, the threshold obtained with LNGPD(A) was found to be adequate in both cases, whereas one case in each of the other methods failed to give an adequate threshold.

(c) This article has described a method for including threshold uncertainty in the calculation of the confidence intervals of high-return period quantiles when the POT method is used. This requires threshold variance, which is easily obtained when the LNGPD(A) parameters are fit. However, it was observed that including threshold variance did not significantly widen the confidence intervals of the quantiles.

2.A Appendixes.

2.A.1 Estimation procedure

To estimate the parameters of the LNGPD distributions by maximum likelihood, the negative log-likelihood function is minimized using the BFGS method (quasi-Newton method, see, e.g., Nocedal and Wright (2006) chap. 6) implemented in the MATLAB ® optimization toolbox. Prior to the estimation of the parameters, the redistribution of the data was performed. This involves taking the original data, truncated with precision of $0.1\,m$, and distributing them uniformly in they corresponding symmetrical intervals $(X - 0.05, X + 0.05)$ (see Solari and Losada, 2011).

For the two studied series, it was observed that the estimated parameters provided a satisfactory fit with the data and a statistically significant improvement (evaluated through the AIC and the BIC) with respect to the fit obtained with the LN distribution.

Despite the above considerations, to have greater certainty that the optimization method will find the parameters that maximize the likelihood function, the iso-likelihood curves presented in figure 2.15 were constructed. These curves correspond to the LNGPD(A) model and were calculated by setting u_1 and u_2 and calculating the remaining three parameters (μ_{LN}, σ_{LN} and ξ_2) by maximum likelihood.

From the analysis of these graphs, it is shown that the method used in both cases finds the parameters that maximize the likelihood function. With regard to the presence of the two relative maxima in these functions, it is considered that these maxima correspond to the lack of homogeneity of the population, probably produced by seasonal and/or directional variations in the series. The extension of the LNGPD(A) model to non-stationary conditions is developed in Solari and Losada (2011).

Acknowledgments

This research was funded by the Spanish Ministry of Education through its postgraduate fellowship program FPU, grant AP2009-03235. Partial funding was also received from the Spanish Ministry of Science and Innovation (research project CTM2009-10520), the Spanish Ministry of Public Works (projects CIT-460000-2009-21 and 53/08 –FOM/3864/2008 order–) and the Andalusian Regional Government (research project P09-TEP-4630). The authors also acknowledge Puertos del Estado for providing the wave data records.

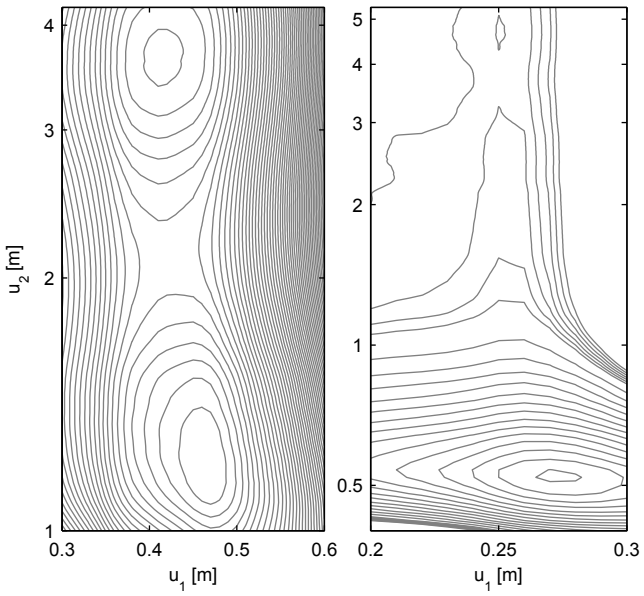


Figure 2.15: Iso-likelihood contours of the LNGPD(A) model, estimated by fixing the parameters u_1 and u_2 for the Cádiz (left) and the Barcelona (right) data series.

References

E. Castillo, A. S. Hadi, N. Balakrishnan, J. M. Sarabia, Extreme Value and Related Models with Applications in Engineering and Science, Wiley Seires in Probability and Statistics, John Wiley & Sons, Inc., ISBN 978-0471-67172-5, 2005.

S. Coles, An Introduction to Statistical Modeling of Extreme Values, Springer Series in Statistics, Springer, Berlin, ISBN 978-1852-33459-8, 2001.

A. Davison, R. Smith, Models for Exceedances over High Thresholds, Journal of the Royal Society Series B 52 (3) (1990) 393–442.

Y. Goda, Random seas and design of maritime structures, World Scientific Publishing Co. Pte. Ltd., 2nd edn., ISBN 978-981-02-3256-6, 2000.

L. H. Holthuijsen, Waves in Oceanic and Coastal Waters, Cambridge University Press, ISBN 978-0-521-86028-4, 2007.

M. A. Losada, ROM 0.0: General procedure and requirements in the design of harbor and maritime structures. PART I, Puertos del Estado, Spain, ISBN 84-88975-30-9, 2002.

F. Mazas, L. Hamm, A multi-distribution approach to POT methods for determining extreme wave heights, Coastal Engineering 58 (5) (2011) 385–394, ISSN 03783839, doi:10.1016/j.coastaleng.2010.12.003.

E. T. Mendoza, J. A. Jimenez, J. Mateo, A coastal storms intensity scale for the Catalan sea (NW Mediterranean), Natural Hazards and Earth System Science 11 (9) (2011) 2453–2462, ISSN 1684-9981, doi:10.5194/nhess-11-2453-2011.

J. Nocedal, S. J. Wright, Numerical Optimization, Springer, 2nd edn., ISBN 978-0387-30303-1, 2006.

M. K. Ochi, Ocean Waves. The Stochastic Approach, Cambridge ocean technology series 6, Cambridge University Press, ISBN 978-0-521-01767-1, 1998.

J. I. Pickands, Statistical Inference Using Extreme Order Statistics, The Annals of Statistics 3 (1) (1975) 119–131.

S. Ponce de Leon, C. Guedes Soares, Sensitivity of wave model predictions to wind fields in the Western Mediterranean sea, Coastal Engineering 55 (11) (2008) 920–929, ISSN 03783839, doi:10.1016/j.coastaleng.2008.02.023.

S. Sato, K. Torii, T. Takagi, Performance-Based Design of Beach as a Shore Protection Facility, in: G. Y., T. S. (Eds.), Proceedings of International Workshop on Advanced Design of Maritime Structures in the 21st Century (ADMS21), Port and Harbour Research Institute. Ministry of Land, Infrastructure and Transport, Japan, 2001.

S. Solari, M. Losada, Non-stationary wave height climate modeling and simulation, Journal of Geophysical Research 116 (C09032) (2011) 1–18, ISSN 0148-0227, doi: 10.1029/2011JC007101.

S. Takahashi, K. Shimosako, M. Hanzawa, Performance Design for Maritime Structures and Its Application to Vertical Breakwaters—Caisson Sliding and Deformation-Based Reliability Design, in: G. Y., T. S. (Eds.), Proceedings of International Workshop on Advanced Design of Maritime Structures in the 21st Century (ADMS21), Port and Harbour Research Institute. Ministry of Land, Infrastructure and Transport, Japan, 2001.

P. Thompson, Y. Cai, D. Reeve, J. Stander, Automated threshold selection methods for extreme wave analysis, Coastal Engineering 56 (10) (2009) 1013–1021, ISSN 03783839, doi:10.1016/j.coastaleng.2009.06.003.

Chapter 3

Non-stationary wave height climate modeling and simulation

3.1 Abstract

The most popular methods of simulating time series for wave heights and other meteorological and oceanic variables are based on the use of autoregressive models and the transformation of variables to make them normal and stationary. Generally, when these models are used, attention is centred on their capacity to represent the autocorrelation of the series.

In this article, a simulation model is proposed that is based on the following: (i) a non-stationary parametric mixture model for the marginal distribution of the variable, that combines a log-normal distribution for main-mass regime and generalised Pareto distributions for upper and lower tail regimes, and (ii) the use of copulas to model the time dependency of the variable. The model has been evaluated by comparing the original series and the simulated series in terms of the autocorrelation function, the mean, the annual maxima and peaks-over-threshold regimes, and the persistences regime. It has also been compared to an ARMA model and found to yield more satisfactory results.

3.2 Introduction

The verification of coastal and harbour structures may require the use of Level III verification methods. These methods are usually complex and require the use of numerical simulation techniques (e.g., Monte Carlo techniques) (*Losada*, 2002).

In coastal engineering, the main variables to be simulated are sea-state variables such as significant wave height, wind, and sea level, which characterise the sea state in a time domain in which processes are assumed to be stationary. For this purpose, generally speaking, the duration should not exceed O(1hr). This research focuses on the evolutionary behaviour of the sea-state variables, i.e., on long-term analysis.

From a physical point of view, the temporal evolution of sea-state variables is conditioned by phenomena operating on different time scales.

Table 3.1: Outline of the relationships of dependence.

Connection with	NO	YES
Other variables	Univariate	Multivariate
Same variable	Without auto-correlation	With auto-correlation
Time	Stationary	Non-stationary

Processes with a time scale of O(day)-O(weeks), such as synoptic phenomena and the cycles of spring and neap tides, produce dependence among the variables that originate and autocorrelation in each variable. The clearest example related to sea states is the passage of a storm. The storm will generate wind speeds and wave heights that are larger than average, and therefore, it is expected that these variables will be correlated during a storm. At the same time, the evolution of these variables (and others) over time is determined by the intensity and path of the storm, so there are physical reasons to expect that these variables will present significant autocorrelation within the time scale of the storm.

O(year) scale processes, such as seasons, produce variations in the intensity and frequency of the O(day)-O(week) scale phenomena and thus cause temporal variations in sea-state variables. In the same way, O(>year) scale processes, such as interannual variability, influence the characteristics of each year (e.g., they create drier or wetter years and years with more or less wave action) and also produce temporal variations in sea-state variables.

Regarding the statistical tools used in the long-term analysis of sea-state variables, it is important to note that such studies can be univariate or multivariate, may or may not include auto-correlation, and can be stationary or non-stationary. Table 3.1 summarises the characteristics of a study: whether the variables are dependent on other variables (i.e., whether they are correlated with other variables), whether the variables are self-dependent (i.e., exhibit autocorrelation or time dependence), or whether they are dependent on time (i.e., whether their distribution is non-stationary). The long-term (climate) behaviour of sea-state variables includes such characteristics and, consequently, should be studied using non-stationary multivariate models that represent the time dependence (or auto-correlation) of the variables.

In figure 3.1, various physical phenomena evolving in different time scales are associated with statistical models that have been used in this study to appropriately model the sea-state variables for these time scales.

The maximum time scale that the simulation must take into account to be applied to engineering is the period used to verify the system. This period is generally the useful life of the system, which is 10-50 years, although it can be a shorter duration when the aim is to verify construction processes or evaluate other short-term phenomena.

With regard to the simulation of times series for significant wave heights (H_s or H_{m0}), there are currently two lines of research: one that focuses on simulating storms and another that simulates complete series of values.

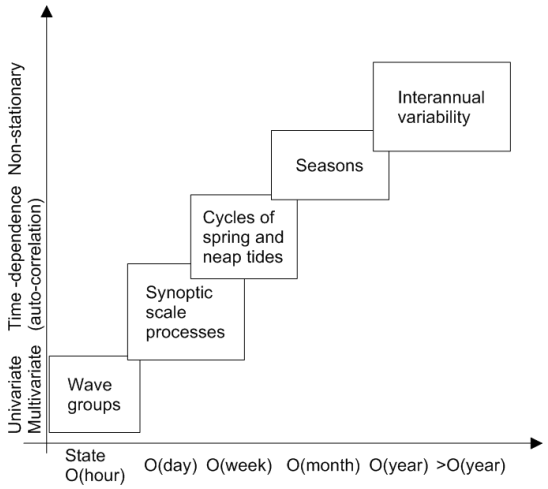


Figure 3.1: Physical phenomena evolving in different time scales, and statistical models for the appropriate modelling of the sea-state variables.

The method most widely used to simulate storms involves developing joint or conditioned distributions for the random variables of storm occurrence, intensity, and duration. Based on these distributions, new time series are simulated assuming a standard shape for the storm.

In general, storm occurrence is modelled using a Poisson distribution and storm intensity using a generalised Pareto distribution (GPD). It is common to condition the duration of a storm to its intensity. Some examples of this type of approximation are presented in *DeMichele et al.* (2007); *Payo et al.* (2008); *Callaghan et al.* (2008). Although stationary functions are generally used for this purpose, non-stationary functions can also be employed, such as those proposed in *Luceño et al.* (2006); *Méndez et al.* (2006, 2008); *Izaguirre et al.* (2010). A less frequent alternative in storm simulation is to assume that it is a Markov process and to use a multivariate distribution of extremes to model the time dependence of the variable while the storm lasts (*Coles*, 2001, chap. 8). This technique is used in *Smith et al.* (1997); *Fawcett and Walshaw* (2006); *Ribatet et al.* (2009).

Monbet et al. (2007) review simulation methods for complete time series applied to wind and waves. The methods currently used can be classified as parametric and non-parametric.

The Translated Gaussian Process (TGP) method (*Walton and Borgman*, 1990; *Borgman and Scheffner*, 1991; *Scheffner and Borgman*, 1992) is the most widely used non-parametric method. This method uses the spectrum of the normalised variable. According to *Monbet et al.* (2007), non-parametric methods such as those based on resampling (called resampling methods) are less frequently used and are not discussed in this article.

The most frequently used parametric methods are based on autoregressive models. Studies employing such methods include *Guedes Soares and Ferreira* (1996); *Guedes*

Soares et al. (1996); *Scotto and Guedes Soares* (2000); *Stefanakos* (1999); *Stefanakos and Athanassoulis* (2001); *Cai et al.* (2007) for univariate series; for multivariate series, relevant studies include *Guedes Soares and Cunha* (2000); *Stefanakos and Athanassoulis* (2003); *Stefanakos and Belibassakis* (2005); *Cai et al.* (2008). As in the TGP, before autoregressive models can be used, the series must be normalised. For this purpose, non-stationary models of the mean and the standard deviation, like those proposed by *Athanassoulis and Stefanakos* (1995); *Stefanakos* (1999); *Stefanakos et al.* (2006), are used.

The current methods present the following limitations:

(a) Methods of normalising variables are either stationary (e.g. *Cai et al.*, 2007, 2008) or non-stationary. However, they focus on the centre of the data distribution, generally using the non-stationary mean and standard deviation for normalisation (e.g. *Guedes Soares et al.*, 1996; *Athanassoulis and Stefanakos*, 1995).

(b) Parametric time dependence models are linear (e.g *Guedes Soares et al.*, 1996), piecewise linear (e.g. *Scotto and Guedes Soares*, 2000), or non-linear but are limited to the extremes (e.g. *Smith et al.*, 1997).

(c) Generally speaking, the simulation is only evaluated using the mean, the standard deviation and the autocorrelation.

This article proposes a simulation method for non-stationary univariate series with time dependence. This method involves the use of a non-stationary parametric mixture distribution to model the univariate distribution of the variable and of copulas to model their time dependence.

The rest of this paper is structured in three sections and seven annexes. In Section 4.4, the proposed model is presented together with the procedure for simulating new time series. In Section 3.4, the model parameters are fitted to a data series of significant wave heights, new series are simulated and the results obtained are discussed. Finally, in Section 3.5, the conclusions are summarised. The derivation of the equations associated with the presented model is illustrated in the appendices at the end of the paper, along with a list of the abbreviations used throughout the paper (Appendix 3.A.7).

3.3 Methodology

The non-stationary model (Section 3.3.1) includes variations of the order of months to years. Because it is a mixture distribution, it can be used to model both medium and extreme generation processes; i.e. this distribution is able to accurately model medium (or main-mass) states and extreme (or tails) states. The time dependence model (Section 3.3.2) models processes whose time scale is composed of various states. Because it is copula-based, this model makes it possible to use various non-linear dependence structures that can be either symmetrical or asymmetrical.

This section also describes the method used to simulate new data series (Section 3.3.3) and the structure of the ARMA models (Section 3.3.4), which are used to compare the results obtained with those obtained using the copula-based time-dependence model.

3.3.1 Non-stationary distribution function

S. Solari (Simulation of time series of geophysical variables; application to harbor engineering (in Spanish), doctoral thesis, University of Granada, Spain, submitted 2011) present a mixture model

$$f(x) = \begin{cases} f_m(x)F_c(u_1) & x < u_1 \\ f_c(x) & u_1 \leq x \leq u_2 \\ f_M(x)\left(1 - F_c(u_2)\right) & x > u_2 \end{cases} \tag{3.1}$$

where F_c is the log-normal distribution (LN), F_m is the GPD of minima, and F_M is the GPD of maxima. When continuity is imposed to the probability density function and the lower bound of the GPD has a value of zero, the GPD distributions are

$$f_m(x|x < u_1) = \frac{1}{\sigma_1}\left(1 - \frac{\xi_1}{\sigma_1}(x - u_1)\right)^{-\frac{1}{\xi_1}-1} \quad \xi_1 \neq 0 \tag{3.2a}$$

$$f_M(x|x > u_2) = \frac{1}{\sigma_2}\left(1 + \frac{\xi_2}{\sigma_2}(x - u_2)\right)^{-\frac{1}{\xi_2}-1} \quad \xi_2 \neq 0 \tag{3.2b}$$

with

$$\sigma_1 = -\xi_1 u_1 \qquad \xi_1 = -\frac{F_c(u_1)}{u_1 f_c(u_1)} \qquad \sigma_2 = \frac{1 - F_c(u_2)}{f_c(u_2)} \tag{3.3}$$

This model is similar to that proposed by *Cai et al.* (2007) for ARMA models with the exception that in (3.1), the continuity of the probability density function is assured by the conditions presented in (3.3). Furthermore, *Cai et al.* (2007) do not provide a method of threshold estimation, whereas Solari (submitted thesis, 2011) show that the threshold can be estimated simultaneously with the other parameters.

The five parameters of the model are (μ_{LN}, σ_{LN}, ξ_2, u_1, u_2). To represent annual variations or those of a shorter duration, the parameters (μ_{LN}, σ_{LN}, ξ_2) are approximated using a Fourier series whose main time period is the year:

$$\theta(t) = \theta_{a0} + \sum_{k=1}^{N} \left(\theta_{ak}\cos(2\pi kt) + \theta_{bk}\sin(2\pi kt)\right) \tag{3.4}$$

where t is the time measured in years (see e.g. *Coles*, 2001; *Méndez et al.*, 2006).

The parameters u_1 and u_2 are replaced by Z_1 and Z_2, using $F_c(u_1) = \Phi(Z_1)$ and $F_c(u_2) = \Phi(Z_2)$, where Φ is the standard normal distribution and Z_1 and Z_2 are stationary parameters. However, because the parameters μ_{LN} and σ_{LN} of the central distribution F_c are non-stationary, the thresholds u_1 and u_2 are non-stationary as well.

The distribution parameters are derived using maximum likelihood estimation, minimising the negative log-likelihood function (NLLF) after the redistribution of the data Solari (submitted thesis, 2011). Redistribution involves taking the original data, truncated with precision $0.1\,m$, and distributing them uniformly at symmetrical intervals $(X - 0.05, X + 0.05)$.

The parameters are estimated by progressively increasing the order of approximation of the Fourier series. The parameters obtained for order n ($\theta_{a\,0}$, $\theta_{a\,1}, \theta_{b\,1}, \ldots, \theta_{a\,n}, \theta_{b\,n}$) are the first approximation used to estimate those in order $n+1$, with zero used as the first approximation of the new parameters $(\theta_{a\,n+1}\ ,\ \theta_{b\,n+1}) = (0, 0)$.

To evaluate the significance of the improvement in fit obtained when the order of the Fourier series is increased, the Bayesian Information Criterion $BIC = -2\log(L) + \log(N_d)p$ is used (see e.g. *Fan and Yao*, 2005) where L is the likelihood function, N_d is the number of available observations, and p is the number of model parameters.

Interannual variation (i.e., long-term cycles of over a year) and variation due to covariables (e.g., climatic indices) are incorporated in the distribution function in a manner similar to the way in which seasonal variation is incorporated (see e.g. *Coles*, 2001; *Izaguirre et al.*, 2010). For parameter θ, a series of covariables $C_i(t)$, and interannual variation of period T_j,

$$\begin{aligned}\theta =& \theta_{a0} + \sum_{k=1}^{N_k} \left(\theta_{ak}\cos(2\pi k t) + \theta_{bk}\sin(2\pi k t)\right) + \\ & \sum_{j=1}^{N_j} \left(\theta_{aj}\cos(2\pi t/T_j) + \theta_{bj}\sin(2\pi t/T_j)\right) + \sum_{i=1}^{N_i} f(C_j(t), t)\end{aligned}$$

where long-term trends and other non-cyclic components are included as particular cases of the functions $f(C_j(t), t)$ in which there is no dependence on any covariable.

Once these parameters are estimated, the accumulated probability function for the time period $(t, t+T)$ is calculated as

$$P(H \le H^*) = \frac{1}{T}\int_t^{t+T} P(H \le H^*|t)dt \tag{3.5}$$

where $P(H \le H^*|t)$ is the non-stationary LN-GPD model (3.1) (NS-LN-GPD):

$$P(x_t|t) = \begin{cases} F_m(x_t|t)\Phi(Z_1) & x_t < u_1(t) \\ F_c(x_t|t) & u_1(t) \le x_t \le u_2(t) \\ \Phi(Z_2) + F_M(x_t|t)\left(1 - \Phi(Z_2)\right) & x_t > u_2(t) \end{cases} \tag{3.6}$$

Goodness-of-fit is evaluated using PP and QQ graphs constructed by standardising the variable x_t following the procedure described in Appendix 3.A.1.

3.3.2 Temporal dependence

The NS-LN-GPD model (3.6) can be used to transform the non-stationary series of significant wave heights $\{H_s(t)\}$ into the uniformly distributed stationary series $\{P(t)\} \sim \mathcal{U}(0,1)$ using $P(t) = Prob[H \leq H_s(t) \,|\, t]$. Next, copula theory is used to model the joint distribution of k successive states $(P_t, P_{t-1}, ..., P_{t-k+1})$. For an introduction to copula theory, see *Joe* (1997); *Nelsen* (2006); *Salvadori et al.* (2007). The use of copulas to model Markov chains is demonstrated in *Abegaz and Naik-Nimbalkar* (2008a,b). *Stefanakos* (1999); *Serinaldi and Grimaldi* (2007); *DeMichele et al.* (2007); *Nai et al.* (2004); *de Waal et al.* (2007) apply copula theory to marine climate and other met-ocean variables.

First, the time dependence between two consecutive states is studied. The joint probability $Prob(P_t, P_{t-1})$ is represented by copula C_{12} such that

$$C_{12}(u,v) = Prob[P_t \leq u, P_{t-1} \leq v] \tag{3.7}$$

On this basis, the conditioned probability function is obtained. This function defines the distribution of P_t given P_{t-1} (or vice versa) and thus defines the first-order Markov process:

$$C_{1|2}(u,v) = Prob[P_t \leq u \,|\, P_{t-1} = v] = \frac{\partial C_{12}}{\partial v}(u,v) \tag{3.8}$$

To define a model of a higher order than 1, a copula construction process is used (*Joe*, 1997, chap. 4.5).

Given copula $C_{1...k}$ (which defines the joint probability of k successive states) and, consequently, given the Markov model of order $k-1$, variables $F_{1|2...k} = Prob[P_t|P_{t-1}, \ldots, P_{t-k+1}]$ and $F_{k+1|2...k} = Prob[P_{t-k}|P_{t-1}, \ldots, P_{t-k+1}]$ are constructed. The dependence between two variables is measured using Kendall's τ_k or Spearman's ρ_s statistic (see Appendix 3.A.3). If this dependence is significant, then there is a relationship of dependence between P_t and P_{t-k} that cannot be explained by the Markov model of order $k-1$. In this case, it is necessary to construct a k-order Markov model. This can be accomplished using copula $C_{1...k+1}$

$$\begin{aligned} C_{1...k+1}(u_1, ..., u_{k+1}) =& Prob[P_t \leq u_1, ..., P_{t-k} \leq u_{k+1}] \\ =& \int_{-\infty}^{u_2} ... \int_{-\infty}^{u_k} C_{1k+1}(F_{1|2...k}, F_{k+1|2...k}) \\ & C_{2...k}(dx_2, ..., dx_k) \end{aligned} \tag{3.9}$$

where C_{1k+1} is a bivariate copula fit to the variables $F_{1|2...k}$ and $F_{k+1|2...k}$. This procedure is repeated until the value of k at which the dependence between variables $F_{1|2...k}$ and $F_{k+1|2...k}$ is not significant.

The procedure described is used to define multivariate copulas (i.e., those higher than the second order) based on a set of bivariate (i.e., second-order) copulas. Appendix 3.A.4 describes how this procedure is used to construct copula C_{1234}, which defines a third-order Markov process.

An alternative procedure that has not been implemented in this study involves using the autocorrelation function of the variable x_t to set the order of the process k as the maximum time lag for which the autocorrelation is significant. Then, the copula construction method described above can be used to construct the multivariate copula $C_{1\ldots k}$.

Copulas families used

This research tested different copula families for the data used. The families selected were those that had the best goodness-of-fit based on the value of their likelihood functions and based on a visual evaluation. The two copula families used in this study were an asymmetric version of the Gumbel-Hougaard family and the Fréchet family (Appendix 3.A.5). A list of copula families, their characteristics, and the different ways to fit them to the data can be found in *Joe* (1997); *Nelsen* (2006); *Salvadori et al.* (2007); *Jaworski et al.* (2010). For a summary of methods and goodness-of-fit tests, see *Genest and Favre* (2007) and references therein.

3.3.3 Simulation methodology

The simulation process consists of two parts. First, the time-dependence model of copulas (3.9) is used to obtain the series of probabilities $\{P_t\}$; then, the non-stationary model (3.1) is used to transform the probabilities into wave heights. To simulate the realisation P_t of the Markov process of order $k-1$, once the previous realisations P_{t-1} to P_{t-k+1} are known, $u_t \sim \mathcal{U}(0,1)$ is simulated and P_t obtained, resolving the following equation

$$\begin{aligned} u_t =& \frac{\partial C_{1\ldots k}}{\partial u_2 \ldots \partial u_k}(P_t, \ldots, P_{t-k+1}) \\ =& \frac{\partial C_{1k}}{\partial F_{k|2\ldots k-1}}\Big(F_{1|2\ldots k-1}(P_t, \ldots, P_{t-k+2}), \\ & \qquad F_{k|2\ldots k-1}(P_{t-1}, \ldots, P_{t-k+1})\Big) \end{aligned} \tag{3.10}$$

where C_{1k} is the bivariate copula fit to $F_{1|2\ldots k-1}$ and $F_{k|2\ldots k-1}$ to construct $C_{1\ldots k}$ and where $F_{1|2\ldots k-1}$ and $F_{k|2\ldots k-1}$ are calculated using the set of bivariate copulas $C_{1k-1}, C_{1k-2}, \ldots, C_{12}$.

When this procedure is used, it is not necessary to use (3.9) to perform the simulations because (3.10) can be resolved using the bivariate copulas. To obtain P_t, equation (3.10) can be numerically solved using the bisection method. The simulation process for a third-order Markov model is described in Appendix 3.A.6.

3.3.4 ARMA models

An ARMA(p,q) model is given by

$$Z_t = \phi_1 Z_{t-1} + \ldots + \phi_p Z_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} \tag{3.11}$$

where ϕ and θ are the coefficients of the autoregressive component and of the moving average, respectively, and ε_t stands for the independent, identically distributed realisations with a null mean and variance σ_ε^2 (a normal distribution is generally assumed). The AR(p) model corresponds to the ARMA(p,0) case.

To estimate the parameters of the ARMA model, the probability series $\{P_t\}$, obtained using the NS-LN-GPD model (3.6), is transformed into a series $\{Z_t\}$ via the inverse of the standard normal distribution. Once $\{Z_t\}$ has been obtained, the parameters ϕ, θ and σ_ε^2 can be estimated using maximum likelihood estimation.

Once the model (3.11) is fitted, white noise is generated with variance σ_ε^2, and a new series $\{Z_t\}$ is simulated using parameters ϕ and θ. After the series $\{Z_t\}$ has been simulated, it is transformed into $\{P_t\}$ using a standard normal distribution and afterwards into $\{H_s\}$ using the inverse of the NS-LN-GPD model (3.6).

3.4 Application

The research study described in this article used a series of 36,496 data records of spectral significant wave height from 13 years and 3 months of sea states with a duration of 3 hours (although there were some gaps in the record). The data were obtained using the WAM numerical model, provided by Puertos del Estado, Spain (www.puertos.es), corresponding to WANA point number 1054046 (36.5°N, 6.5°W, Gulf of Cádiz, Spain). This is the same data series used by Solari (submitted thesis, 2011).

3.4.1 Non-stationary seasonal distribution

In this section, the NS-LN-GPD parameters are estimated. A non-stationary LN distribution (NS-LN) is also fitted (corresponding to the NS-LN-GPD with Z_1 and Z_2 parameters approaching infinity) for use in testing the goodness of fit obtained using the NS-LN-GPD model.

In the first instance, the parameters are only allowed to have seasonal variations (i.e., variation of periods less than or equal to a year (3.4)); interannual variation, covariables and trends were not considered.

Fourier series are evaluated (3.4) with a maximum order of approximation n between 1 and 12. The order 1 represents annual variation, 2 represents semiannual variation, and so on. For each fit distribution, the BIC is estimated.

The models are identified using three digits $[a\,b\,c]$; a is the order of approximation of the Fourier series used for μ_{LN}, b is the order of approximation of the series used for σ_{LN}, and c is the order of approximation of the series used for ξ_2. When a maximum approximation n is allowed, $a, b, c \leq n$ should hold. The total number of parameters of the model $[a\,b\,c]$ is $2(a+b+c)+5$; i.e., there are $2a+1$ parameters to be used in the Fourier series representation of μ_{LN}, $2b+1$ parameters to be used in the Fourier series

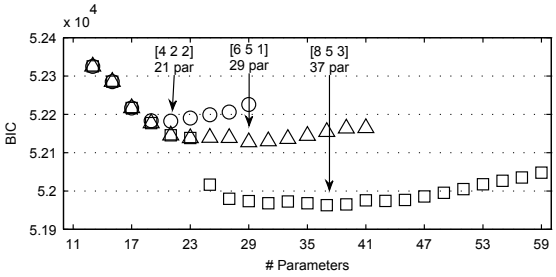


Figure 3.2: Minimum Bayesian Information Criterion obtained for different numbers of parameters in the NS-LN-GPD model, with maximum approximation of the fourth order (◯), 6th order (△) and 9th order (□).

Table 3.2: NS-LN parameters.

	μ		σ	
Ord. (k)	θ_{ak}	θ_{bk}	θ_{ak}	θ_{bk}
0	-0.116	—	0.561	—
1	0.318	0.203	0.100	-0.016
2	-0.024	-0.070	0.021	-0.019
3	0.010	-0.009	-0.004	-0.008
4	0.051	0.001	0.008	0.014

representation of σ_{LN}, $2c+1$ parameters to be used in the Fourier series representation of ξ_2, and the two stationary parameters Z_1 and Z_2.

Figure 3.2 shows the value of the BIC, depending on the total number of parameters when maximum approximations are permitted of order $n = 4, 6, 9$. For each number, only the minimum BIC model is included. The minimum BIC models are identified for each n–order maximum approximation. Although each curve has a relative minimum, the minimum decreases as the maximum allowed order n increases. This finding implies that to use the BIC as a selection criterion for the model, one must first define the maximum allowed order of approximation n.

In this study, the minimum variation period for the parameters has been limited to 3 months. (The maximum allowed order of approximation n is limited to 4.) The minimum BIC model in this case is [4 2 2]: i.e., a Fourier series of order 4 for μ_{LN} and of order 2 for σ_{LN} and ξ_2. Figure 3.3 shows the annual temporal evolution of parameters μ_{LN}, σ_{LN} and ξ_2 from model NS-LN-GPD [4 2 2]. As can be observed, the principal component is the annual period, and the other components provide non-negligible corrections of a lesser order. The only exception is parameter ξ_2, for which the semi-annual component is of the same order of magnitude as the annual one. The fit of the [4 2 2] model obtained using the NS-LN-GPD parameters is compared with that of the model obtained using the NS-LN (also using $n = 4$). Tables 3.2 and 3.3 show the estimated NS-LN-GPD and NS-LN parameters, respectively.

Figure 3.4 shows the quantiles corresponding to the empirical accumulated probability values and those obtained when the NS-LN and NS-LN-GPD models are used. The empirical quantiles have been obtained using a moving window of one month. Gener-

Table 3.3: NS-LN-GPD parameters.

	μ_{LN}		σ_{LN}		ξ_2	
Ord. (k)	θ_{ak}	θ_{bk}	θ_{ak}	θ_{bk}	θ_{ak}	θ_{bk}
0	-0.094	—	0.520	—	-0.006	—
1	0.322	0.199	0.097	-0.019	-0.014	0.076
2	-0.019	-0.073	0.023	-0.012	-0.063	-0.037
3	0.004	-0.011	-	-	-	-
4	0.045	0.004	-	-	-	-
Z_1					Z_2	
-0.734 (23%)					1.078 (86%)	

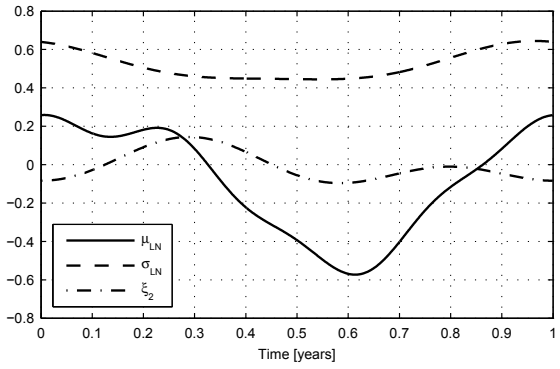


Figure 3.3: Time evolution of μ_{LN}, σ_{LN} and ξ_2 for the NS-LN-GPD [4,2,2] model.

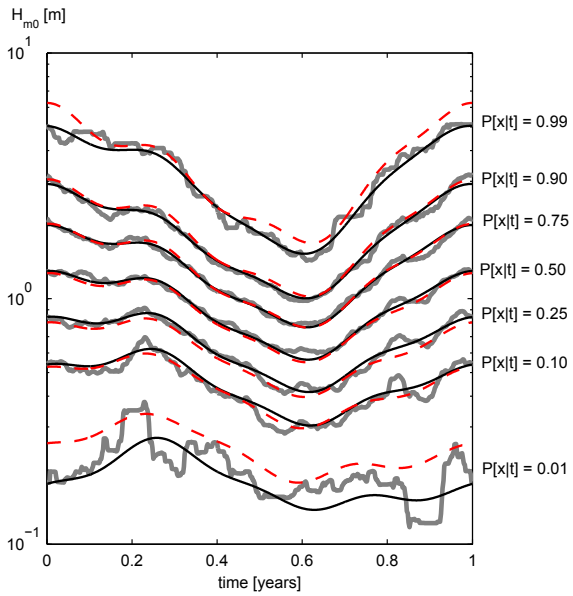


Figure 3.4: Iso-probability quantiles for non-exceeding probability $P[x|t]$ equal to 0.01, 0.1, 0.25 0.5, 0.75, 0.9 and 0.99; empirical (grey continuous line), NS-LN model (red dashed line) and NS-LN-GPD model (black continuous line).

ally speaking, the quantiles calculated using the NS-LN-GPD distribution coincide with the empirical quantiles. As compared with the NS-LN model, the NS-LN-GPD model exhibits superior fit at the tails.

Figure 3.5 (the top graph) shows the annual CDF on log-normal paper. As can be observed, the NS-LN-GPD model exhibits a better fit at the tails than the NS-LN model. Figure 3.5 (the bottom graph) shows the annual PDF. The NS-LN-GPD model fits the mode better than the NS-LN model.

Finally, Figure 3.6 shows the Q-Q and P-P graphs for the two models. These graphs confirm the goodness-of-fit obtained using the NS-LN-GPD model.

3.4.2 Interannual variations

The purpose here is to show how the proposed model can include the interannual variations observed in the series and examine how these interannual variations affect the simulation of new series. The physical basis of the observed interannual variations is not under study here. Moreover, the observed trends are assumed to be cyclical so that the mean value of the long-term simulations is not affected. This also makes it easier to compare the original and simulated series.

It is not our aim to perform an in-depth analysis of the interannual variation in the data series being used; this would mean studying covariables of interest such as the NAO and considering long-term trends and climate cycles, which require longer series than the one available as well as series of covariables (see e.g. *Ruggiero et al.*, 2010; *Izaguirre et al.*, 2010).

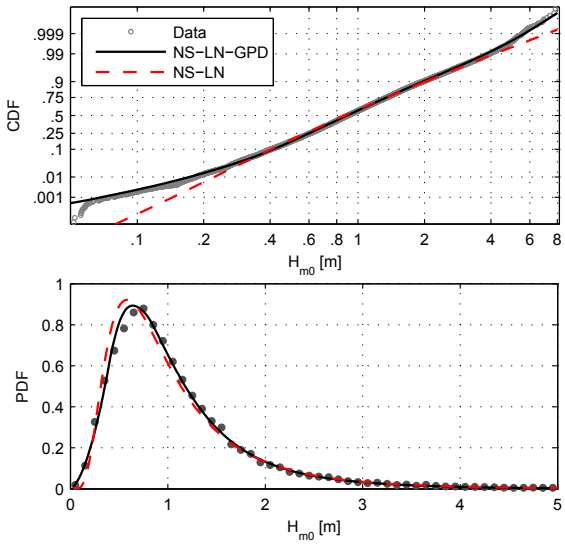


Figure 3.5: Accumulated probability on log-normal paper (top graph) and probability density (bottom graph). Empirical (dots), data from the NS-LN normal model (dashed line), and data from the NS-LN-GPD model (continuous line).

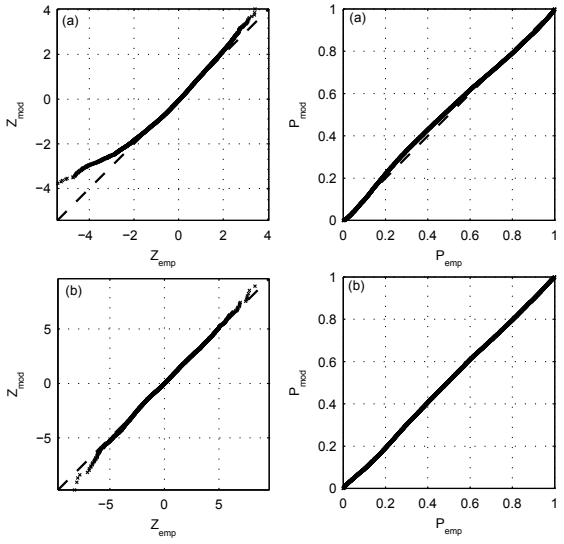


Figure 3.6: Left: Q-Q graph of the non-stationary log normal model (a) and the non-stationary model (b). Right: P-P graph of the non-stationary log normal model (a) and the non-stationary model (b).

When the moving average of the data is displayed on a graph (Figure 3.7), two trends are observed: (i) a cyclical component with a period of approximately 5 years and (ii) a decreasing trend. To analyse both, the following cyclical components are included in the mean:

$$\begin{aligned} \mu_{LN,anual} &= a_{i1}\cos(2\pi t/5) + b_{i1}\sin(2\pi t/5) \\ &\quad + a_{i2}\cos(2\pi t/26) + b_{i2}\sin(2\pi t/26) \end{aligned} \tag{3.12}$$

This is an ad hoc model for long-term trends that assumes that the downward trend in the 13 years of data is part of a 26-year pattern of cyclical variation.

These four parameters and the other parameters of the model are estimated using maximum likelihood estimation with $n = 4$ as the maximum order of approximation for the Fourier series and using the BIC to select the model. The model obtained in this case is [4 2 2 2], where the first three numbers refer to the order of approximation of μ_{LN}, σ_{LN} and ξ_2 and the last refers to the two interannual cyclical components included in μ_{LN} (3.12).

Figure 3.7 shows the moving average of the logarithm of the data obtained using a moving window of 90 days and the mean of NS-LN-GPD model [4 2 2 2]. As can be observed, the μ_{LN} parameter with interannual variation adequately captures the trend in the mean of the logarithm of the data.

Model [4 2 2 2] exhibits a goodness of fit similar to that of model [4 2 2] (as given in Figures 3.5 and 3.6 and therefore not shown here).

3.4.3 Time Dependency. Copulas

To fit the time dependency, different copula families can be tested. In this study, the families with the best fit are selected based on the log-likelihood function (LLF) and a visual evaluation. The following paragraphs describe the data fitting processes, which are conducted based on the probability series $\{P_t\}$ obtained using NS-LN-GPD model [4 2 2 2]. Figure 3.8 shows the mean and standard deviation of P_t as well as their smoothed values on an annual scale. As can be observed, the series may be treated as stationary.

The asymmetric Gumbel-Hougaard copula (3.26) provides a good fit for the time-dependence between P_t and P_{t-1}. The parameters estimated for this copula are $\theta = 5.462$, $\theta_1 = 0.994$ and $\theta_2 = 0.969$. This shows that P_t and P_{t-1} are significantly dependent on each other (high θ) and that the distribution is slightly asymmetrical ($\theta_1 \approx \theta_2$).

Figure 3.9 depicts the empirical function $C(P_t, P_{t-1})$ and that obtained using the asymmetric Gumbel-Hougaard function. It is clear that the modelled and empirical iso-probability curves overlap, except around $P_t \approx P_{t-1} \approx 0.1 - 0.4$, where the data reflect a more marked dependence than that exhibited by the model. In general, the fit is good.

We then estimated the dependence between P_t and P_{t-2}, which was not explained by $C(P_t, P_{t-1})$. For this purpose, the C_{12} copula was used to estimate $F_{1|2}$ and $F_{3|2}$.

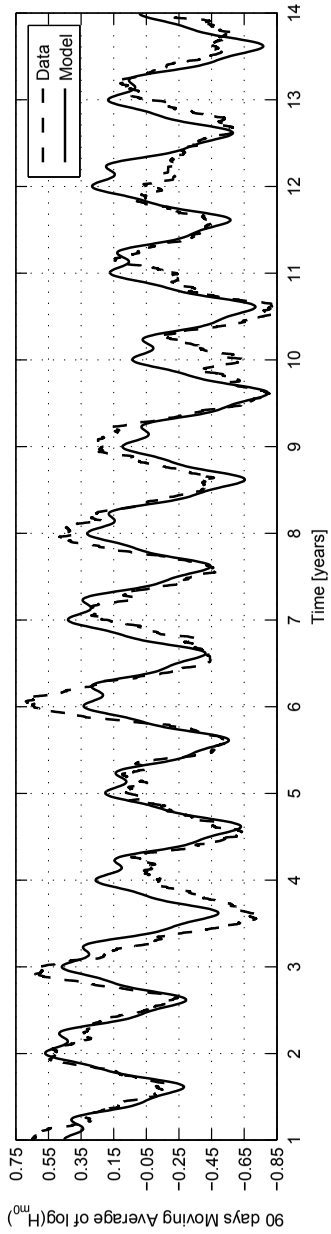


Figure 3.7: Ninety-day Moving Average of H_s and the $\mu_{LN}(t)$ parameter of interannual model.

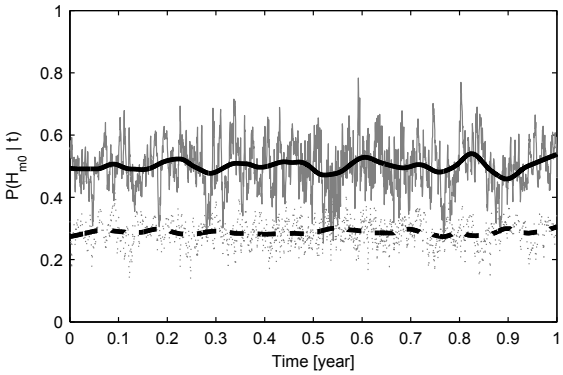


Figure 3.8: Mean and standard deviation of P_t, estimated on an annual scale for each state, and their moving average smooth curves.

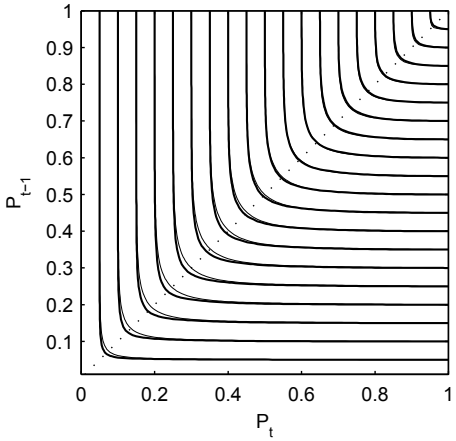


Figure 3.9: Empirical copula $C(P_t, P_{t-1})$ (thick line) and asymmetric Gumbel-Hougaard copula (thin line).

Table 3.4: Copulas parameters fitted using P_t series obtained with the NS-LN-GPD [4 2 2] (SM) and NS-LN-GPD [4 2 2 2] (IM) models.

	C_{12} G-H Asim.			C_{13} Fréchet		C_{14} Fréchet	
	θ	θ_1	θ_2	α	β	α	β
SM	5.697	0.995	0.971	0	0.194	0.005	0
IM	5.462	0.994	0.969	0	0.192	–	–

The dependence between $F_{1|2}$ and $F_{3|2}$ is significant ($\tau_k = -0.133$ and $\rho_s = -0.192$), and thus, the trivariate copula C_{123} was constructed.

To obtain the trivariate copula (3.22), the bivariate copula $C_{13}(F_{1|2}, F_{3|2})$ was fitted. In this case, a good fit was obtained using the Fréchet family. The parameters were fitted using (3.33) and assuming that $\alpha = 0$. A good fit was obtained, although there was some asymmetry in the data that was not captured by the copula.

The copula C_{123} was used to estimate $F_{1|23}$ and $F_{4|23}$. The dependence between these variables was found to be $\tau_k = -1.4 \times 10^{-3}$ and $\rho_s = -1.3 \times 10^{-4}$. Consequently, the variables $F_{1|23}$ and $F_{4|23}$ can be regarded as independent.

Table 3.4 summarises the parameters of the copulas fitted using the probability series $\{P_t\}$ obtained with the NS-LN-GPD [4 2 2] and [4 2 2 2] models (i.e., the seasonal model (SM) and interannual model (IM)). For the SM, the influence of considering the C_{14} copula was not found to be very significant.

3.4.4 Time dependency. ARMA models

High-order AR(p) and ARMA(p,q) models were estimated to compare the results obtained. An optimal number of parameters was not selected; rather a sufficiently high number ($p = q = 23$) was used to take advantage of the capacities of these models. We decided to work with ARMA models because they provided slightly better results than the AR models.

3.4.5 Simulation

A simulation was conducted of 500 years of significant wave height H_s with each of the models fitted to the data: (a) the SM and the dependence model based on copulas (SM-C); (b) the IM and the dependence model based on copulas (IM-C); (c) the SM and the ARMA(23,23) model (SM-A); and (d) the IM and the ARMA(23,23) model (IM-A).

Figure 3.10 shows a five-year data series and another five-year series simulated using the IM-C model. The next step was to evaluate the results obtained using the different models, differentiating between the medium or main-mass regime and the extreme or upper-tail regime.

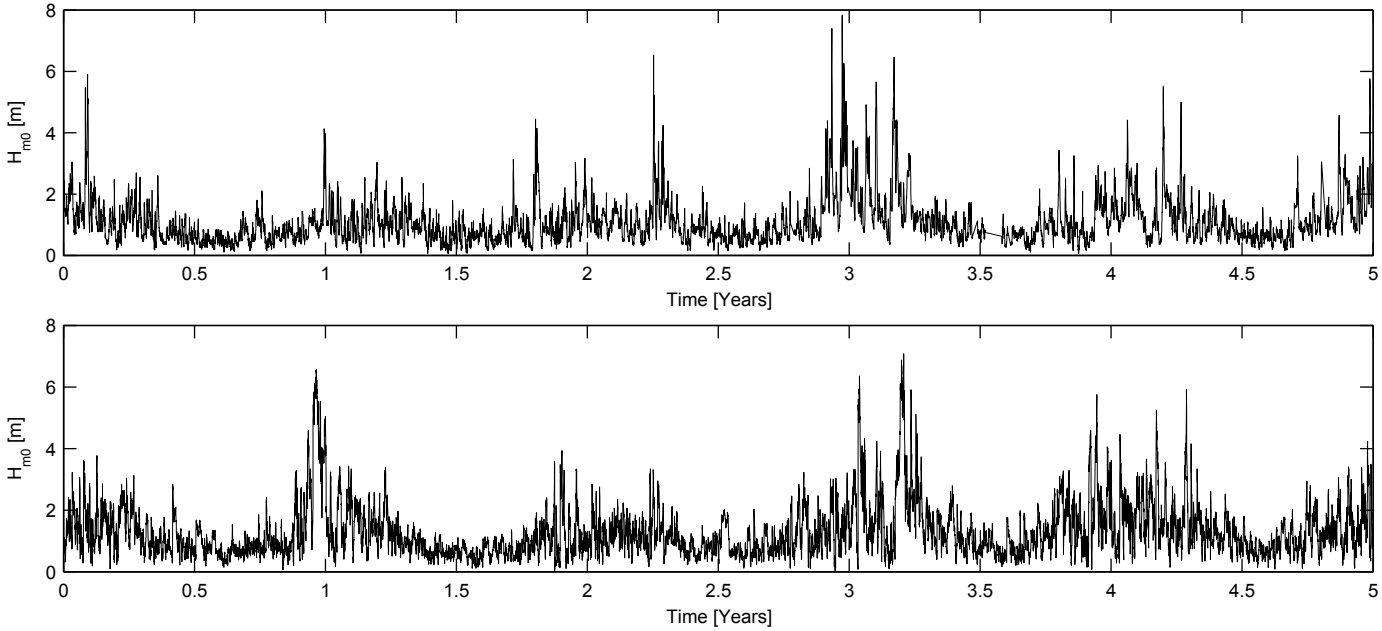


Figure 3.10: Five years of measured significant wave heights (top) and simulated significant wave heights (bottom).

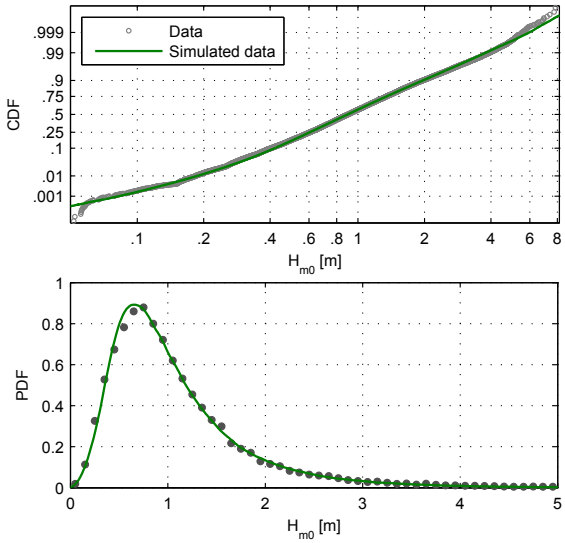


Figure 3.11: Accumulated probability on log-normal paper (top graph) and probability density (bottom graph). Original (dots) and simulated (green line) data series.

Table 3.5: Statistics obtained from the first four central moments.

	Data	SM-C	IM-C	SM-A	IM-A
Mean	1.088	1.077	1.086	1.090	1.093
Variance	0.548	0.521	0.538	0.539	0.556
Skewness	2.127	2.106	2.275	2.159	2.410
Kurtosis	10.006	10.468	12.290	10.846	14.326

Medium or main-mass regime

The medium regime obtained using the four simulated series are very similar. In fact, it is practically impossible to differentiate between the four series in the PDF and CDF plots. Therefore, Figure 3.11 presents the results only for model SM-C. By comparing Figure 3.11 with Figure 3.5 , it is clear that the distribution of the simulated data series (Figure 3.11) is equal to the theoretical distribution (Figure 3.5). This finding is because the simulated series is very long (500 years).

Table 3.5 shows the values of the statistics derived from the first four moments of the distribution: mean, variance, skewness, and kurtosis. As can be observed, all of the models properly represent the mean and the variance. Regarding skewness and kurtosis, the best approximations were obtained using the SM-C and SM-A models. The IM-C and IM-A models yielded overestimated figures for kurtosis, particularly when the ARMA model was used for time dependence.

Figure 3.12 shows the autocorrelation function (ACF) for the data and the four simulated series. For a time lag of less than three days, the SM-C and IM-C models fit the data better than the SM-A and IM-A models. In contrast, for longer time-lags, the SM-A and IM-A models provide a better fit. The main reason for this is that

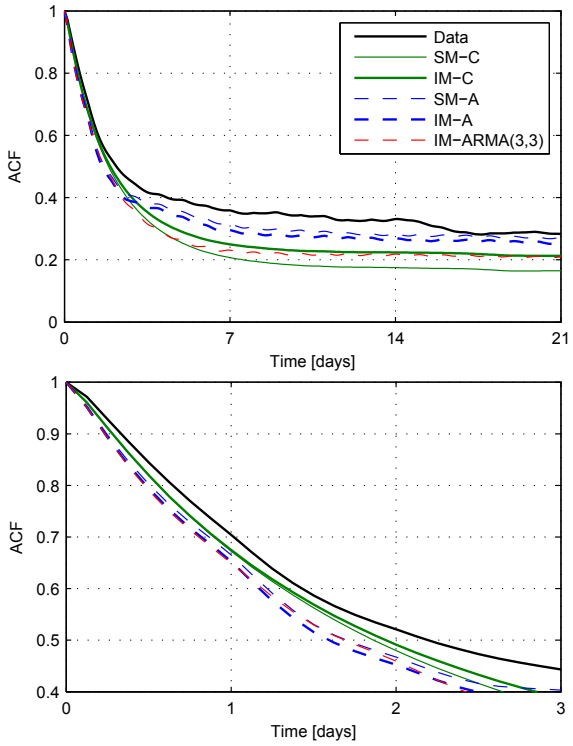


Figure 3.12: Autocorrelation function (ACF) for the four dependence models used and for a simulation run using an ARMA(3,3) model.

the ARMA model is a 23rd-order model, whereas the copula-based models correspond to second-order and third-order Markov models for the IM-C and SM-C, respectively. When third-order ARMA models are used (as indicated by the red dashed line referred to as IM-ARMA (3,3) in Figure 3.12), the long-term fit of the ACF is equivalent to that obtained using copula-based models, whereas the short-term fit is roughly the same as that obtained using a 23rd-order ARMA model.

Figure 3.13 shows the PDF of the persistences over thresholds ($0.5m$, $1.0m$, $1.5m$, $2.0m$, $2.5m$, $3.0m$). In many cases, there are discrepancies between the persistence regimes for the original and simulated data series. For a threshold of $0.5\,m$, the simulated series show a lower than observed frequency of persistence of short duration (6 hours); i.e., the simulations overestimate persistence over $0.5\,m$. For thresholds greater than $2\,m$, the simulations (particularly those obtained using ARMA-based models) show a higher than observed frequency of persistence of short duration (6 hours); i.e., both the copula-based and the ARMA models underestimate persistence, but the extent of the underestimation by the ARMA model is greater. Nevertheless, for thresholds greater than $1.5\,m$, the series obtained using the copula-based models (SM-C and IM-C) show a better fit with regard to the persistence than that obtained using the ARMA model. In contrast, for the thresholds $0.5\,m$ and $1\,m$, the data series simulated using the ARMA model exhibits a better fit with regard to the persistence than the series simulated using

the copula model.

Extreme or upper-tail regime

This study has analysed two aspects of the extreme regime: (i) annual maxima and (ii) storms and peaks over the threshold (POT regime).

Annual maxima

Figure 3.14 shows the annual maxima of the empirical data and of the simulated series for different return periods. Wide dispersion can be observed for high return periods: e.g., for 50-year return period, the values of obtained from the simulated series are between $7.5\,m$ for the model SM-C and more than $10\,m$ for the model IM-A. Generally speaking, the ARMA model has overestimated the annual maxima, whereas the data obtained via the copula-based model are underestimates. Nevertheless, the series simulated using the IM-C model appropriately fit the empirical regime of annual maxima.

Additionally, the effect of including interannual variations (via the IM-C and IM-A models) was to increase the value of H_s for a given return period. This finding occurred independent of the time-dependence model used.

Storms and peaks over threshold (POT)

This study focused on the mean number of storms per year, their distribution throughout an average year, their duration, and the maximum significant wave height reached during the storm (i.e., the POT regime). Storms were identified following Solari (submitted thesis, 2011); the value of the threshold was $u = 3.58\,m$, and the minimum time between the storms was $T_{min} = 2$ days; this minimum time assured that the peaks came from different storms or independent events. The mean number of storms per year based on these data was $\nu = 3.08$. The mean numbers of storms based on the simulated series were $\nu_{\mathrm{SM-C}} = 3.15$, $\nu_{\mathrm{IM-C}} = 3.46$, $\nu_{\mathrm{SM-A}} = 6.16$, and $\nu_{\mathrm{IM-A}} = 6.66$.

Figure 3.15 shows the variation in parameter ν throughout the year. The values were obtained by dividing the year into 24 subsets of 1/2 month each[1], calculating the mean number of storms in each subset, and multiplying them by 24 so that the unit used would be the number of storms per year. The integral of the curve in the year is the mean number of storms per year. The results obtained via the SM-C and IM-C models are within the confidence limits obtained from the original data. In contrast, the results obtained using the SM-A and IM-A models include a significantly greater number of storms than was actually recorded, particularly in the winter.

Figure 3.16 reflects the distribution of storm durations (i.e., persistence exceeding the threshold u). The results obtained via the SM-C and IM-C models were found to provide a slightly better fit of the data than the SM-A and IM-A models, although the four models tended to overestimate the frequency of short durations (approx. 5 hours), and underestimate frequency of long durations (approx. 30 hours).

[1] This two-week time scale corresponds to the variation between spring and neap tides. Even though this was not previously considered, it is another of the variation scales of the system, forced in this case by astronomical phenomena. One might ask if these variations have any effect on the occurrence or intensity of the storms.

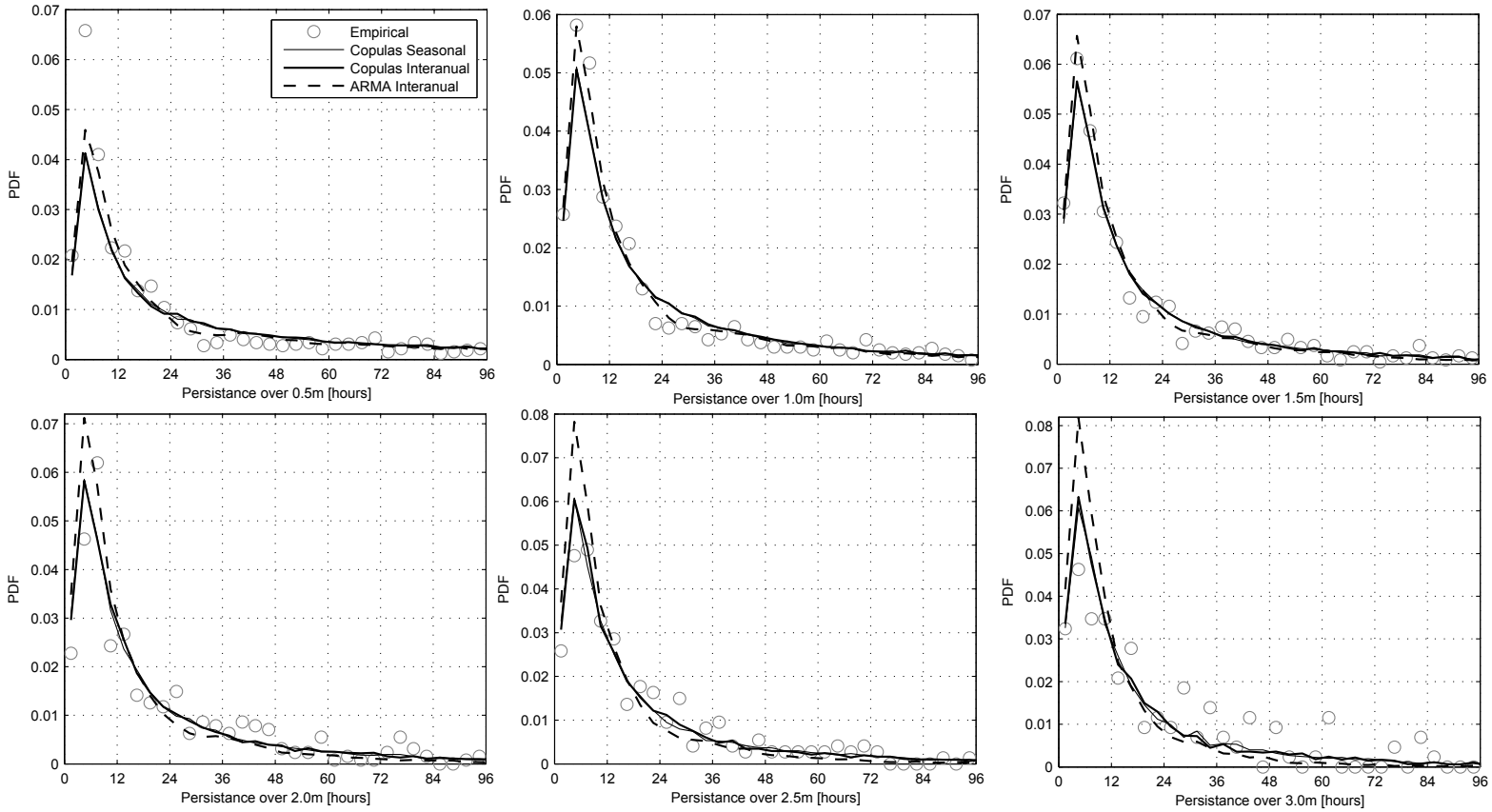


Figure 3.13: Persistence over thresholds 0.5, 1, 1.5, 2, 2.5 and 3m.

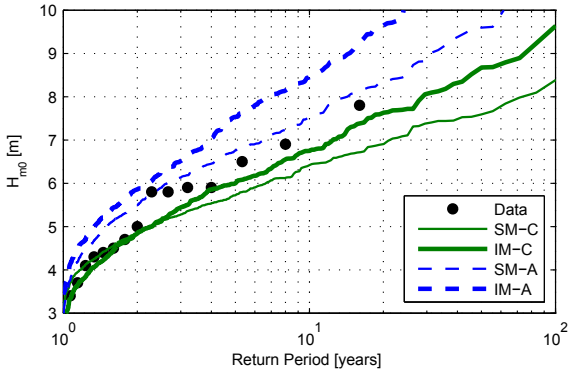


Figure 3.14: Annual maxima H_s: empirical data (dots), data from the copula models (green lines) and data from the ARMA models (blue lines).

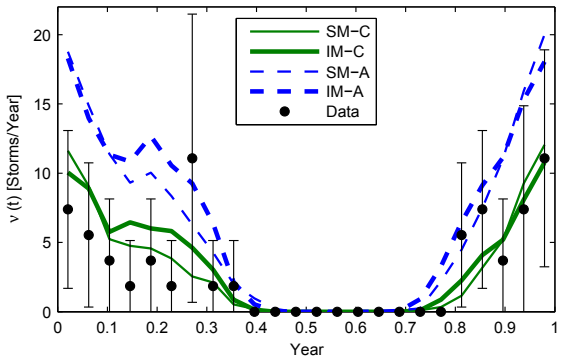


Figure 3.15: Storm occurrence: empirical data with 90% confidence intervals (black lines with dots), data from the copula models (green lines), and data from the ARMA models (blue lines).

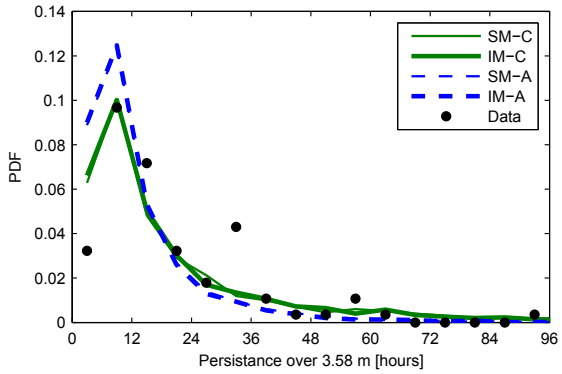


Figure 3.16: Persistence of the storms above 3.58m in days: empirical data (dots), data from the copula models (green lines) and data from the ARMA models (blue lines).

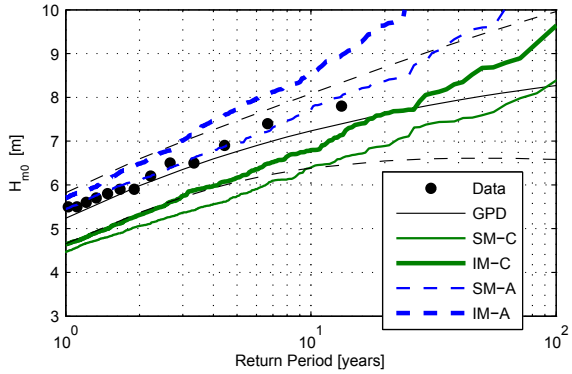


Figure 3.17: POT regime for H_s: empirical data (dots), annual GPD with confidence intervals (grey line), data from the copula models (green lines) and data from the ARMA models (blue lines).

Finally, Figure 3.17 shows the values of H_s corresponding to different return periods as obtained from the POT regime. It also displays the fit of the GPD obtained in Solari (submitted thesis, 2011) for that regime. In this case, the simulated series that best fit the data is that obtained via the SM-A model. In contrast, the series obtained using the IM-A model contains significant overestimates and reflects a long-term tendency that is very different from the tendency indicated by the GPD. On the other hand, although the IM-C model underestimated the data for return periods of less than 10 years, the series obtained exhibit a long-term trend that lies within the GPD confidence limits.

3.4.6 Discussion

With regard to the marginal distribution, all of the simulated series have approximated the original data quite well. The differences between the models become evident when the autocorrelation and persistence regimes are analysed. As compared to the ARMA model, the copula-based time-dependency model provides a better fit to persistence data for thresholds higher than $1\,m$.

With respect to autocorrelation, it appears that in the long term (with time-lags longer than 3 days), the high-order autoregressive models (23) provide better fitting data than do the models based on copulas. However, when low-order autoregressive models (of order 3) are used, the long-term behaviour of the autocorrelation is similar to that obtained using copula-based models (which are also low-order models). If only short-term behaviour is considered (with a time lag of less than 3 days), the copula-based models show a slightly better fit in terms of autocorrelation than that obtained using autoregressive models.

For the extreme regime, the IM-C model provided the best fit in every way. The exception was the POT regime, for which the IM-C model provided the second-best fit. The analysis of the extreme values in terms of the return period clearly indicated the effect of including interannual variations in the model. For particular return periods, the series obtained using the IM model include greater values of H_s than those obtained using the SM model. The data from the ARMA-based models indicate that there

was a much larger mean number of storms per year than was actually recorded. The data from these models also underestimate the duration of the storms. In contrast, the results derived using the copula-based models appropriately fit the recorded data regarding the mean number of storms per year, their distribution throughout the year and their duration.

Based on these findings, copula-based models can be deemed more suitable for use than are ARMA-based models given the frequency and persistence of the storms, which are important parameters to consider when studying systems such as beaches or ports. Even though the copula-based model yielded simulated series with characteristics that are very similar to those of the original series, there are certain differences between the series with regard to the POT regime.

The effect of interannual variability is especially evident in the values for the upper tail even though it was only included in the parameters for the mean of the distribution. This is one of the advantages of using an integral model that covers the entire range of values of the variable. Performing a more in-depth analysis of interannual variation by taking into account the effect of covariables could improve the results obtained. Furthermore, it would provide more information regarding the long-term behaviour of the variable.

3.5 Conclusions

This article has described a non-stationary univariate model for the long-term distribution of sea-state variables that is valid for the entire range of values of the variable. The model includes seasonal variation using a Fourier-series approximation of the parameters and can also take into account climate cycles, trends, and covariables.

The results of this study indicate that this non-stationary model can be used to transform the original non-stationary variable ($H_s(t)$ in this article) into a stationary one $P(t) = Prob[H_s < H_s(t)|t]$. Using this variable ($P(t)$), it is possible to study the time dependence or autocorrelation of the original variable (H_s). For this purpose, in this research, a copula-based model was developed based on the assumption that the process being examined was a Markov process.

The application of the models to a data series for hindcast significant wave height indicated that the simulations obtained via the copula-based time-dependence model were better than those obtained using an ARMA model. However, some related considerations require further study. The long-term autocorrelation data generated by the copula-based models (with time-lags larger than 3 days) is inferior to that obtained using the high-order ARMA models. The possibility of improving these results by using other families of copulas should be investigated. It will also be necessary to more rigorously study how including long-period variation and covariables in the non-stationary model influences the simulated series.

This study has shown that from an engineering viewpoint, it is not appropriate to evaluate simulation methods exclusively in terms of the ACF of the simulated series. A good ACF fit does not ensure that the model will behave suitably in representing

persistence regimes, storm regimes and annual maxima.

3.A Appendixes

3.A.1 Data standardization

To build the PP and QQ plots of the NS-LN-GPD model, the standardized variable Z_e is used.

$$Z_e = \begin{cases} Z1 - Z_{min} & H(t) < u_1(t) \\ Z_{LN} & u_1(t) \leq H(t) \leq u_2(t) \\ Z_2 + Z_{max} & H(t) > u_2(t) \end{cases} \tag{3.13}$$

where Z_1 and Z_2 are the parameters of the model; u_1 and u_2 are the thresholds calculated with the model; and Z_{LN}, Z_{min} and Z_{max} are calculated as

$$Z_{LN} = \frac{\log(H(t)) - \mu_{LN}(t)}{\sigma_{LN}(t)} \tag{3.14}$$

$$Z_{min} = \frac{1}{\xi_1(t)} \log\left(1 - \frac{\xi_1(t)}{\sigma_1(t)} \left(H(t) - u_1(t)\right)\right) \tag{3.15}$$

$$Z_{max} = \frac{1}{\xi_2(t)} \log\left(1 + \frac{\xi_2(t)}{\sigma_2(t)} \left(H(t) - u_2(t)\right)\right) \tag{3.16}$$

This takes into account that when $H(t)$ has a log-normal distribution, Z_{LN} has a standard normal distribution; and when $H(t)$ has a GPD distribution of minima (maxima), Z_{min} (Z_{max}) has a unit-parameter exponential distribution.

After calculating the standardized variable Z_e this variable was used to calculate empirical probability P_e. The modeled values of Z_m quantiles and of probability P_m were calculated from Z_e and P_e as

$$Z_m(P_e) = \begin{cases} Z_1 + \log(P_e/\Phi(Z_1)) & P_e < \Phi(Z_1) \\ \Phi^{-1}(P_e) & \Phi(Z_1) \leq P_e \leq \Phi(Z_2) \\ Z_2 - \log(1 - \frac{P_e - \Phi(Z_2)}{1 - \Phi(Z_2)}) & P_e > \Phi(Z_2) \end{cases} \tag{3.17}$$

$$P_m(Z_e) = \begin{cases} \Phi(Z_1)\exp(Z_e - Z_1) & Z_e < Z_1 \\ \Phi(Z_e) & Z_1 \leq Z_e \leq Z_2 \\ \Phi(Z_2) + (1 - \Phi(Z_2))(1 - \exp(Z_2 - Z_e)) & Z_e > Z_2 \end{cases} \tag{3.18}$$

Finally, graph QQ was built with (Z_e, Z_m) and graph PP was built with (P_e, P_m).

3.A.2 Copula definition

A copula is a function $C : [0,1] \times [0,1] \to [0,1]$ such that for all $u, v \in [0,1]$, it holds that $C(u,0) = 0$, $C(u,1) = u$, $C(0,v) = 0$ and $C(1,v) = v$; and for all $u_1 \leq u_2, v_1 \leq v_2 \in [0,1]$ it holds that

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0$$

The use of copulas to define multivariate distribution functions is based on the Sklar's theorem: when F_{XY} is a two-dimensional distribution function with marginal distribution functions F_X y F_Y, there is then a copula C such that $F_{XY} = Prob[X \leq x, Y \leq y] = C(F_X(x), F_Y(y))$.

3.A.3 Measures of association

For a bivariate series (x, y). , the most widely used measurements of association are Kendall's τ_k and Spearman's ρ_s (*Salvadori et al.*, 2007). A sample version of these parameters are

$$\tau_k = \frac{c - d}{c + d} \tag{3.19}$$

$$\rho_s = 1 - \frac{6 \sum_{i=1}^{n} (R_i - S_i)^2}{n^3 - n} \tag{3.20}$$

where c (d) are the number of concordant (discordant) pairs (x_i, y_i) (x_j, y_j),defined as $(x_i - x_j)(y_i - y_j) < 0$ (> 0); $R_i = Rank(x_i)$; $S_i = Rank(y_i)$; n is the sample size.

3.A.4 Copula-based second-order and third-order Markov Models

Variables $F_{1|2}$ and $F_{3|2}$ are calculated using the bivariate copula C_{12} that defines the first-order Markov process:

$$F_{1|2}(u, v) = Prob[P_t \leq u \,|\, P_{t-1} = v] = \frac{\partial C_{12}}{\partial v}(u, v) \tag{3.21a}$$

$$F_{3|2}(v, w) = Prob[P_{t-2} \leq w \,|\, P_{t-1} = v] = \frac{\partial C_{23}}{\partial v}(v, w) \tag{3.21b}$$

Where it is assumed that the time-dependence structure is stationary, and thus $C_{12} \equiv C_{23}$.

If these variables are dependent on each other (a dependence measured with τ_k or ρ_s), a trivariate copula C_{123} is then built that contemplates this dependence and which defines the second-order Markov process

$$C_{123}(u, v, w) = Prob[P_t \leq u, P_{t+1} \leq v, P_{t+2} \leq w]$$

Where marginal distributions C_{12} and C_{23} are given by the copula $C_{12} \equiv C_{23}$, and where marginal C_{13} represents the dependence of P_t and P_{t-2} that is not explained by C_{12}. A copula of this type can be found in (*Joe*, 1997, chap. 4.5)

$$C_{123}(u,v,w) = \int_{-\infty}^{v} C_{13}(F_{1|2}(u,x), F_{3|2}(x,w))F_2(dx) \tag{3.22}$$

Where C_{13} is fit based on the sample of $F_{1|2}$ and $F_{3|2}$.

Similarly, $F_{1|23}$ and $F_{4|23}$ are calculated using C_{123}

$$\begin{aligned} F_{1|23}(u,v,w) =& Prob[P_t \le u \,|\, P_{t-1} = v, P_{t-2} = w] \\ =& \frac{\partial^2 C_{123}}{\partial v \partial w} \Big/ \frac{\partial^2 C_{23}}{\partial v \partial w} \\ =& \frac{\partial C_{13}}{\partial F_{3|2}}(F_{1|2}(u,v), F_{3|2}(v,w)) \end{aligned} \tag{3.23a}$$

$$\begin{aligned} F_{4|23}(v,w,y) =& Prob[P_{t-3} \le y \,|\, P_{t-1} = v, P_{t-2} = w] \\ =& \frac{\partial^2 C_{123}}{\partial u \partial v} \Big/ \frac{\partial^2 C_{12}}{\partial u \partial v} \\ =& \frac{\partial C_{24}}{\partial F_{2|3}}(F_{2|3}(v,w), F_{4|3}(w,y)) \end{aligned} \tag{3.23b}$$

Where $C_{12} \equiv C_{23} \equiv C_{34}$ and $C_{123} \equiv C_{234}$.

If the dependence between $F_{1|23}$ and $F_{4|23}$, measured with Kendall's τ_k or Spearman's ρ_s, is significant, there is a significant degree of dependence between P_t and P_{t-3} that is not explained by C_{123}, and copula C_{1234} is built, which defines the fourth-order Markov process

$$\begin{aligned} C_{1234}(u,v,w,y) =& Prob[P_t \le u, P_{t+1} \le v, P_{t+2} \le w, P_{t+3} \le y] \\ =& \int_{-\infty}^{w} \int_{-\infty}^{v} C_{14}(F_{1|23}(u,x_1,x_2), F_{4|23}(x_1,x_2,y)) \\ & C_{23}(dx_1, dx_2) \end{aligned} \tag{3.24}$$

Where copula C_{14} is fit, based on the sample of variables $F_{1|23}$ and $F_{4|23}$.

The distribution of P_t conditioned to $P_{t-1} = v$, $P_{t-2} = w$ and $P_{t-3} = y$ is then obtained by deriving (3.24)

$$\begin{aligned} C_{1|234}(u,v,w,y) =& Prob[P_t \le u \,|\, P_{t-1} = v, P_{t-2} = w, P_{t-3} = y] \\ =& \frac{\partial^3 C_{1234}}{\partial v \partial w \partial y} \Big/ \frac{\partial^3 C_{234}}{\partial v \partial w \partial y} \\ =& \frac{\partial C_{14}}{\partial F_{4|23}}(F_{1|23}(u,v,w), F_{4|23}(v,w,y)) \end{aligned} \tag{3.25}$$

3.A.5 Copulas families

The Gumbel-Hougaard family is the same as the logistic family used in the multivariate theory of extremes (see e.g. *Coles*, 2001, chap. 8 or *Salvadori et al.*, 2007, app. C). This study used an asymmetric version of this family (see e.g.: *Ribatet et al.*, 2009).

$$C_{12}(u,v) = Prob[x \leq u, y \leq v] = \exp\{-V(u,v)\} \tag{3.26}$$

with

$$V(u,v) = (1-\theta_1)\hat{u} + (1-\theta_2)\hat{v} + \left[(\theta_1\hat{u})^\theta + (\theta_2\hat{v})^\theta\right]^{1/\theta} \tag{3.27}$$

where $\hat{u} = -\log(u)$ and $\hat{v} = -\log(v)$, $\theta \geq 1$, $0 \leq \theta_1, \theta_2 \leq 1$.

The conditioned distributions are given by

$$\begin{aligned} C_{1|2}(u,v) =& Prob[x \leq u \,|\, y = v] = \frac{\partial C}{\partial v}(u,v) \\ =& \frac{C(u,v)}{v}\left[1 - \theta_2 + \theta_2\left(1 + \frac{(\theta_1\hat{u})^\theta}{(\theta_2\hat{v})^\theta}\right)^{\frac{1}{\theta}-1}\right] \end{aligned} \tag{3.28}$$

$$\begin{aligned} C_{2|1}(u,v) =& Prob[y \leq v \,|\, x = u] = \frac{\partial C}{\partial u}(u,v) \\ =& \frac{C(u,v)}{u}\left[1 - \theta_1 + \theta_1\left(1 + \frac{(\theta_2\hat{v})^\theta}{(\theta_1\hat{u})^\theta}\right)^{\frac{1}{\theta}-1}\right] \end{aligned} \tag{3.29}$$

whereas the density is

$$\begin{aligned} c_{12}(u,v) &= Prob[x = u, y = v] = \frac{\partial^2 C_{12}}{\partial u \partial v}(u,v) \\ &= \frac{C(u,v)}{uv}\Big\{\left[C_{1|2}(u,v)\right]\left[C_{2|1}(u,v)\right] + \\ &\quad \theta_1\theta_2(\theta-1)\left((\theta_1\hat{u})^\theta + (\theta_2\hat{v})^\theta\right)^{\frac{1}{\theta}-2}(\theta_1\theta_2\hat{u}\hat{v})^{\theta-1}\Big\} \end{aligned} \tag{3.30}$$

The parameters of this copula are estimated by means of maximum likelihood using (3.30).

The Fréchet copula family is given by

$$C_{12}(u,v) = \alpha M_2(u,v) + (1-\alpha-\beta)\Pi_2(u,v) + \beta W_2(u,v) \tag{3.31}$$

where $M_2(u,v) = \min(u,v)$is the Fréchet-Hoeffding upper bound; $\Pi_2(u,v) = uv$ is the independent copula; and $W_2(u,v) = \max(u+v-1,0)$ is the Fréchet-Hoeffding lower bound. The following relations are used to fit the parameters of the Fréchet family (*Salvadori et al.*, 2007)

$$\tau_K(\alpha,\beta) = \frac{(\alpha-\beta)(\alpha+\beta+2)}{3} \tag{3.32}$$

$$\rho_S(\alpha,\beta) = \alpha - \beta \tag{3.33}$$

3.A.6 Simulation procedure of the third-order Markov process

For the third-order Markov process., the simulation procedure is:

(i) At $t = 1$, $u_1 \sim \mathcal{U}(0,1)$ is simulated, and $P_1 = u_1$ is taken.

(ii) For $t = 2$, $u_2 \sim \mathcal{U}(0,1)$ is simulated, and P_2 is calculated conditioned to P_1, solving the following equation

$$u_2 = C_{2|1}(P_1, P_2) \tag{3.34}$$

(iii) For $t = 3$, $u_3 \sim \mathcal{U}(0,1)$ is simulated, and P_3 is calculated conditioned to P_1 and P_2, solving the following equation

$$u_3 = C_{3|1}\Big(C_{1|2}(P_1, P_2), C_{3|2}(P_2, P_3)\Big) \tag{3.35}$$

(iv) for $t \geq 4$, $u_t \sim \mathcal{U}(0,1)$ is simulated, and P_t is calculated conditioned to P_{t-1}, P_{t-2} and P_{t-3}, solving the following equation

$$\begin{aligned} u_t = & C_{4|1}\Bigg(C_{1|23}\Big(C_{1|2}(P_{t-3}, P_{t-2}), C_{3|2}(P_{t-2}, P_{t-1})\Big), \\ & C_{4|23}\Big(C_{2|3}(P_{t-2}, P_{t-1}), C_{4|3}(P_{t-1}, P_t)\Big)\Bigg) \end{aligned} \tag{3.36}$$

(v) Once the series $\{P_t\}$is simulated, the series $\{H_t\}$ is constructed, using the inverse of the NS-LN-GPD (3.6).

In steps ***(ii)*** to ***(iv)***, the expressions of the conditioned copulas are analytically resolved, whereas equations (3.34), (3.35) and (3.36) are numerically solved with the bisection method.

3.A.7 List of abbreviations

Table 3.6 lists the abbreviations used throughout the article.

Table 3.6: List of abbreviations.

Abbreviation	Description
BIC	Bayesian Information Criterion
GPD	Generalised Pareto distribution
IM	NS-LN-GPD model fitted to the data allowing the parameters to have interannual variations
IM-A	Combination of IM model for marginal distribution and ARMA model for time dependency
IM-C	Combination of IM model for marginal distribution and copulas-based model for time dependency
LLF	Log-likelihood function
LN	Log-normal distribution
NLLF	Negative log-likelihood function
NS-LN	Non-stationary log-normal distribution
NS-LN-GPD	Non-stationary mixture model composed by a log-normal distribution for the main-mass regime and two generalised Pareto distributions for the tails regimes
SM	NS-LN-GPD model fitted to the data without allowing for interannual variations of the parameters
SM-A	Combination of SM model for marginal distribution and ARMA model for time dependency
SM-C	Combination of SM model for marginal distribution and copulas-based model for time dependency

Acknowledgments

This research was funded by the Spanish Ministry of Education through its postgraduate fellowship program, grant AP2009-03235. Partial funding was also received from the Spanish Ministry of Science and Innovation (research project CTM2009-10520) and the Andalusian Regional Government (research project P09-TEP-4630). The authors also wish to thank Puertos del Estado for providing the wave record data.

References

Abegaz, F., and U. Naik-Nimbalkar (2008a), Modeling statistical dependence of markov chains via copulas models, *Journal of Statistical Planning and Inference*, *138*, 1131–1146, doi:10.1016/j.jspi.2007.04.028.

Abegaz, F., and U. Naik-Nimbalkar (2008b), Dynamic copula-based markov time series, *Communications in Statistics - Theory and Methods*, *37(15)*, 2447–2460, doi:10.1080/03610920801931846.

Athanassoulis, G., and C. Stefanakos (1995), A nonstationary stochastic model for long-term time series of significant wave height, *Journal of Geophysical Research*, *100*(C8), 149–162.

Borgman, L. E., and N. W. Scheffner (1991), Simulation of time sequences of wave

height, period, and direction, *Tech. Rep. TR-DRP-91-2*, Coastal Engineering Research Center, U.S. Army Engineer Waterways Experiment Station, Vicksburg, Miss.

Cai, Y., B. Gouldby, P. Dunning, and P. Hawkes (2007), A simulation method for flood risk variables, in *2nd Institute of Mathematics and its Applications International Conference on Flood Risk Assessment, 4th September 2007 , University of Plymouth, England.*

Cai, Y., B. Gouldby, P. Hawkes, and P. Dunning (2008), Statistical simulation of flood variables: incorporating short-term sequencing, *Journal of Flood Risk Management*, *1*, 3–12.

Callaghan, D., P. Nielsen, A. Short, and R. Ranasinghe (2008), Statistical simulation of wave climate and extreme beach erosion, *Coastal Engineering*, *55*, 375–390.

Coles, S. (2001), *An Introduction to Statistical Modeling of Extreme Values*, Springer Series in Statistics, Springer, Berlin.

de Waal, D., P. van Gelder, and A. Nel (2007), Estimating joint tail probabilities of river discharges through the logistic copula, *Environmetrics*, (May 2006), 621–631, doi:10.1002/env.

DeMichele, C., G. Salvadori, G. Passoni, and R. Velozzi (2007), A multivariate model of sea storms using copulas, *Coastal Engineering*, *54*, 734–751.

Fan, J., and Q. Yao (2005), *Nonlinear Time Series. Nonparametric and Parametric Methods*, Springer Science+Business Media, Inc.

Fawcett, L., and D. Walshaw (2006), Markov chain models for extreme wind speeds, *Environmetrics*, *17*, 795–809.

Genest, C., and A.-C. Favre (2007), Everything you allways wanted to know about copula modeling and were afraid to ask, *Journal of Hydrologic Engineering*, *12*, 347–367.

Guedes Soares, C., and C. Cunha (2000), Bivariate autoregressive models for the time series of sognificant wave height and mean period, *Coastal Engineering*, *40*, 297–311.

Guedes Soares, C., and A. M. Ferreira (1996), Representation of non-stationary time series of significant wave height with autoregessive models, *Probabilistic Engineering Mechanics*, *11*, 139–148.

Guedes Soares, C., A. M. Ferreira, and C. Cunha (1996), Linear models of the time series of significant wave height on the southwest coast of portugal, *Coastal Engineering*, *29*, 149–167.

Izaguirre, C., F. J. Mendez, M. Menendez, A. Luceño, and I. J. Losada (2010), Extreme wave climate variability in southern europe using satellite data, *Journal of Geophysical Research*, *115 (C04009)*, doi:10.1029/2009JC005802.

Jaworski, P., F. Durante, H. Wolfgang, and T. Rychlik (Eds.) (2010), *Copula Theory and Its Applications, Proceeding of the Workshop Held in Warsaw, 25-26 September 2009*, Springer.

Joe, H. (1997), *Multivariate Models and Dependence Concepts*, Monographs on Statistics and Applied Probability 73, 1 ed., Chapman & Hall.

Losada, M. A. (2002), *ROM 0.0: General procedure and requirements in the design of harbor and maritime structures. PART I*, Puertos del Estado, Spain.

Luceño, A., M. Menéndez, and F. Méndez (2006), The effect of temporal dependence on the estimation of the frequency of extreme ocean climate events, *Proccedings of the Royal Society A*, *462*, 1638–1697.

Méndez, F. J., M. Menéndez, A. Luceño, and I. J. Losada (2006), Estimation of the long-term variability of extreme significant wave height using a time-dependent peak over threshold (pot) model, *Journal of Geophysical Research*, *111 (C07024)*, 1–13.

Méndez, F. J., M. Menéndez, A. Luceño, R. Medina, and N. E. Graham (2008), Seasonality and duration in extreme value distributions of significant wave height, *Ocean Engineering*, *35*, 131–138.

Monbet, V., P. Ailliot, and M. Prevosto (2007), Survey of stochastic models for wind and sea state time series, *Probabilistic Engineering Mechanics*, *22*, 113–126.

Nai, J., P. van Gelder, P. Kerssens, Z. Wang, and E. van Beek (2004), Copula approach for flood probability analysis of the huangpu river during barrier closure, in *Proceeding of the 29th International Coastal Engineering Conference. Lisbon, Portugal*, edited by J. McKee Smith, pp. 1591–1603, World Scientific.

Nelsen, R. B. (2006), *An Introduction to Copulas*, Springer Series in Statistics, 2 ed., Springer.

Payo, A., A. Baquerizo, and M. A. Losada (2008), Uncertainty assessment: Application to the shoreline, *Journal of Hydraulic Research*, *46*(Extra Issue 1), 96–104.

Ribatet, M., T. B. M. J. Ouarda, E. Sauquet, and J.-M. Gresillon (2009), Modeling all exceedances above a threshold using an extremal dependence structure: Inference on several flood characteristics, *Water Resources Research*, *45*, doi:10.1029/2007WR006322.

Ruggiero, P., P. D. Komar, and J. C. Allan (2010), Increasing wave heights and extreme value projections: The wave climate of the u.s. pacific northwest, *Coastal Engineering*, doi:10.1016/j.coastaleng.2009.12.005.

Salvadori, G., C. De Michele, N. T. Kottegoda, and R. Rosso (2007), *Extreme in Nature. an Approach Using Copulas*, Water Science and Technology Library 56, 1 ed., Springer.

Scheffner, N. W., and L. E. Borgman (1992), Stochastic time-series representation of wave data, *Journal of Waterway, Port, Coastal, and Ocean Engineering*, *118*(4), 337–351.

Scotto, M., and C. Guedes Soares (2000), Modelling the long-term series of significant wave height with non-linear threshold models, *Coastal Engineering*, *40*, 313–327.

Serinaldi, F., and S. Grimaldi (2007), Fully nested 3-copula: Procedure and application on hydrological data, *Journal of Hydrologic Engineering*, *12*, 420–430.

Smith, R. L., J. A. Tawn, and S. G. Coles (1997), Markov chain models for thereshold exceedances, *Biometrika*, *84*(2), 249–268.

Stefanakos, C. (1999), Nonstationary stochastic modelling of time series with applications to environmental data, Ph.D. thesis, Technical University of Athenas.

Stefanakos, C., and G. Athanassoulis (2001), A unified methodology for the analysis, completion and simulation of nonstationary time series with missing values, with application to wave data, *Applied Ocean Research*, *23*(4), 207–220, doi:10.1016/S0141-1187(01)00017-7.

Stefanakos, C., and G. Athanassoulis (2003), Bivariate stochastic simulation based on nonstationary time series modelling, in *13th International Offshore and Polar Engineering Conference, ISOPE*, vol. 5, pp. 46–50.

Stefanakos, C., G. Athanassoulis, and S. F. Barstow (2006), Time series modeling of significant wave height in multiple scales, combining various sources of data, *Journal of Geophysical Research*, *111*(C10), 1–12, doi:10.1029/2005JC003020.

Stefanakos, C. N., and K. A. Belibassakis (2005), Nonstationary stochastic modelling of multivariate long-term wind and wave data, in *Proceeding of 24th International Conference on Offshore Mechanics and Arctic Engineering (OMAE2005)*, Halkidiki, Greece.

Walton, T. L., and L. E. Borgman (1990), Simulation of nonstationary, non-gaussian water levels on great lakes, *Journal of Waterway, Port, Coastal, and Ocean Engineering*, *116*(6), 664–685.

Chapter 4

On the use of Vector Autoregressive (VAR) and Regime Switching VAR models for the simulation of sea and wind state parameters

4.1 Abstract

The simulation of long (several years) time series of multivariate wave and wind state parameters has many applications in coastal and ocean engineering, including coastal morphology, transportation and energy exploitation studies, amongst others.

In this work the use of vector autorregresive (VAR) and Regime Switching VAR models for the simulation of wave height, period and direction, and wind speed and direction, is studied.

In order to normalize and stationarize the series, non-stationary mixture uni-variate distributions are fitted to the above five variables. Then three different VAR models (one standard model and two regime switching models) are fitted and new time series are simulated. Finally, an in depth analysis of the long term simulations is performed, in order to study its ability to reproduce the behavior of the original series.

It is found that VAR models are able to capture main features of the original series, but they fail in reproducing some of the persistence regimes and some aspects of the bi-variate distributions. On the other hand, although Regime Switching VAR models improve some aspects of the simulations, they produce some unexpected behavior in the correlation of the simulated series.

4.2 Introduction

Simulation of time series of wave and wind parameters has many applications. Some of them are the design and management of harbors and waterways, the study of coastal

morphology and the design of shore protection structures, the design, construction and operation of off-shore structures, etc. Guedes Soares and Cunha (2000); Stefanakos and Belobassakis (2005).

This work focuses on the study of a methodology for the simulation of new time series of the variables that defines the sea and wind states on deep waters, i.e.: spectral significant wave height H_{m0}, peak period T_p, mean wave direction θ_M, wind speed V_W and wind direction θ_W.

For the application and verification of the methodology a hindcasted time series is used. It corresponds to 13 years of 3 hours states, taken at the Gulf of Cádiz (36.5°N, 6.5°W), Spain.

First, the five variables are normalized and stationarized. For this non-stationary parametric marginal distributions functions of the variables are used. Then, for modeling time dependence and interdependence of the normalized variables, the use of three Vector Autoregressive (VAR) models is studied: a standard VAR model, and two Regime Switching VAR models, the Self Exiting Threshold VAR model (TVAR) and the Markov Switching VAR model (MSVAR).

The objective is to study the ability of the different VAR models to reproduce the behavior of the original time series when they are used for long term simulations (several years). For this, simulated time series are compared with original ones in terms of its marginal distributions, its auto- and cross- correlation functions, and its persistence regimes over different thresholds.

The rest of the document is organized as follows. Section 4.3 presents a brief revision of previous work on the use of autoregressive models for the simulation of multivariate met-ocean variables. In sections 4.4 to 4.6 the methodology used for the simulation is introduced as well as the structure of the different models used. Estimation of the parameters of the models for the study case is dealt with in sections 4.7 to 4.9. Once parameters are estimated new time series are simulated. The comparison of the new series with the original ones is done at section 4.10. Finally, a discussion of the results is done at section 4.11, while main conclusions of this work are summarized at section 4.12.

For the sake of readability of this document, many of the graphical results that partially justify the conclusions, as well as some details of the models used, are skipped.

4.3 Background

This works focus on the multivariate simulation of time series of met-ocean variables by means of autoregressive models.

The use of univariate autoregressive (AR) models for the simulation of significant wave height was presented in Guedes Soares and Ferreira (1996) and Guedes Soares et al. (1996). In Scotto and Guedes Soares (2000) the analysis is extended to the use of univariate self exiting threshold autoregressive (TAR) models. More recently Cai et al. (2007) analyze the use of AR models for the simulation of time series of environmental variables.

With regards to multivariate simulation, Guedes Soares and Cunha (2000) use bivariate autoregressive models for the simulation of wave height and peak period time series. Stefanakos and Belobassakis (2005) use a vector autoregressive moving average model to the simulation of wave height, peak period and wind speed, while Cai et al. (2008) use a bivariate AR model for the study of wave heights and storm surges.

Main differences of this work with previous ones are: (a) the analysis of TVAR and MSVAR models for 5-variate met-ocean variables, (b) the method used for normalization of the variables, and (c) the in depth analysis performed on the simulated series. In our work this analysis comprises several aspects not usually covered in previous works: probability distribution of the persistence regimes, marginal bivariate distributions of the normalized as well as of the original variables, and the ability of the model to reproduce the variability actually observed on the climatic variables (i.e. some years are more severe than others). This in depth analysis gives a better idea of the applicability and the limitation of the VAR models for met-ocean variables simulation than that obtained by only comparing autocorrelations and first moments of the original and simulated distributions.

4.4 Methodology

The proposed methodology comprises three steps:

(a) Non-stationary distributions functions and normalization of the variables. For each one of the variables under study a non-stationary distribution function is fitted $V_i(t) \sim F_i(V_i|t)$. Using this function, variables are normalized by means of

$$Z_i(t) = \Phi^{-1}\left(F_i\left(V_i\left(t\right)|t\right)\right) \tag{4.1}$$

where $\Phi(x)$ is the standard normal univariate distribution. Non-stationary functions used in this work are introduced in section 4.5, the fitting of the functions to the data is presented on section 4.7, and the analysis of the normalized variables is shown in section 4.8.

(b) Vector Autoregressive Models (VAR). These models are used to explain the time dependence and the inter-dependence of the normalized variables. First a order p VAR linear model is fitted (VAR(p)). Then, two different regime switching versions are fitted: a Self Exiting Threshold VAR model (TVAR(K_R, p)), and a Markov Switching VAR model (MSVAR(K_R, p)), where K_R is the number of regimes of the models. The structure of the different models is introduced in section 4.6, while the estimation of its parameters is shown in section 4.9.

(c) Simulation. The simulation of new time series comprises two steps. First, one of the VAR models is used for the simulation of a new time series of the normalized variables$\{Z_i\}$. Then normalized variables are transformed to the original ones by means of the non-stationery distributions $V_i = F_i^{-1}\left(\Phi\left(Z_i\right)|t\right)$. The simulated time series obtained with the different models are analyzed on section 4.10.

4.5 Probability distributions

Here the univariate marginal distribution functions used for normalization of the variables are described.

Variables under study are H_{m0}, T_p, θ_M, V_W and θ_W. Four different stationary and non-stationary distribution functions are used for the normalization of these five variables. These distributions are shown next. The parameters of these distributions are estimated through maximum likelihood.

4.5.1 C-GPD Model

The distribution used for H_{m0} and V_W is the same that was used by Solari and Losada (2011c). This distribution consists of a mixture of a truncated central distribution function for the central part, and two generalized Pareto distributions (GPD) for the tails. The distribution is

$$f(x) = \begin{cases} f_m(x)F_c(u_1) & x < u_1 \\ f_c(x) & u_1 \leq x \leq u_2 \\ f_M(x)\left(1 - F_c(u_2)\right) & x > u_2 \end{cases} \tag{4.2}$$

where f_c is the density function selected for the central part, f_m is the lower tail GPD and f_M is the upper tail GPD; u_1 and u_2 are lower and upper thresholds where the central distribution is truncated; $F_c(u_1)$ and $1 - F_c(u_2)$ are scale constants for the lower and upper GPD respectively.

For the density function (4.2) to be continuous and to have lower bound equal to zero, the following relations must be fulfilled Solari and Losada (2011a,b)

$$\sigma_1 = -\xi_1 u_1 \qquad \xi_1 = -\frac{F_c(u_1)}{u_1 f_c(u_1)} \qquad \sigma_2 = \frac{1 - F_c(u_2)}{f_c(u_2)}$$

For modeling wave heights H_{m0} a Log-Normal (LN) distribution is taken for the central function f_c, and the resulting model is called LN-GPD. For wind speeds V_W central distribution f_c is taken to be a biparametric Weibull distribution (WB), and the resulting model is called WB-GPD.

LN distribution has position parameter μ_{LN} and scale parameter $\sigma_{LN} > 0$; the WB distribution has scale parameter $\alpha_{WB} > 0$ and shape parameter $\beta_{WB} > 0$; the minimum GPD f_m has shape parameter ξ_1, scale parameter $\sigma_1 \geq 0$ and position parameter u_1; the maximum GPD f_M has shape parameter ξ_2, scale parameter $\sigma_2 \geq 0$ and position parameter u_2.

In order to simplify the analysis the position parameters of both GPD, u_1 and u_2, are replaced by Z_1 and Z_2, where $u_i = F_c^{-1}(\Phi(Z_i))$, being Φ the univariate standard normal distribution (Solari and Losada, 2011c, see). Then Z_1 and Z_2 are assumed to be constants, while the remaining parameters are time varying.

Parameters of the LN-GPD distribution are (μ_{LN}, σ_{LN}, ξ_2, Z_1, Z_2), while those of the WB-GPD distribution are (α_{LN}, β_{LN}, ξ_2, Z_1, Z_2). Parameters μ_{LN}, σ_{LN}, α_{WB}, β_{WB} and ξ_2 are time varying, and are expressed as Fourier series

$$\theta = \theta_{a0} + \sum_{k=1}^{N_k} (\theta_{ak} \cos(2\pi kt) + \theta_{bk} \sin(2\pi kt)) \tag{4.3}$$

where t is on annual scale, and as a consequence only variations of periods less or equal to one year are taken into account (seasonal variations).

4.5.2 Bi Log-Normal Model

For the peak period T_p a mixture model composed by two LN distributions is defined. The model is

$$f(x) = \alpha f_{LN_1}(x) + (1 - \alpha) f_{LN_2}(x) \tag{4.4}$$

The parameters of the distribution are (μ_1, σ_1, μ_2, σ_2, α), with $\sigma_1, \sigma_2 > 0$ and $0 \leq \alpha \leq 1$. All parameters are time varying and are expressed as Fourier series using (4.3).

4.5.3 Tetra Truncated Normal Model

For both mean wave direction θ_M and wind direction θ_W a mixture of four stationary normal distributions, truncated at 0° and 360°, is used

$$f(x) = \sum_{i=1}^{4} \alpha_i f_{N\,i}(x) \left[F_{N\,i}(360) - F_{N\,i}(0)\right]^{-1} \tag{4.5}$$

where f_N and F_N are the probability density function (PDF) and the cumulative distribution function (CDF) of the normal distribution, and $\sum_{i=1}^{4} \alpha_i = 1$. The distribution has position parameters μ_{Ni} and scale parameters $\sigma_{Ni} > 0$, with $i = 1, 2, 3, 4$, and proportion parameters α_i with $i = 1, 2, 3$, being $\alpha_4 = \sum_{i=1}^{3} \alpha_i$.

Circular distributions for the direction variables will not be considered in this phase.

4.6 Vector Autoregressive models

Autoregressive models give the value of the current observation as a linear function of past observations and a white noise. Vector Autoregressive models are an extension of autoregressive models for multivariate data. Here, three different models are considered. First, the classical Vector Autoregressive model of finite order p (VAR(p)). Secondly two regime switching non-linear models constructed using the classical one: the Self Exciting Threshold Vector Autoregressive model of K_R regimes and order p (TVAR(K_R, p)), and the Markov Switching Vector Autoregressive model of K_R regimes and order p (MSVAR(K_R, p)). These two model are based on the definition of different regimes.

For each regime a VAR(p) model is used, and as a consequence these two models are piecewise linear.

For a description of vector autoregressive models the reader is referred to Lütkepohl (2005).

4.6.1 VAR(p) model

The Vector Autoregressive model of order p is given by (Lütkepohl, 2005, see e.g.).

$$y_t = \nu + \sum_{i=1}^{p} A_i y_{t-i} + u_t \tag{4.6}$$

where $y_t = (y_{1t}, \ldots, y_{Kt})'$ is a vector of dimensions $(K \times 1)$, being K the number of variables; each A_i is a matrix of autoregressive coefficients, of dimensions $(K \times K)$; $\nu = (\nu_1, \ldots, \nu_K)'$ is a vector of dimension $(K \times 1)$ that allows for a non zero mean $E(y_t)$; and $u_t = (u_{1t}, \ldots, u_{Kt})'$ is a $K-$dimensional white noise, also called innovation process or error, that must fulfill $E(u_t) = 0$, $E(u_t u_t') = \Sigma_u$ and $E(u_t u_s') = 0$ for $s \neq t$.

In this work the parameters of the VAR(p) model are estimated through Least Square (see Lütkepohl, 2005, Ch. 3). For this the model is expressed on matrix notation as $Y = BZ + U$, where $Y = (y_1, \ldots, y_T)$, $B = (\nu, A_1, \ldots, A_p)$, $Z_t = [1, y_t, \ldots, y_{t-p+1}]'$, $Z = (Z_0, Z_1, \ldots, Z_{T-1})$ and $U = (u_1, \ldots, u_T)$, being T the number of observations available for estimation. Then, autoregressive parameters are estimated as $\hat{B} = YZ^{-1}(ZZ')^{-1}$; while the covariance matrix Σ_u of the withe noise u_t is estimated through the errors $\hat{U} = Y - \hat{B}Z$ as $\hat{\Sigma}_u = \hat{U}\hat{U}'/(T - Kp + 1)$.

For defining the order p of the model that should be used it is possible to use the Bayesian Information Criteria $BIC = -2LLF + \log(T)N_p$, where LLF is the log likelihood function and N_p is the number of parameters of the model. Procedure is as follows: first model parameters are estimated for a series of orders p; then, LLF and BIC are estimated for each one of the models and the one with the lower BIC is selected as the "optimum" model.

Assuming that the white noise follows a multivariate normal distribution of zero mean and covariance $\hat{\Sigma}_u$, the LLF is

$$LLF = \sum_{t=1}^{T} log\left(f_{MVN}(\hat{u}_t|0, \hat{\Sigma}_u)\right)$$

where $f_{MVN}(\hat{u}_t|0, \hat{\Sigma}_u)$ is the density function of the multivariate normal distribution.

4.6.2 TVAR(K_R, p) model

Threshold VAR models assume that there exists more than one possible regime for the system, and that at each time t the regime is defined by the value taken by the variable z at time $t - d$, where d is the delay. When z is one of the variables of the regression, the model is called Self Exiting. This last case is the one studied here.

The structure of the TVAR(K_R, p) model is

$$y_t = \nu^{(j)} + \sum_{i=1}^{p} A_i^{(j)} y_{t-i} + u_t^{(j)} \text{ if } r_{j-1} < z_{t-d} \leq r_j \tag{4.7}$$

where the set r_j are the thresholds that defines the different regimes.

Once that the number of regimes K_R, the set of thresholds r_j, the delay d, and the variable z used to identify the regimes are all defined, it is possible to estimate the autoregressive parameter and the covariance matrix of each regime in the same way it was done for the VAR model, using the Least Square method.

Here BIC is used for estimation of z, d, r_j. To calculate the BIC of the TVAR model: (a) the LLF is estimated using a multivariate normal distribution for each regime, and (b) the number of parameters N_p includes the parameters of all regimes.

Then, given the number of regimes K_R, for each one of the possible variables z a set of thresholds r_j and a set of delays d are defined. Then, BIC is estimated for each possible model, and the one with the lower BIC is selected. This is repeated with different number of regimes, and again the one with the lower BIC is taken as the "optimum" model. This procedure is similar to that used by Tsay (1998), but in this case BIC is used for selecting the model, while Tsay used the mean square error.

4.6.3 MSVAR(K_R, p) model

In the Markov Switching VAR model it is assumed the existence of an unobserved variable s_t that determines the regime at each time steps, and that this variables follows a discrete Markov process. Again, for each regime j a VAR model is defined. Then, the MSVAR(K_R, p) is defined as

$$y_t = \nu^{(j)} + \sum_{i=1}^{p} A_i^{(j)} y_{t-i} + u_t^{(j)} \text{ if } s_t = j \tag{4.8}$$

where the unobserved variable s_t follows a Markov process with transition matrix P_s, which has to be estimated on basis of the observed variables y_t.

There are two possible approaches for parameter estimation of MSVAR models: the use of maximum likelihood method, through a EM algorithm (see e.g. Hamilton (1990)), or the use of Bayesian estimation procedure, through the use of Markov Chain Monte Carlo (MCMC) methods (see e.g. Albert and Chib (1993); Harris (1999)). In this work the later approach is used.

Again, the BIC is used for the selection of number of regimes K_R and of the order p of the model. However in this case the likelihood function of the joint distribution of the observed and unobserved variables is used for the calculation of the BIC, and as a consequence the BIC obtained here can not be compared with those obtained for the VAR and TVAR models. The joint likelihood function is

$$f(Y_n, s_1, \ldots, s_n, \lambda) = f(Y_n | S_n, \lambda) P(s_1) \prod_{t=2}^{n} P(s_t | s_{t-1}) \tag{4.9}$$

where

$$f(Y_n|S_n,\lambda) = f(Y_r|S_r,\lambda)\prod_{t=r+1}^{n} f(y_t|Y_{t-1},s_t,\lambda) \tag{4.10}$$

with $f(Y_r|S_r,\lambda)$ being the likelihood of the first r observations and $f(y_t|Y_{t-1},s_t,\lambda)$ is the likelihood of the remaining observations conditional to the regimes and the previous observations, $P(s_1)$ is the marginal probability of the regimes and $P(s_t|s_{t-1})$ is the transition probability from s_{t-1} to s_t.

It should be noted that the estimation of the absolute likelihood of the observed variables would require the integration of (4.9) on all the possible realizations of s_t.

4.7 Distributions fitting

The parameters of the marginal distributions of the five variables are obtained through maximum likelihood. For the non-stationary distributions different models are fitted, varying the order of approximation of the Fourier series between 0 and 4. For each model the BIC is estimated, and the one with the lower BIC is selected. The procedure is similar to that used in Solari and Losada (2011c).

Models give a very good fitting for the five variables under study. Figures 4.1 and 4.2 show the model obtained for H_{m0}. In figure 4.1 the annual mean probability density function (PDF) and cumulative distribution function (CDF) are presented. It is noticed that the model fits very well the data, except for the mode of the model, which is approximately $0.1\,m$ smaller than the empirical mode. To verify that the model is able to capture the seasonal behavior of the variable, non-stationary empirical and modeled quantiles are plotted in figure 4.2, where agreement between both is noticeable.

A similar analysis (not shown here) is performed for the other four variables with similar results.

4.8 Normalized data series

By means of (4.1), and using the marginal distributions fitted in the previous section, the data series are normalized, obtaining the normalized variables Z_H, Z_T, $Z_{\theta M}$, Z_V, and $Z_{\theta W}$.

First what can be noticed is that the normalization procedure is actually capable of producing standard normal distributions. Figure 4.3 shows the CDF of the normalized variables on normal probability paper. It is observed that they follow a standard normal distribution expect for probabilities lower than 0.1% or higher than 99.9%.

Figure 4.4 shows the time series of the normalized variable Z_H. It is noticed that, although there are no seasonal variations in the normalized time series (middle and bottom panels show the variable and the moving average of the mean and the standard deviation in annual scale), there are some important inter-annual variations that the

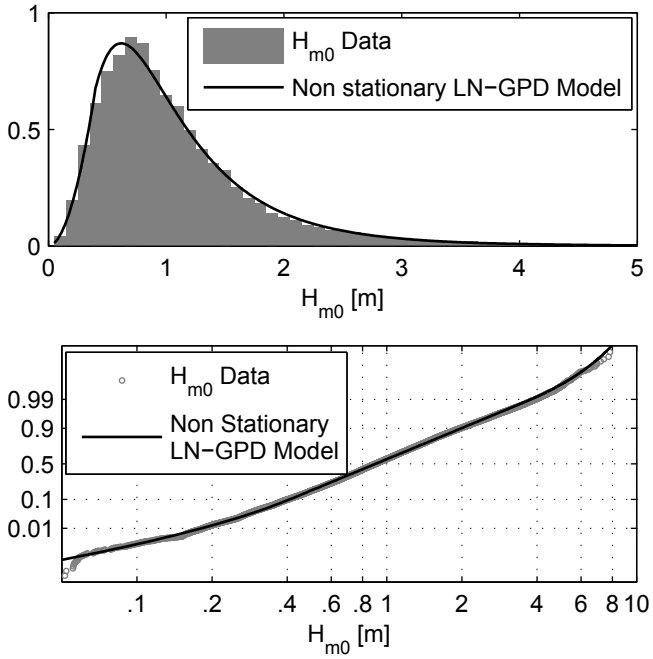


Figure 4.1: Annual probability density functions PDF (top) and cumulative distribution function CDF (bottom) for H_{m0}.

transformation is not able to cope with. For example, the first two years of the series show higher values than the others (top panel).

A complete analysis of the time series of the five normalized variables was also performed. Stationarity of the mean, the standard deviation and the auto- and cross-correlation were studied using the methodology described in van Gelder et al. (2007).

In general terms no seasonal variations are observed in the mean and the standard deviation of the variables Z_H, Z_T are Z_V, i.e. they can be assumed weakly stationary. However, some seasonal variations are observed in the time series of $Z_{\theta M}$ and $Z_{\theta W}$. These variations are not considered significant and are overlooked. However, they could be avoided by using a non-stationary version of the tetra truncated normal model (4.5).

With regards to correlations, it is noticed that autocorrelations of variables Z_H and Z_T show seasonality, while crosscorrelations (Z_H, Z_T), (Z_H, Z_V) and (Z_T, Z_V) show some non-stationary behavior but no clear seasonal pattern can be recognized.

It is concluded that the marginal non-stationary distributions can be used for the transformation of the original non-stationary variables into standard normal weakly stationary variables. However, there is non-stationarity remaining in the time dependence structure of the normalized variables that may justify the use of time varying models when modeling the time-dependence structure of the normalized variables.

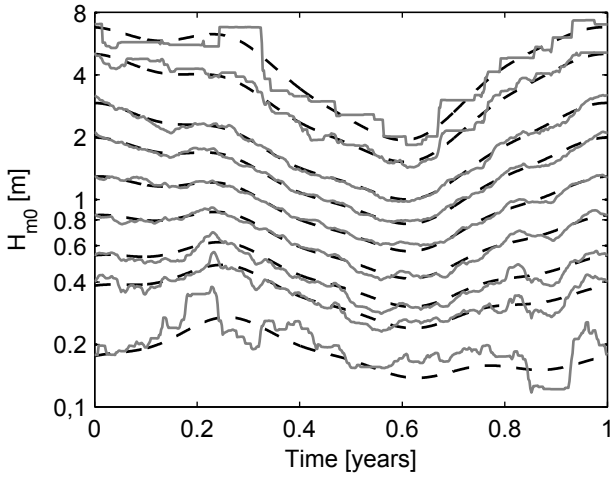


Figure 4.2: Empirical (gray) and modeled (black) quantiles of 1, 5, 10, 25, 50, 75, 90, 99 and 99.9% for H_{m0} as a function of time.

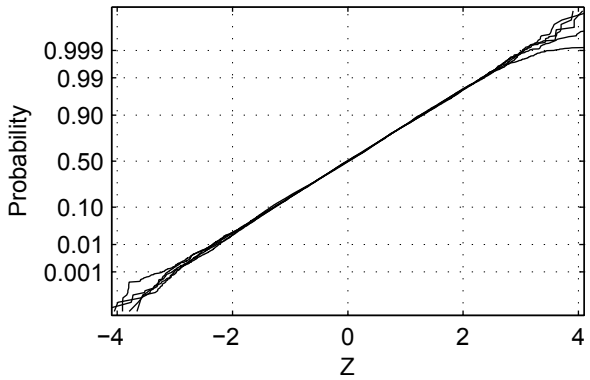


Figure 4.3: CDF of the normalized variables in normal probability paper.

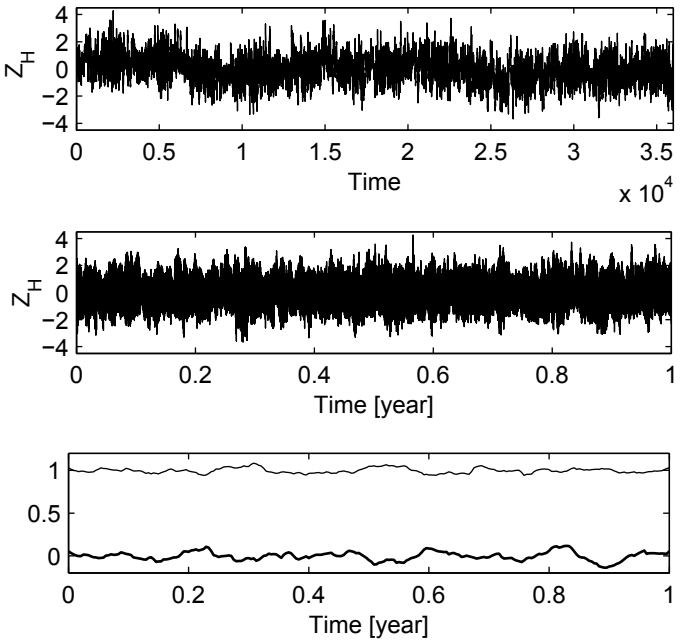


Figure 4.4: Time series of the normalized variable Z_H (top), normalized variable on annual scale (middle), and 90 days moving average of the mean and the standard deviation (bottom).

4.9 Autoregressive models parameters estimation

4.9.1 VAR model

The parameters of the VAR(p) model are estimated through the least square method described on section 4.6, for order p between 1 and 8. For each model the BIC is calculated, and the lower BIC is obtained with $p = 7$.

4.9.2 TVAR model

First a two regimes model ($K_R = 2$) is fitted using several different variables z for defining the regimes: wave height H_{m0}, normalized wave height Z_H, peak period T_p, normalized period Z_T, wind speed V_W, normalized wind speed Z_V, wave steepness H_{m0}/T_p^2 and pseudo normalized wave steepness Z_H/Z_T.

According to the BIC the best fit model is obtained when z is wind speed V_W, although good fits are also obtained when taking z as the normalized wind speed or the wave height. In all the three cases optimum delay d is 1 and optimum order p is 7.

After that a three regime ($K_R = 3$) model is fitted, but in this case only significant wave height H_{m0}, wind speed V_W and normalized wind speed Z_V are evaluated as regime defining variables z. Again best fit, evaluated through BIC, is obtained when $z = V_W$ with delay $d = 1$ and order $p = 7$.

BIC of the three regimes model TVAR$(3, 7)$ is smaller than the BIC of the two

regimes model TVAR(2, 7), which in turn is smaller than the BIC of the standard model VAR(7). Therefore model TVAR(3, 7) is selected, with $z = V_W$ and $(r_1, r_2) = (3.8\,m/s, 7.1\,m7s)$.

Results obtained indicate that the time dependence structure of the variables depends on the intensity of the wind. Three different structures are identified: one for soft winds, with $V_W < 3.8\,m/s$ (Beaufort under 4), other for relatively strong winds $V_W > 7.1\,m/s$ (Beaufort over 5), and a last one for intermediate winds speeds $3.8\,m/s < V_W < 7.1\,m/s$ (Beaufort between 4 and 5).

It was noticed that for every studied TVAR model the optimum order p was 7 and the optimum delay d was 1, except when the pseudo-steepness was used for z, for which optimum delay d was 3.

4.9.3 MSVAR model

MSVAR models with two and three regimes are studied. In both cases order p varying between 1 and 8 is evaluated and BIC is estimated using (4.9). The model with the minimum BIC is MSVAR(3, 7).

In the MSVAR model the vector $\nu^{(j)}$ gives an idea of the physical meaning of each regime. Table 4.1 presents the three $\nu^{(j)}$ vectors obtained for the MSVAR(3, 7). It is seen that regime one corresponds to wave heights lower than the mean, with peak periods over the mean and wind speeds lower than the mean, all of them characteristics of swell conditions. Regime three on the other hand corresponds to wave heights and wind speeds higher than the mean, and peak period lower than the mean, all characteristics of sea conditions. Regime two can be seen as an intermediate regime, with average wave heights and periods, and moderate to high wind speeds.

Table 4.1: ν vector for the three regimes of the model MSVAR(3, 7).

Reg.	Z_H	Z_T	$Z_{\theta M}$	Z_V	$Z_{\theta W}$
1	-0.105	0.102	0.215	-0.282	-0.078
2	0.023	0.038	-0.091	0.188	0.053
3	0.308	-0.842	-0.694	0.228	0.006

4.9.4 Residuals analysis

A usual way to evaluate the quality of the autoregressive models is to study the residuals. In this work residuals where supposed to follow a normal distribution in order to estimate the LLF of the autoregressive models.

Figure 4.5 shows the residuals of of Z_H obtained with the VAR(7) model. It is notice that residuals do not show significant autocorrelation (middle panel), and that 80% of them follows a normal distribution (lower panel). The lower and upper 10% of the residuals also tend to follow a normal distributions, but with higher variance. This is considered to be a consequence of the non-stationarity observed on the variance of the residual (upper panel).

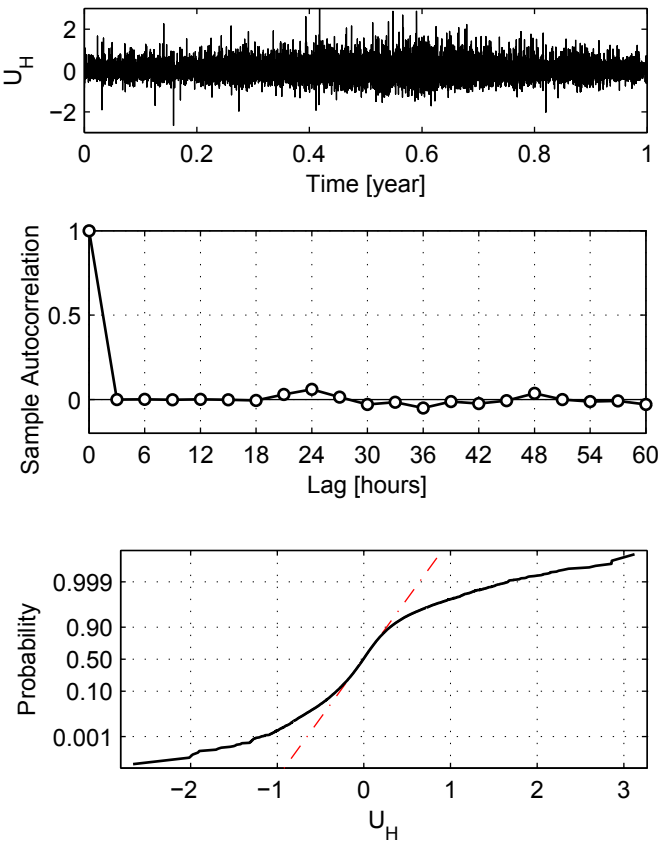


Figure 4.5: Error $\hat{U_H}$ obtained with VAR(7) model (top), autocorrelation of $\hat{U_H}$ (middle), CDF of $\hat{U_H}$ in normal probability paper (bottom).

Also an analysis on the stationarity and independence of all the remaining residuals was performed. It was observed that residuals show some non-stationarity. This is coherent with the non-stationarity observed in the dependence structure of the normalized variables.

However, from the point of view of this work, i.e. the long term simulation of new series for engineering applications, it is considered that the best way of evaluating adequacy of the models is by studying the new simulated time series. This is conducted in the next section.

4.10 Simulation

Three time series of the normalized variables, of 500 years each, are simulated using the three autoregressive models fitted in previous section. Then, these series are transformed to the original variables by means of the marginal distributions.

Next, the simulated series are compared with the original ones in terms of its marginal distributions (uni- and bi-variate) and its interannual variability, of its persistence regimes and of its auto- and cross-correlations.

4.10.1 Univariate Marginal Distributions and Interannual variability

Ferreira and Guedes Soares (2002) pointed out that probability models for sea state parameters should be able to reproduce the interannual variability that is characteristic of environmental variables. This variability is evident when one compares the PDF of different measured years.

Here, it is verified that the simulated series share the same mean annual PDF as the original series, and that they also reproduce most of the interannual variability registered on the original time series.

In figure 4.6 the annual PDF of each of the 500 years simulated with the VAR(7) model are presented, along with the PDF of the 13 measured years and its mean annual PDF. Results obtained for the other variables, as well that those obtained with models TVAR$(3,7)$ and MSVAR$(3,7)$, have the same behavior as those presented here.

First, it is noticed that the simulated series are able to reproduce the mean annual PDF of the measured series. On the other hand, the simulations show a significant variation in the annual PDF. The annual PDF of the simulated series produce a cloud around the annual mean PDF of the measured data that includes most of the measured annual PDF. However, there are at least two years of measured data whose PDF can not be reproduce by the simulated series. Those two years correspond to the first two years of the measured series, for which severer weather was observed, i.e. higher wave heights and wind speed and lower peak period.

4.10.2 Bivariate Distributions

It is important that the simulated series reproduce not only the marginal bivariate distributions of the measured data but also its marginal multivariate distributions, since the

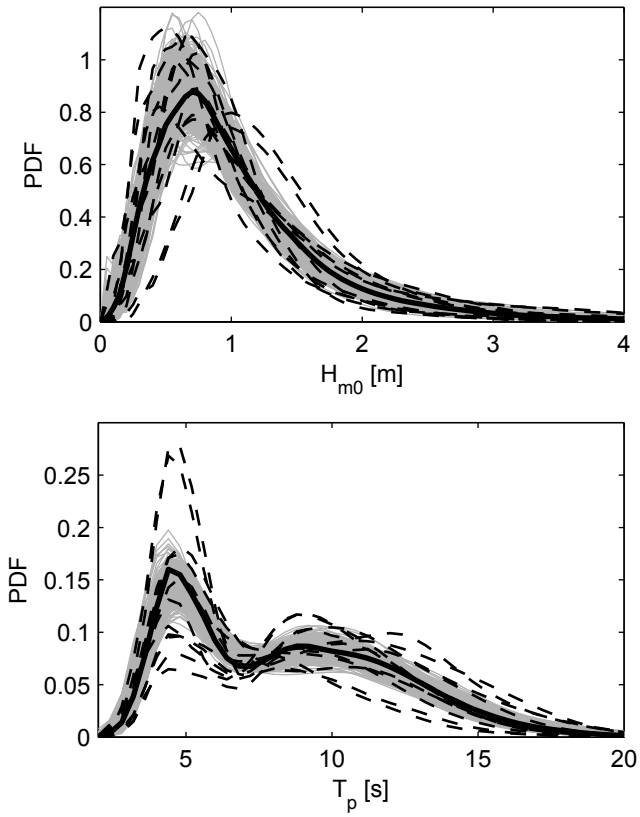


Figure 4.6: Interanual variability of the anual PDF for H_{m0} (top) and T_p (bottom). Grey: annual PDF for each simulated year; Broken Black: anual PDF for each measured year; Continuous Black: mean anual measured PDF.

latter contain information about the joint occurrence of values of the variables. Among the multivariate distributions, bivariate distributions are the the easiest to evaluate graphically and are the most familiar for the coastal engineer, therefore here the ability of the simulated series to reproduce the original bivariate distributions is analyzed.

The 10 possible bivariate distributions were analyzed. Here only the two most commonly used bivariate distributions are included. Figures 4.7 and 4.8 show bivariate distributions of (H_{m0},T_p) and (H_{m0},V_W) respectively, for both the original and the normalized variables.

It was observed that data series simulated with VAR(7) model reproduce well those bivariate distributions of the normalized variables whose behavior is similar to that of a multivariate normal distribution, i.e. with only one mode and with constant dependence structure for the whole range of values of the variables. When the bivariate distribution shows multimodality or its dependence structure depends on the value of the variables, as is the case of (Z_H,Z_T) and (Z_H,Z_V) respectively, the model is unable to capture this behavior.

Model TVAR$(3,7)$ has a better performance capturing the varying dependence structure of the (Z_H,Z_V) distribution, but it is unable to reproduce the bimodality of the (Z_H,Z_T) distribution. MSVAR$(3,7)$ on the other hand is capable of reproducing the bimodality of the (Z_H,Z_T) distribution and also improves the results obtained with the TVAR$(3,7)$ in the (Z_H,Z_V) distribution.

However, the improvement obtained with the regime switching models in the representation of the bivariate distribution of the normalized variables, does not translate into a significant improvement in the representation of the bivariate distributions of the original variables. It is observed that the bivariate distributions obtained with the three autoregressive models are very similar. All of them reproduce the main features of the bivariate distribution of the measured data, but still all of them fail in reproducing some details of the distributions, as can be the increase on the dependence between H_{m0} and T_p for storm conditions (high wave height and low periods).

4.10.3 Persistence regimes

Persistence regimes over different thresholds are useful for the estimation of the operationality of navigation channels, for the planning of marine operations, or for the estimation of the availability of renewable energy resources as wind and waves. Therefore, it is important that the simulation is as accurate as possible on the representation of the persistence regimes of the variables.

Persistence regimes of the three simulated series are very similar. In general terms it is observed that persistence regimes are well reproduced by the simulated series for thresholds close to the mean of the variables. As the threshold increases the difference between the persistence regimes of the measured and the simulated series increases too. The general trend is to produce shorter persistences than those observed in the measured data.

As an example of this, figure 4.9 shows persistence regimes of H_{m0} and V_W over thresholds corresponding to nonexceedance probabilities 0.5 and 0.9.

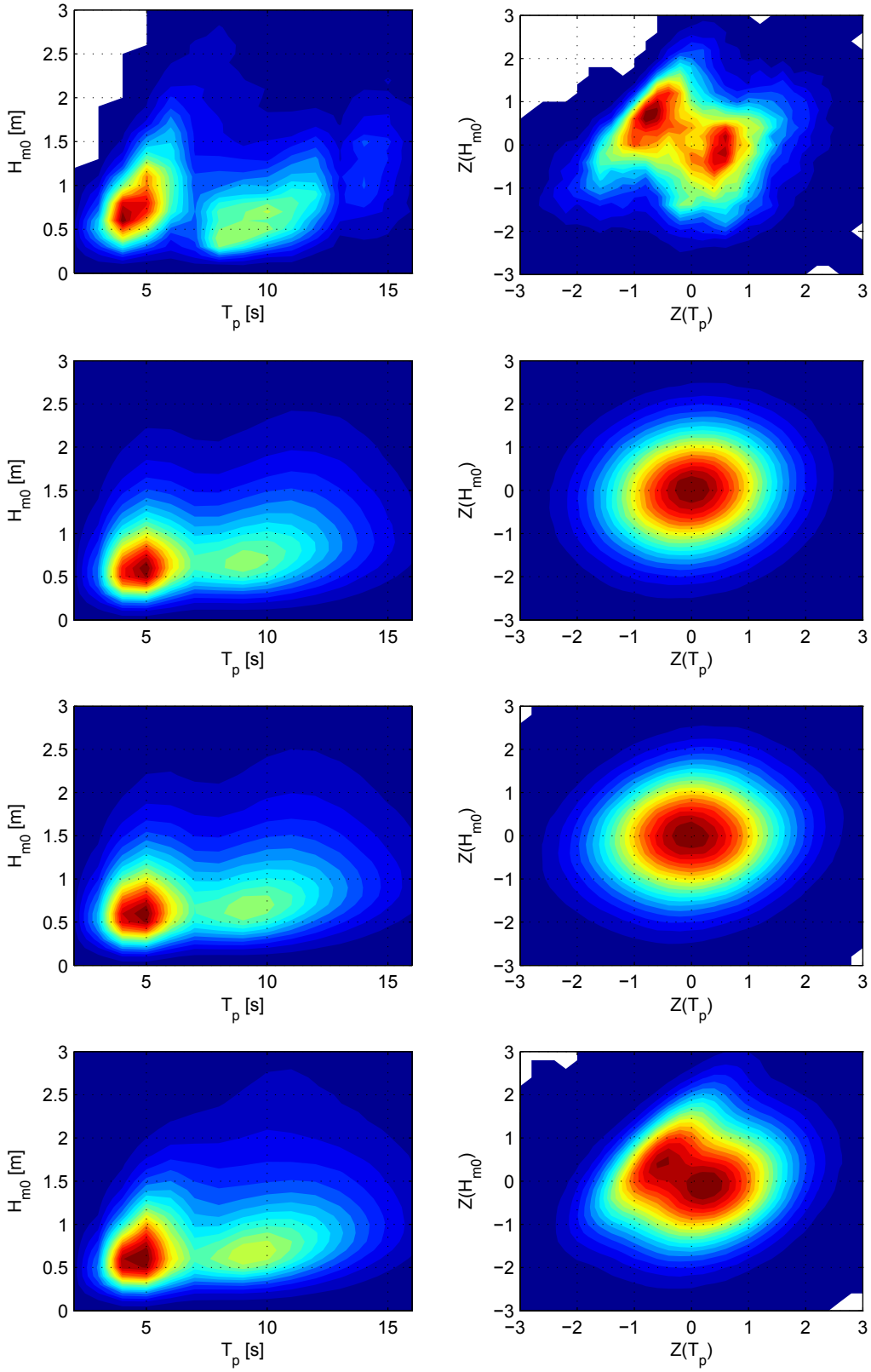


Figure 4.7: Bivariate distribution of $H_{m0} - T_p$ (left) and of $Z_H - Z_T$ (right), obtained with the measured data (top) and with the data simulated with the VAR (second from top), the TVAR (third from top) and the MSVAR (bottom) models.

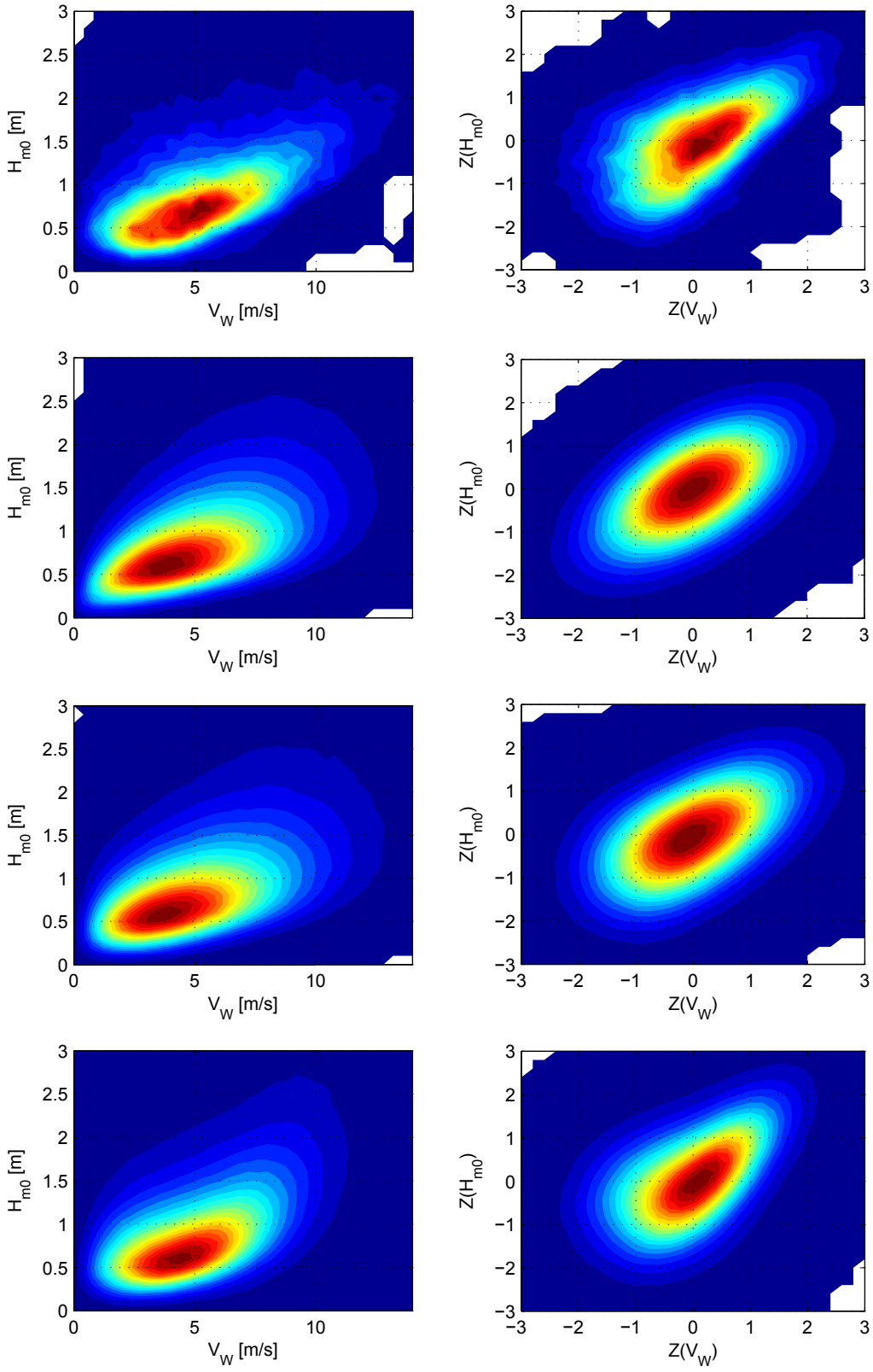


Figure 4.8: Bivariate distribution of $H_{m0} - V_W$ (left) and of $Z_H - Z_V$ (right), obtained with the measured data (top) and with the data simulated with the VAR (second from top), the TVAR (third from top) and the MSVAR (bottom) models.

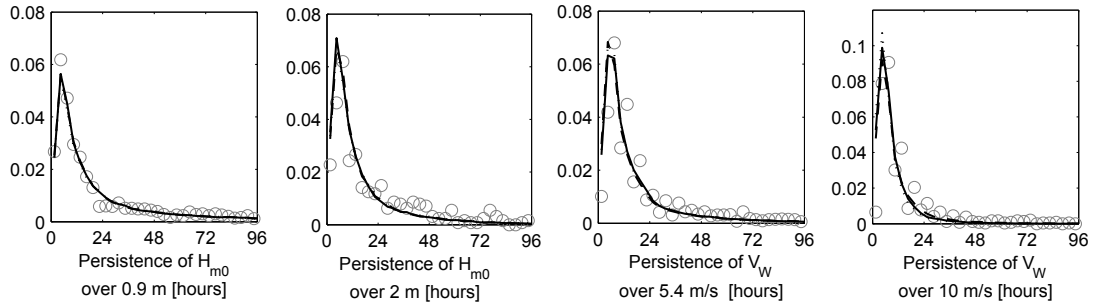


Figure 4.9: Persistence of H_{m0} and V_W over thresholds corresponding to mean annual probability 0.5 and 0.9.

Although it has not been verified, it is suspected that the observed trend to produce shorter persistences than observed may be partially caused by the inability of the model in reproducing all the observed interannual variability of the variables, i.e. if more severe years could be simulated it is expected that also longer persistence over high thresholds would occur.

4.10.4 Auto- and Cross-correlation

Figure 4.10 shows auto- and cross- correlation functions, for lag up to 48 hours, obtained with the measured and the simulated series.

It is observed that VAR(7) model is the one that better reproduce the correlation structure of the measured series. In the case of the normalized variables the correlations obtained with the simulated series are almost identical to those obtained with the measured series. However, in case of the original variables, some significant difference are observed for those correlations that involves any of the direction variables. This may be because in this work direction variables are treated as linear variables, when in fact they are circular variables. Maybe the correlation structure of the simulated series could be improved by using circular model for direction variables.

On the other hand, regime switching models TVAR$(3,7)$ and MSVAR$(3,7)$ produce unexpected effects on the correlation structure of the simulated series. In particular it is noted that the series simulated with the MSVAR$(3,7)$ model have higher cross-correlation between H_{m0} and the variables T_p, θ_M and θ_W, than that observed on the measured series. Additionally, it shows positive cross-correlation between T_p and V_W, while the measured data shows negative cross-correlation (with higher winds sea condition is observed, while with lower winds swell prevails).

4.11 Discussion

The proposed univariate marginal distribution functions provide a good fit to the measured data series, and gives a valid alternative to the methods used by other authors Cunha and Guedes Soares (1999); Stefanakos and Belobassakis (2005) for the normal-

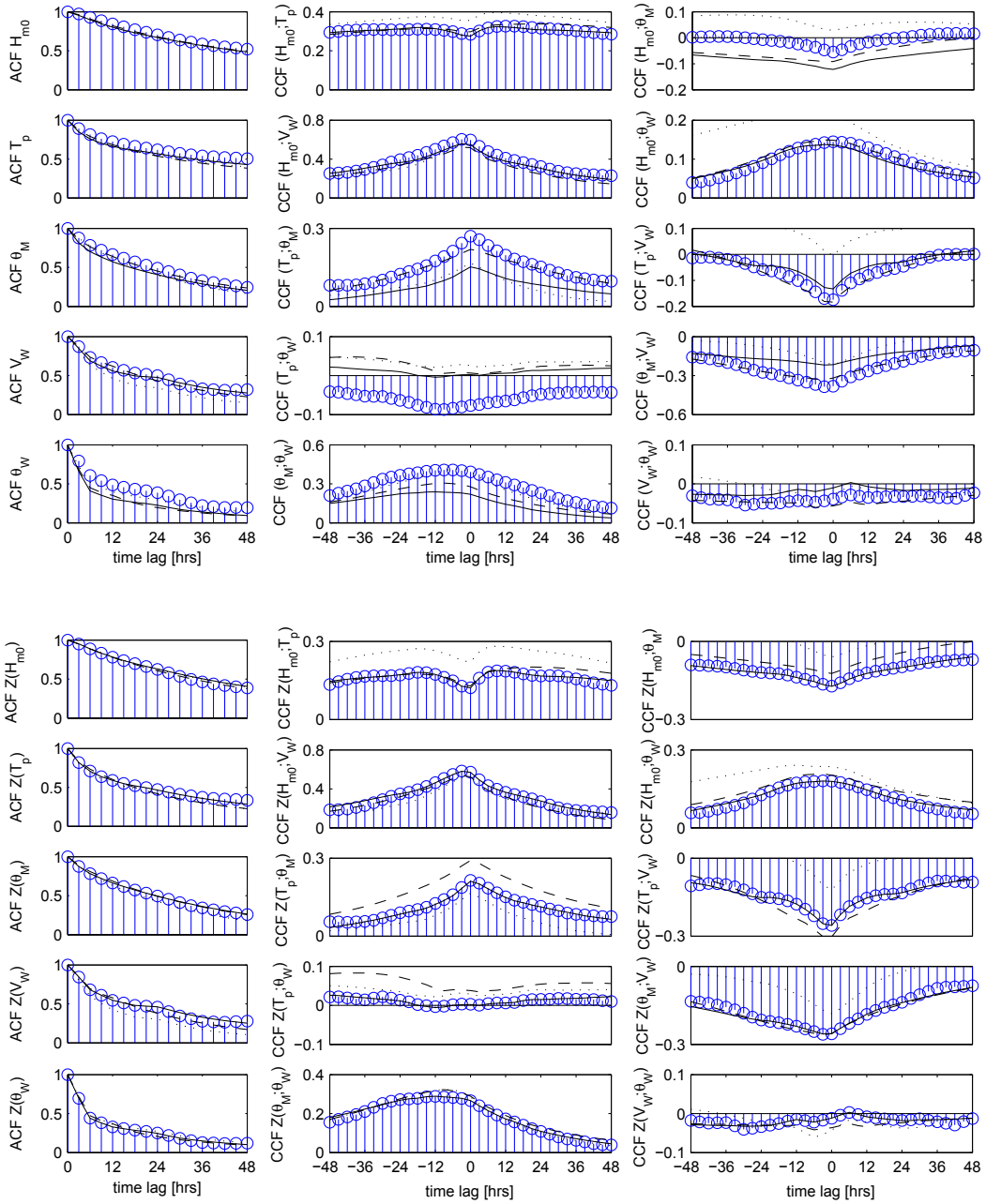


Figure 4.10: Auto- and cross correlation functions of the original (top) and normalized (bottom) variables. Lags expressed in hours. Blue circles: measured data; continuos line: VAR model simulated series; dashed line: TVAR model simulated series; dotted line: MSVAR model simulated series.

ization and stationarization of the data series. However, the proposed method has two limitations. First, it does not account for interannual variations and trends on the series. Secondly, it is unable to produce fully non-stationary series.

First mentioned limitation can easily be overcome by allowing the parameters of the distributions to have different time variation structure. Without background modifications of the proposed models, just by including additional factors on equation (4.3), the parameters can be allowed to vary with periods greater than one year (e.g.: multi year cyclic variations), to have nonperiodic dependence on time (e.g.: to have trends), and even to be a function of covariables (e.g.: climatic indexes).

On the other hand, the second mentioned limitation does not have a straightforward solution. The detailed study done about the stationarity of the normalized time series shows that these can only be considered weakly non-stationary, and that they have time varying dependence structure. This last aspect can not be taken into account with the proposed normalization procedure. In order to cope with it, it would be necessary to use time varying models for modeling the dependence structure (e.g. time varying VAR models).

It was found that, when using the VAR model for modeling the time dependence structure of the series, most of its behavior can be explained. New simulated series obtained with the fitted VAR model reproduce satisfactorily the auto- and crosscorrelation structure of the original series, as well as its univariate marginal distributions, and to some extend its interannual variability and its persistence regimes.

Studied regime switching models (TVAR and MSAR) were found more able in reproducing the behavior of the bivariate marginal distributions of the normalized variables. Particulary the MSVAR is able to capture both bimodal and dependence varying bivariate distributions. This however does not translate into a significant improvement of agreement between the measured and the simulated bivariate distributions of the original variables.

4.12 Conclusions

In summary, a methodology based on the use of non-stationary distributions and autoregressive models was introduced, which can be used for the simulation of long-term series of 5-variate met-ocean variables.

Through an in depth analysis of the simulated series main limitations of the simulation procedure, when used for engineering applications, were identified. Amongst them is the inability of the models to reproduce the observed persistence regimes for high thresholds of the variables.

It has been shown that the use of regime switching models do not necessarily produce better simulated series, although fitting errors are reduced, and better agreement is achieved between the measured and simulated bivariate distributions of the normalized variables. Given the unexpected behavior of the correlations observed in the series simulated with the regime switching models, its use in our case study is discouraged.

At least two possible work lines were identified and discussed in previous sections

that could improve the simulation of met-ocean variables: (a) to introduce long term variations, trends and covariables into the parameters of the marginal univariate distributions, and (b) to use time varying VAR models for modeling time dependence structure of the normalized series.

Acknowledgments

Sebastián Solari would like to acknowledge to *Ministerio de Educación* (Spanish Ministry of Education), for the financial support provided through the FPU scholarship (reference number AP2009-03235), and to the *Consejería de Economía, Innovación y Ciencia* of *Junta de Andalucía* (Ministry of Economy, Innovation and Science of the Andalusian Government, Spain) for financial supporting his stay at Delft University of Technology.

Wave and wind data series used in this work were kindly provided by Puertos del Estado (Spanish Port Authority).

References

Albert, J. H. and S. Chib (1993). Bayes inference via gibbs sampling of autoregressive time series subject to markov mean and variance shifts. *Journal of Business and Economic Statistics 11*(1), 1–15.

Cai, Y., B. Gouldby, P. Dunning, and P. Hawkes (2007). A simulation method for flood risk variables. In *2nd Institute of Mathematics and its Applications International Conference on Flood Risk Assessment, 4th September 2007 , University of Plymouth, England.*

Cai, Y., B. Gouldby, P. Hawkes, and P. Dunning (2008). Statistical simulation of flood variables: incorporating short-term sequencing. *Journal of Flood Risk Management 1*, 3–12.

Cunha, C. and C. Guedes Soares (1999). On the choice of data transformation for modelling time series of significant wave height. *Ocean Engineering 26*, 489–506.

Ferreira, J. and C. Guedes Soares (2002). Modelling bivariate distributions of significant wave height and mean wave period. *Applied Ocean Research 24*, 31–45.

Guedes Soares, C. and C. Cunha (2000). Bivariate autoregressive models for the time series of sognificant wave height and mean period. *Coastal Engineering 40*, 297–311.

Guedes Soares, C. and A. M. Ferreira (1996). Representation of non-stationary time series of significant wave height with autoregessive models. *Probabilistic Engineering Mechanics 11*, 139–148.

Guedes Soares, C., A. M. Ferreira, and C. Cunha (1996). Linear models of the time series of significant wave height on the southwest coast of portugal. *Coastal Engineering 29*, 149–167.

Hamilton, J. D. (1990). Analysis of time series subject to change in regime. *Journal of Econometrics 45*, 39–70.

Harris, G. R. (1999). Markov chain monte carlo estimation of regime switching vector autoregressions. *Astin Bulletin 29*(1), 47–79.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer-Verlag.

Scotto, M. and C. Guedes Soares (2000). Modelling the long-term series of significant wave height with non-linear threshold models. *Coastal Engineering 40*, 313–327.

Solari, S. and M. A. Losada (2011a). A unified statistical model for hydrological variables including the selection of threshold for the POT method. *Submitted to Water Resource Research*.

Solari, S. and M. A. Losada (2011b). Unified distribution models for met-ocean variables: application to series of significant wave height. *Submitted to Coastal Engineering*.

Solari, S. and M. A. Losada (2011c). Non-stationary wave height climate modeling and simulation. *Journal of Geophysical Research 116 C09032*, doi:10.1029/2011JC007101.

Stefanakos, C. N. and K. A. Belobassakis (2005, June 12-16). Nonstationary stochastic modelling of multivariate long-term wind and wave data. In *Proceeding of 24th International Conference on Offshore Mechanics and Arctic Engineering (OMAE2005)*. Halkidiki, Greece.

Tsay, R. S. (1998). Testing and modeling multivariate threshold models. *Journal of the American Statistical Association 93*, 1188–1202.

van Gelder, P., W. Wang, and J. Vrijling (2007). *Extreme Hydrological Events: New Concepts for Security*, Volume 78 of *Earth and Environmental Sciences*, Chapter Statistical estimation methods for extreme hydrological events, pp. 199–252. Springer.

Chapter 5

Diseño de la profundidad de un canal de navegación con base en riesgo y uso y explotación

5.1 Resumen

Se presenta el desarrollo y la aplicación de un modelo de simulación para el diseño del calado y el ancho de un canal de navegación, atendiendo los modos de fallo toque de fondo y salidas de márgenes, y la parada operativa del canal. El modelo simula los tránsitos de entrada y salida del puerto, y en cada tránsito calcula los tiempos de espera y la probabilidad de que el barco toque fondo. Los resultados de las simulaciones se usan para calcular el riesgo en la vida útil del canal, así como el desempeño del mismo en cuanto a tiempos de espera y operatividad. El modelo se aplica para la optimización del calado del canal de acceso de la futura terminal de contenedores del Puerto de la Bahía de Cádiz, España.

5.2 Introducción

El diseño de un canal de navegación implica definir su traza, calado, ancho y política de uso, teniendo en cuenta los principales modos de fallo que condicionan la seguridad y el servicio del canal, así como el dragado inicial. Los principales modos de fallo en un canal están asociados a los buques: toque de fondo y salida de márgenes durante el tránsito. Cuando el oleaje es el agente predominante, los movimientos verticales del buque durante el tránsito determinan la seguridad y el servicio de los canales.

Encontrar un diseño óptimo del canal tal que minimice los costos conjuntos de dragado inicial, esperas por paradas operativas y pérdidas por toque de fondo durante el tránsito es un ejercicio complejo, en el cual se debe trabajar con diversos agentes climáticos, de distintas escalas temporales y espaciales, teniendo en cuenta que la posición de los buques dentro del puerto varía con el tiempo.

El diseño en base a riesgo -definido como el producto de la probabilidad de ocurrencia de un suceso y la valoración económica de sus consecuencias- o con base en beneficios, proporciona una alternativa viable para realizar esta optimización. Tanto el riesgo como el beneficio permiten resumir los diversos costos involucrados, de modo de seleccionar la alternativa de diseño que presenta el menor costo global en la vida útil del canal, i.e. se selecciona la alternativa de menor riesgo (o de mayor beneficio).

A su vez, el diseño de las obras marítimas debe cumplir con estándares predefinidos en cuanto a seguridad y servicio. Cuando su impacto económico, social y ambiental es alto, las mismas deben verificarse mediante técnicas de Nivel III (*Losada*, 2002), i.e. se deben realizar predicciones del desempeño de la obra a lo largo de su vida útil. En general estas verificaciones -o predicciones- no pueden hacerse de forma directa y debe recurrirse a técnicas numéricas, como la simulación o integración de Monte Carlo (*Losada*, 2002; *Reeve*, 2009).

En este trabajo se aborda el diseño en base a riesgo del calado y la política de uso de un canal de navegación, mediante técnicas de simulación de Monte Carlo, calculando el desempeño de la obra en su vida útil. Se propone una metodología general y se implementa para un caso de estudio en el Puerto de la Bahía de Cádiz.

El artículo se estructura en 6 secciones. En la sección 2 se resumen los antecedentes relativos a simulación y cálculo de riesgo en canales de navegación. La sección 3 presenta una descripción general de la metodología. La sección 4 describe los principales módulos de cálculo que componen el modelo. En la sección 5 se describe la implementación y aplicación del modelo para un caso de estudio. Por último en la sección 6 se resumen las conclusiones del trabajo.

5.3 Antecedentes

Las aproximaciones existentes estudian la seguridad y el servicio de los canales de forma desacoplada, y en general no utilizan una aproximación basada en riesgos o beneficios para optimizar su diseño.

Shabayek and Yeung (2000), *Shabayek and Yeung* (2001) y *Ng and Wong* (2006) usan modelos de simulación para estudiar la capacidad de las áreas de navegación y los muelles de las terminales portuarias desde un punto de vista de la investigación de operaciones. La seguridad se tiene en cuenta en forma de políticas predefinidas para el uso del sistema. *Pachakis and Kiremidjian* (2003) plantean una metodología general para realizar este tipo de estudios.

Spencer et al. (1990), *Briggs et al.* (2003) y *Quy et al.* (2008) estudian la seguridad durante el tránsito, centrando la atención en el modo de fallo por toque de fondo. *Spencer et al.* (1990) describe métodos para evaluar la probabilidad de toque de fondo debido a movimientos verticales del buque inducidos por el oleaje. *Quy et al.* (2008) diseñan en base a riesgo el calado del canal de navegación de una terminal granelera.

A su vez, existen algunas herramientas diseñadas para facilitar el cálculo de la probabilidad de fallo por toque de fondo y la operatividad de los canales (e.g. CADET y HARBORSIM, del cuerpo de ingenieros del U.S. Army), pero siempre trabajando de

forma desacoplada.

5.4 Metodología

En este trabajo se construye un modelo para el cálculo de la seguridad y el servicio en un canal de navegación mediante simulación de Monte Carlo. El modelo calcula la probabilidad de fallo, la operatividad y el riesgo en la vida útil del canal, así como los tiempos de espera de los buques.

Los componentes del modelo se resumen en el esquema de la figura 5.1. Previo a la aplicación del modelo es necesario realizar estudios de agente y de costos, y definir una alternativa de proyecto en términos de su geometría y su política de uso y explotación. El modelo consta de tres módulos: simulación de agentes, cálculo de respuesta del sistema, y cálculo de riesgos y/o beneficio. En función de los resultados obtenidos puede ser necesario modificar la alternativa de diseño, o profundizar en el estudio de agentes o costos para reducir la incertidumbre en la respuesta del sistema.

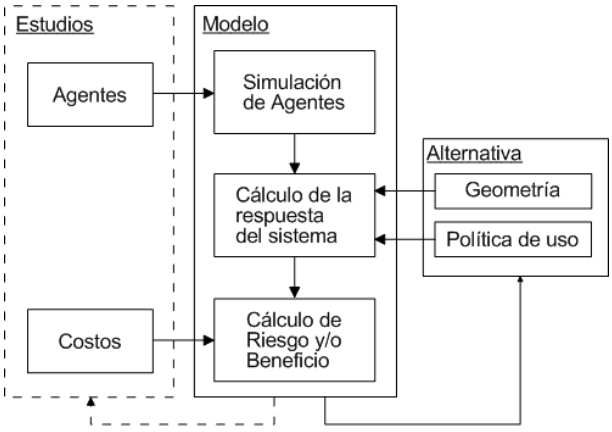


Figure 5.1: Esquema del modelo y su integración en el proceso de diseño.

La figura 5.2 muestra el esquema interno de trabajo del modelo, basado en simulación de Monte Carlo (ver e.g. *Losada* (2009)). El procedimiento consiste en realizar varios experimentos para obtener la distribución de las variables aleatorias de interés. En este caso el experimento es la vida útil del sistema y las variables aleatorias son la probabilidad de fallo, la operatividad, el riesgo (o el beneficio), y el valor esperable de los índices de desempeño, todos medidos en la vida útil (i.e. cada variable es un estadístico del experimento). Por lo tanto el modelo conjuga un aspecto de investigación de operaciones (cálculo de tiempos de espera) y un aspecto de cálculo de probabilidad de fallo mediante ecuaciones de estado.

El módulo de simulación de agentes genera M series de agentes cuya duración es igual a la vida útil del sistema (N años). En cada año se suceden una serie de ciclos de solicitación, que en este caso corresponden a los tránsitos por el canal. Cada tránsito es dividido en estados.

El módulo de respuesta del sistema calcula: (a) las variables instantáneas en cada estado (probabilidad de fallo en el estado, etc.), y (b) las variables agregadas en la vida útil (probabilidad de fallo en la vida útil $P_{F,VU}$, etc.).

El módulo de Riesgo/Beneficios calcula la distribución de las variables aleatorias de interés a partir de los resultados de los M experimentos, i.e. calcula la distribución de $P_{F,VU}$, etc.

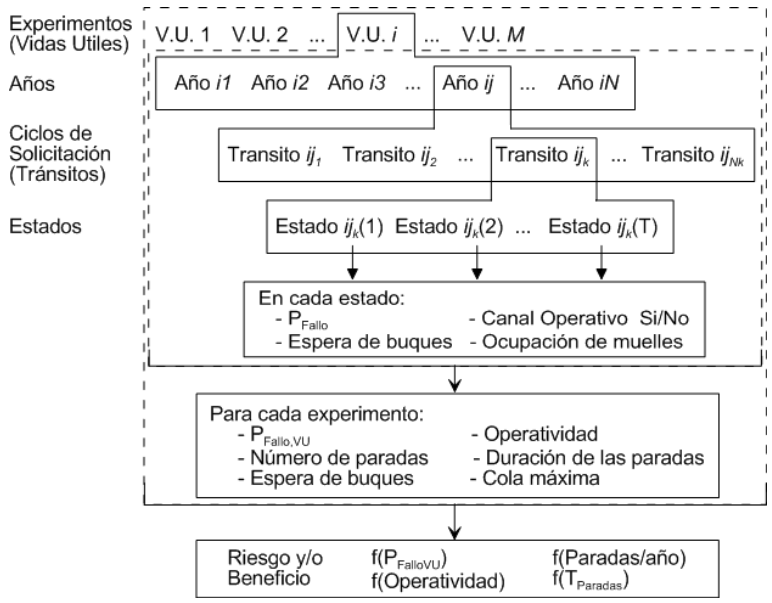


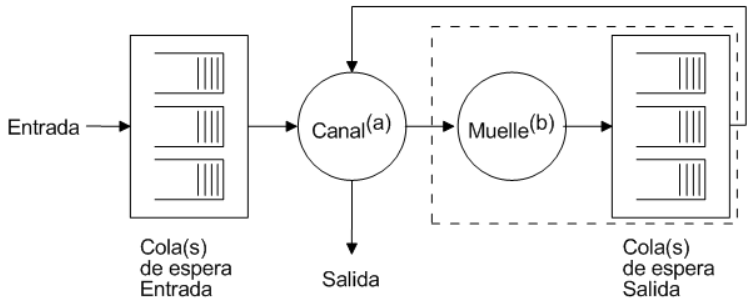
Figure 5.2: Esquema interno de trabajo del modelo (modificado de [10]).

La figura 5.3 presenta el flujo de los buques dentro del puerto como un esquema de colas y servidores. Existen dos conjuntos de colas first-in-first-out (cada cola corresponde a un nivel de prioridad). Un conjunto corresponde a la espera de entrada y otro a la espera de salida. Los servicios a los buques se prestan en el canal y en el muelle. Durante el servicio en el canal se calcula la probabilidad de fallo por toque de fondo, siguiendo el procedimiento que se describe más adelante. La operatividad del canal es la que determina el servicio del sistema, ya que la calidad del servicio se mide en término de los tiempos de espera de entrada y salida.

El núcleo del modelo es el cálculo de la probabilidad de fallo por toque fondo durante el tránsito. El esquema de cálculo seguido es el presentado en *Losada et al.* (2010), y se basa en dos conceptos: el tramo y el estado, los cuales se definen a continuación

5.4.1 Definición de los tramos de canal

El canal de acceso se divide en tramos en los que el nivel de las acciones es uniforme. Esta condición puede expresarse también en términos de los agentes. Así, para los fallos del barco en tránsito, en cada tramo deben ser uniformes la altura y la dirección del oleaje, la intensidad y la dirección de la corriente, la velocidad y la dirección del viento, el nivel de mar, y la velocidad y la dirección del buque.



(a)Canal se refiere al uso del canal de acceso y el área de maniobras, así como al servicio de remolcadores y prácticos.
(b)Muelle se refiere al uso de los muelles y al servicio de carga y descarga de mercancías.

Figure 5.3: Esquema de flujo de los barcos dentro del modelo.

5.4.2 Definición del estado climático

El estado es el período de tiempo durante el cual el nivel de las acciones sobre el buque es estacionario o estadísticamente estacionario. De forma análoga a los tramos, esto se define en término de los agentes.

Cada agente tiene un tiempo característico, o duración del estado, en el cual se asume estacionario o estadísticamente estacionario. Para los agentes oleaje y viento el estado dura entre una y tres horas. La amplitud de la marea es constante en una escala diurna o semidiurna; dependiendo de esta amplitud, el nivel de mar se considera estacionario en una escala de orden O(min) u O(horas).

La escala de tiempo que se utiliza en el modelo para definir el estado climático es la menor de las escalas de tiempo de los distintos agentes.

5.4.3 Definición del estado de tránsito

El estado de tránsito es la unidad de tiempo en la que se realiza el cálculo de la probabilidad de fallo. Durante un estado de tránsito el barco se mueve en un tramo de condiciones uniformes y todos los agentes y acciones son estacionarios o estadísticamente estacionarios.

La duración de un estado de tránsito $T_{Trnsito}$ es el tiempo mínimo necesario para que ocurra uno de los dos sucesos posibles: (a) uno de los agentes cambia de estado $\triangle\, T_{Estado}$, o (b) el barco cambia de tramo $\triangle\, T_{Tramos}$

$$T_{Trnsito} = min\{\triangle\, T_{Estado}, \triangle\, T_{Tramos}\}$$

5.4.4 Cálculo de la probabilidad de fallo

La probabilidad de fallo durante un estado de tránsito $P_{F,E}$ se calcula utilizando una ecuación de estado, la cual depende del modo de fallo.

La probabilidad de fallo en una tránsito completo $P_{F,T}$ se calcula como el complemento de la probabilidad de no fallo en el tránsito (5.1), siendo N_E el número de estados de tránsito que componen un tránsito completo.

$$P_{F,T} = 1 - \Pi_{E=1}^{N_E} (1 - P_{F,E}) \tag{5.1}$$

La probabilidad de fallo en la vida útil $P_{F,VU}$ se calcula como el complemento de la probabilidad de no fallo en la vida útil (5.2), en donde N_T es el número de tránsitos que ocurren en la vida útil del canal.

$$P_{F,VU} = 1 - \Pi_{T=1}^{N_T} (1 - P_{F,T}) \tag{5.2}$$

Los estados de tránsito que componen un tránsito no son independientes entre sí, así como los tránsitos no son independientes unos de otros. Las ecuaciones (5.1) y (5.2) son válidas siempre que el modelo de simulación contemple las dependencias entre estados y entre tránsitos.

Los estados de tránsito que componen un tránsito dado tienen dependencia temporal heredada de la dependencia temporal de los agentes climáticos. Por otro lado, la secuencia de tránsitos depende de las colas de espera, del estado operativo del canal y de los tiempos de servicio; las colas de espera están parcialmente determinadas por el nivel de ocupación de los muelles, las llamadas a puerto y los tránsitos anteriores. Estos aspectos se discuten en la sección 5.5.

5.5 Descripción e implementación del modelo

Siguiendo la metodología de trabajo planteada en la sección 5.4, se programa un modelo de simulación para diseñar el calado del canal de navegación del Puerto de la Bahía de Cádiz.

En esta sección se describe el caso de estudio, los principales aspectos teóricos del modelo y su implementación.

5.5.1 Ampliación del puerto de la Bahía de Cádiz

El puerto de la Bahía de Cádiz está ubicado en el Golfo de Cádiz, al sur de España, sobre el Océano Atlántico. En el mismo se proyecta la construcción de una nueva terminal de contenedores, para lo cual es necesario profundizar el canal de navegación actual (ver figura 5.4).

Se programa el modelo con el objetivo de definir el nuevo calado del canal de acceso y la zona de maniobras, de forma tal que minimice el riesgo en la vida útil de la obra.

En función de los estudios preliminares de agentes climáticos y de maniobras de buque se definen cuatro tramos en el canal de acceso (ver figura 5.5): tramo 1 exterior recto, tramos 2 en curva, tramo 3 interior recto y tramo 4 área de maniobras.

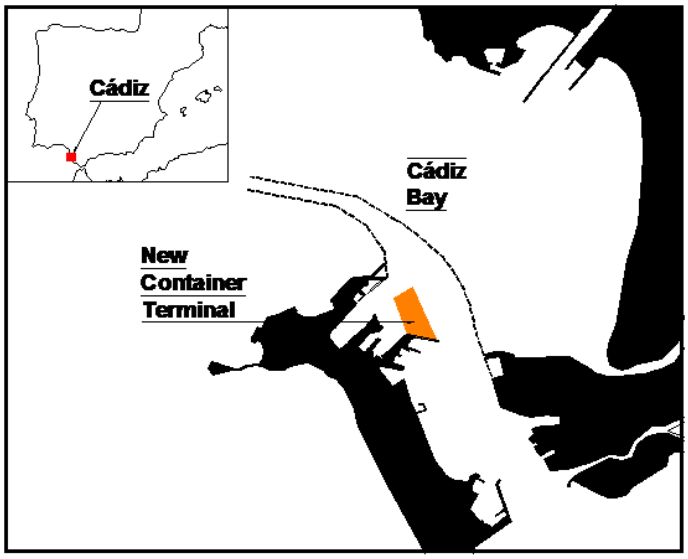


Figure 5.4: Localización del Puerto de la Bahía de Cádiz, de la futura terminal de contenedores y del canal de acceso a la misma.

5.5.2 Respuesta del sistema

El módulo de respuesta del sistema (figura 5.1) es el núcleo de cálculo del modelo. Este módulo recibe como entrada los valores de los agentes climáticos en aguas profundas y las llamadas a puerto de los buques. Dentro del módulo se calcula el estado operativo del canal, la probabilidad de fallo en cada estado y en cada tránsito, y se mantiene un registro de los buques en espera y del nivel de ocupación de los muelles.

A continuación se describen algunos aspectos específicos de la implementación de este módulo.

Propagación de oleaje

Se construye una base de datos de propagaciones de oleaje, compuesta por la propagación de 500 estados de mar desde aguas profundas hasta el canal. Dadas unas condiciones de oleaje exteriores, se interpola entre las propagaciones de la base de datos y se obtiene el oleaje en cada tramo del canal.

Para construir la base de datos se utiliza Oluca-SP. Éste es un modelo de propagación de oleaje espectral basado en la versión parabólica de la ecuación de pendiente suave, incluido en la herramienta SMC (*Gonzalez et al.*, 2007). Oluca-SP está basado en el modelo REF/DIF-S (*Kirby and Özkan*, 1994).

Niveles y corrientes

El nivel del agua se asume uniforme en el dominio, e igual al nivel exterior calculado sumando marea astronómica y meteorológica.

Para el cálculo de las corrientes se procede en tres pasos:

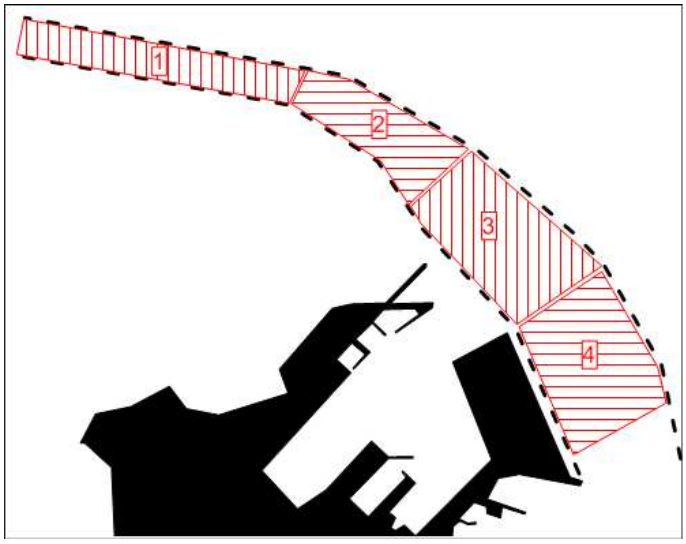


Figure 5.5: Esquema de tramos de canal de acceso (tramos 1 a 3) y del área de maniobras (tramo 4).

(a) Usando datos batimétricos actuales y tres semanas de corrientes medidas en tres puntos del dominio, se calibra un modelo hidrodinámico bidimensional, integrado en vertical.

(b) Se simulan tres meses de corrientes usando los coeficientes calibrados en (a) y la nueva batimetría propuesta.

(c) Se usan los tres meses de datos de corrientes para identificar (calibrar y verificar) modelos tipo caja negra de Hammerstein-Wiener.

Una vez calibrados, los modelos tipo caja negra calculan la velocidad de corriente en los distintos tramos del canal en función de la serie temporal de nivel de mar.

Tiempos de servicio

Se asume que los tiempos de servicio son función de la cantidad de mercancía a cargar/descargar y del rendimiento de las grúas. La cantidad de mercancía a cargar/descargar es una variable del barco y se simula en conjunto con los agentes de uso y explotación (ver sección 5.5.3). El rendimiento de las grúas es una variable del sistema que se simula dentro del módulo de respuesta del sistema.

En España el rendimiento de las grúas oscila entre 18 y 30 contenedores por hora (*Llorca*, 2009). Se asume una distribución uniforme entre los límites anteriores para el rendimiento de cada grúa, y una asignación de tres grúas por barco. Para cada barco se calcula el rendimiento simulando de forma aleatoria el rendimiento de cada grúa. Estudios detallados de los tiempos de servicio pueden realizarse mediante técnicas de simulación (ver e.g. *Sgouridis et al.* (2003) y *Sacone and Siri* (2009), en donde se aplican modelos de simulación para estudiar el desempeño de las playas de contenedores y los tiempos de servicio).

Condiciones de operatividad preliminares

Se realiza un estudio en planta, mediante un modelo de piloto automático, para definir en qué condiciones climáticas (viento, oleaje y corrientes) y con qué número de remolcadores, el barco es capaz de realizar los tránsitos de entrada y salida sin salirse de los márgenes del canal. La respuesta del modelo de piloto automático es determinista y por lo tanto el análisis en planta es determinista, i.e. dadas unas condiciones climáticas solo hay una respuesta posible: el barco se sale de márgenes o no.

Se simula un total de 18.000 tránsitos de entrada y otros tantos de salida, cubriendo un amplio rango de condiciones climáticas: 50 estados de mar, con dirección entre NW y SW, tomada cada 22.5°, altura de ola igual a 2, 4 y 6 m, y período igual a 7, 9, 11 y 13 s; 9 estados de corriente, incluyendo condiciones extremas de llenante y vaciante; 40 estados de viento, cubriendo todo el espectro de direcciones discretizado cada 45°, con velocidades entre 6 y 22 m/s tomadas cada 4 m/s. Estas simulaciones se repiten con distinto número (entre 3 y 5) y configuración de remolcadores.

A partir de este análisis en planta se definen las condiciones climáticas para las cuales no es posible realizar el tránsito. Estas condiciones marcan los límites preliminares de operatividad del canal y se incluyen en el modelo como política de uso del canal.

Respecto al uso del modelo de piloto automático caben dos aclaraciones: (1) se está acotando la ocurrencia del modo de fallo salida de márgenes, i.e. en el modelo no existe este fallo una vez definidas las condiciones de no operatividad; (2) de incluir las componentes aleatorias en el modelo de piloto automático (factores humanos, respuesta aleatoria de buque), el ancho del canal podría diseñarse en base a riesgo del mismo modo que se procede a diseñar el calado.

Probabilidad de fallo por toque de fondo

Para calcular la probabilidad de toque de fondo en un estado de tránsito $P_{F,E}$ se asume que la amplitud de los movimientos verticales del barco sigue una distribución de Rayleigh y que cada oscilación es independiente de las demás (ver *Puertos del Estado* (1999) y *Newland* (2005)). Luego, la probabilidad de toque de fondo en el estado es

$$P_E = 1 - \left(1 - \exp\left\{-\frac{a^2}{2m_0}\right\}\right)^{\frac{T_E}{2\pi}\sqrt{\frac{m_2}{m_0}}} \tag{5.3}$$

en donde T_E es la duración del estado de tránsito; a es el espacio bajo quilla disponible para las oscilaciones del buque debidas al oleaje, una vez restados todos los demás efectos (ver abajo); m_0 y m_2 son los momentos de orden cero y orden dos del espectro de oscilaciones verticales del buque; y $\sqrt{m_2/m_0}$ es el período medio de oscilación del buque.

El cálculo de se hace según *Puertos del Estado* (1999), e incluye los siguientes factores: el calado del barco; factores relacionados con el nivel de aguas, i.e. marea meteorológica y astronómica; factores relacionados con la profundidad del canal y las imprecisiones de la batimetría; factores relacionados con el barco, i.e. trimado dinámico, escora por viento, escora por corrientes, márgenes de maniobrabilidad.

El cálculo de todos estos factores también puede hacerse utilizando otras recomendaciones, como ser *Briggs* (2006), sin que ello afecte el procedimiento general de cálculo del modelo.

El espectro de los movimientos verticales del barco depende de: la forma del casco, su distribución de masa, la velocidad y la dirección del barco, así como del oleaje, las corrientes y el nivel de aguas. Todos estos aspectos pueden tenerse en cuenta modelando numéricamente el comportamiento del barco (ver e.g. *Spencer et al.* (1990), *Briggs et al.* (2003) y *Journée* (2001)), o bien puede hacerse uso de expresiones simplificadas *Puertos del Estado* (1999). Aquí se han utilizado ambas aproximaciones.

Por un lado se calculó la función de transferencia del oleaje a los movimientos verticales totales en proa/popa usando el modelo numérico WASIM de DNV *DNV* (2006) en su versión lineal. Se modeló una geometría simplificada (prisma con calado, manga y eslora del buque de diseño), con distintos calados del canal (14, 16 y 18m) y con distintas velocidades relativas (de 0 a 3,5m/s cada 0,5m/s). Luego, usando espectros TMA unidireccionales (ver e.g. *Massel* (1996)) de altura de ola unitaria, se construye una base de datos de momentos m_0 y m_2. Esta base de datos se incluye en el modelo y, dadas las condiciones en un estado de tránsito, se interpolan los momentos m_0 y m_2. Finalmente los momentos se escalan con la altura de ola en el tramo.

Por otro lado se ha incluido en el modelo el método simplificado de *Puertos del Estado* (1999), el cual da la amplitud significante del movimiento vertical del buque. Esta amplitud significante es equivalente a $2\sqrt{m_0}$. Este método no proporciona el período medio de oscilación del buque T_B. Para mantenerse del lado de la seguridad T_B se toma igual al período pico del oleaje T_P. Una alternativa es usar el período característico del buque (ver e.g. *Vossers* (1962)).

5.5.3 Generación de series aleatorias

El módulo de generación de series aleatorias simula nuevas series de agentes climáticos y de uso y explotación. Estas series entran al módulo de cálculo de respuesta del sistema.

Agentes de uso y explotación

Los agentes de uso y explotación del modelo son los buques que llegan a puerto. Para definir este agente se debe definir: (a) serie de llamadas a puerto, (b) tipo de buque (dimensiones, capacidad de maniobra, etc.), (c) servicios requeridos dentro del puerto.

Las series de llamadas a puerto pueden determinarse a partir de datos del puerto si el sistema en estudio es un sistema preexistente (*Pachakis and Kiremidjian*, 2003), o pueden estimarse en función de la demanda esperada por parte del promotor de la obra.

En el modelo se tiene en cuenta un único tipo de buque, i.e. el buque de diseño. Éste es un portacontenedores de $300\,m$ de eslora entre perpendiculares, $42\,m$ de manga y $12.5\,m$ de calado. El procedimiento de análisis sin embargo puede ampliarse a una flota conformada por diversos tipos de buque sin mayores dificultades teóricas.

Según diversos autores (ver *Pachakis and Kiremidjian* (2003), *Shabayek and Yeung* (2002) y *Noritake and Kimura* (1990)) la distribución de la variable aleatoria *tiempo entre llamadas a puerto* sigue una distribución exponencial. En este caso, como no se dispone de datos específicos el tiempo entre llamadas a puerto se modela con una distribución exponencial cuyo parámetro está dado por la demanda esperada por el promotor[1]: un barco por día.

La cantidad de mercancía a cargar/descargar en cada buque, medida en TEUs, se modela con dos distribución uniformes entre 450 y 750 TEUs (i.e.: se simulan dos valores $\mathcal{U}(450, 750)$ y se suman, obteniéndose el volumen total de TEUs a cargar y descargar).

El tiempo de carga/descarga de cada barco se calcula en función de la cantidad de TEUs a cargar/descargar y de los rendimientos de las grúas (ver sección 5.5.2), simulados para cada una de las grúas que atienden a cada barco.

Agentes climáticos

Los agentes climáticos que intervienen son oleaje, viento, niveles y corrientes.

El oleaje y el viento se simulan de forma aleatoria. Los niveles tienen una componente determinista (marea astronómica) y una aleatoria. Esta última se relaciona con la altura de ola significante mediante una regresión no lineal, obtenida específicamente para el Golfo de Cádiz (*Grupo de Puertos y Costas*, 2008). Las corrientes se modelan de forma determinista en función de la variación de nivel de mar total (ver sección 4.2.2).

Simulación de oleaje y viento

El procedimiento utilizado para la simulación de series de oleaje (altura de ola significante, período de pico espectral y dirección media) y viento (velocidad y dirección media) consiste en (*Solari and van Gelder*, 2011):

i Ajustar modelos no estacionarios mixtos univariados a cada una de las variables.

ii Des-estacionalizar y normalizar cada una de las variables usando los modelos anteriores.

iii Ajustar un modelo autorregresivo vectorial (VAR) a las variables normalizadas que explique su interdependencia y su evolución temporal.

iv Simular nuevas series de las variables normalizadas usando el modelo VAR, y luego usar los modelos univariados para obtener los valores de las variables originales.

Las funciones de distribución que se usan en la simulación son: (a) Una distribución mixta de función central biparamétrica (Log-Normal para la altura de ola significante y Weibull de mínimos biparamétrica) junto con dos Pareto generalizadas, una de mínimos para la cola inferior y una de máximos para la cola superior (5.4) (ver *Solari and Losada*

[1] Es conveniente señalar que estos agentes también pueden estar sujetos a variaciones temporales importantes en el plazo de la V.U. de la obra (*Losada*, 2002). La determinación de estas variaciones requiere un análisis de la evolución económica, entre otros factores (ver e.g. *van Dosser et al.* (2010) y *Shabayek and Yeung* (2001)). Esto no se tiene en cuenta en este trabajo.

(2011b), *Solari and Losada* (2011c), *Solari and Losada* (2011b) y *Solari and Losada* (2011a)).

$$f(x) = \begin{cases} f_m(x)F_c(u_1) & x < u_1 \\ f_c(x) & u_1 \leq x \leq u_2 \\ f_M(x)\left(1 - F_c(u_2)\right) & x > u_2 \end{cases} \tag{5.4}$$

(b) Una distribución mixta compuesta de dos Log-Normal no estacionarias para el período de pico espectral (5.5).

$$f(x) = \alpha f_{LN1}(x) + (1-\alpha) f_{LN2}(x) \tag{5.5}$$

(c) Una distribución mixta estacionaria compuesta de cuatro distribuciones Normal para las direcciones (5.6).

$$f(x) = \sum_{i=1}^{4} \alpha_i f_{Ni} \left[F_{Ni}(360) - F_{Ni}(0)\right]^{-1} \tag{5.6}$$

Las distribuciones (5.4) y (5.5) son no estacionarias. La no estacionareidad se modela a través de los parámetros de las distribuciones, los cuales se representan mediante series de Fourier (5.7), cuyo período varía entre un año y tres meses.

$$\theta(t) = \theta_{a0} + \sum_{k=1}^{N} \left(\theta_{ak}\cos(2\pi kt) + \theta_{bk}\sin(2\pi kt)\right) \tag{5.7}$$

Las figuras 5.6 y 5.7 presentan el ajuste de los modelos no estacionarios para la altura de ola significante, el período de pico espectral y la velocidad de viento.

La figura 5.8 presenta las distribuciones conjuntas altura de ola significante-período y altura de ola significante-velocidad de viento media. La figura 5.9 presenta las persistencias de la altura de ola y del viento por sobre distintos umbrales. La figura 5.10 muestra las autocorrelaciones y las correlaciones cruzadas de las series original y simulada.

En general se observa que las series simuladas presentan las mismas características que las originales.

Simulación de la marea meteorológica

La marea meteorológica MM en el Golfo de Cádiz se aproximada mediante una distribución normal, con media μ y desviación estándar σ, ambas funciones cuadráticas de la altura de ola significante en aguas profundas *Grupo de Puertos y Costas* (2008).

$$MM \sim \mathcal{N}\left(\mu(H_{m0}), \sigma(H_{m0})\right)$$

5.5.4 Cálculo del riesgo

El riesgo es la probabilidad de ocurrencia de un suceso por las consecuencias del mismo cuantificadas en términos monetarios.

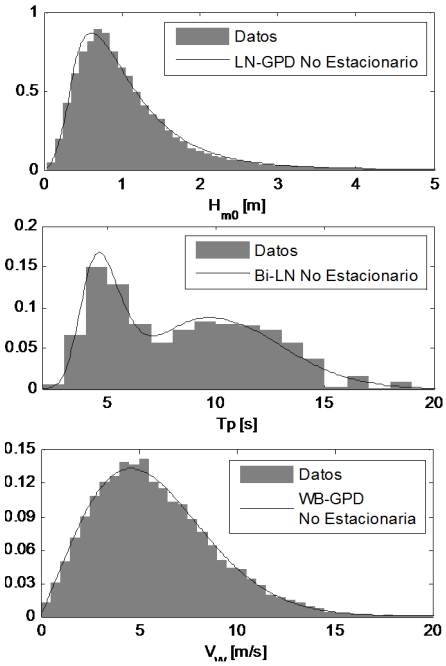


Figure 5.6: Histogramas relativos y PDF anual obtenida a partir de las distribuciones no estacionarias para Hm0, Tp y Vv.

El modelo estudia la ocurrencia de dos sucesos posibles: toque de fondo y espera para uso del canal. Las consecuencias adversas de estos dos sucesos pueden ser diversas.

La consecuencia de la espera se traduce en un costo por hora de espera. Cualquier otra consecuencia adversa derivada de la espera se considera remota y no se analiza. Dado que el objetivo es diseñar el calado del canal, solamente se consideran las consecuencias económicas de las esperas producidas por calado insuficiente, y no las producidas por otras causas: limitantes impuestas para evitar la salida de márgenes, imposibilidad de los prácticos de acceder al barco, no disponibilidad de muelles.

Las consecuencia del toque de fondo son más complejas de analizar, y abarcan desde

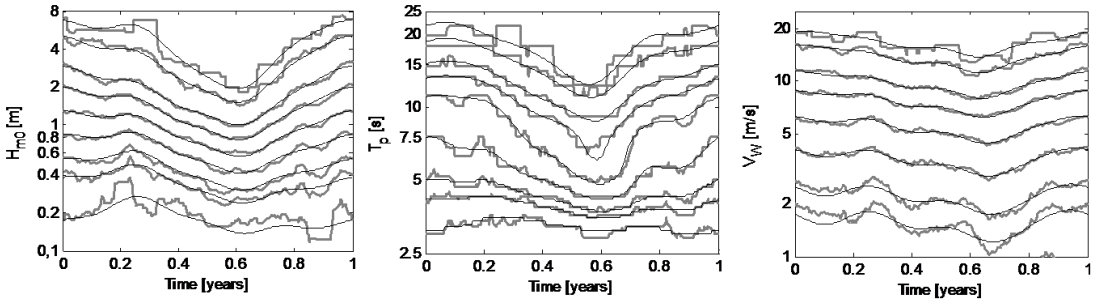


Figure 5.7: Cuantiles empíricos (gris) y modelados (negro) pata Hm0 (izq.), Tp (centro) y Vv (der.). Los cuantiles corresponden a 1, 5, 20, 25, 50, 75, 90, 99 y 99,9% (el cuantíl de 1% no se incluye en Vv).

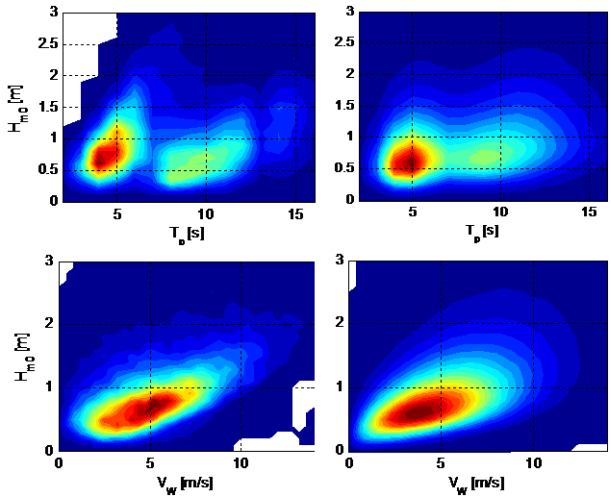


Figure 5.8: Distribución bivariada Hm0-Tp (arriba) y Hm0-Vv (abajo). Datos originales (izq.) y datos simulados (der.).

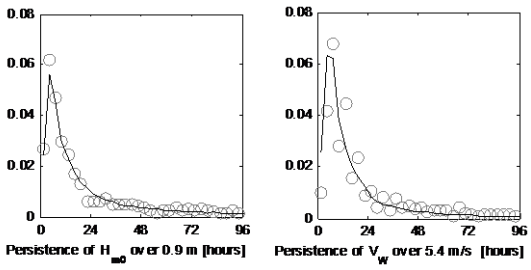


Figure 5.9: PDF de persistencia de oleaje sobre 0.9m (izq.) y viento sobre 5.4m/s (der.). Datos originales (círculos grises) y simulados (línea negra).

las casi despreciables (e.g. el casco no necesita reparación) hasta las más severas (el buque naufraga en el canal). Estas se discuten en la sección 5.5.4 de valoración de consecuencias.

Procedimiento general

El procedimiento general para el cálculo del riesgo asociado al toque de fondo debe tener en cuenta que: (a) las consecuencias en cada tramo y en cada estado pueden ser diferentes, tener diferente valoración, y tener diferente probabilidad condicionada al toque de fondo; (b) la probabilidad de las consecuencias está condicionada no solo a la ocurrencia del fallo, sino también al tramo y al estado de tránsito; (c) la valoración de las consecuencia totales del tránsito es una función no lineal de las consecuencias en cada tramo.

Para calcular el riesgo en este caso se debe incluir dentro del modelo el siguiente esquema de cálculo (figura 5.11). Para cada estado de tránsito, se calcula la probabilidad de fallo $P_{F,E}$; se genera un número aleatorio $u_E \sim \mathcal{U}(0,1)$ y si $u_E \leq P_{F,E}$ entonces

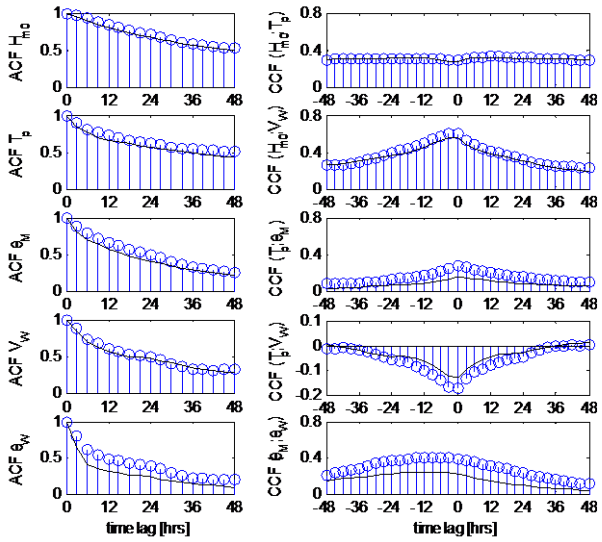


Figure 5.10: Autocorrelación y correlación cruzada de las series originales (azul) y de las series simuladas (negro).

existe fallo (i.e. toque de fondo). Si existe fallo entonces se calcula la probabilidad de las distintas consecuencias, condicionadas al estado y a la ocurrencia del fallo, y su valoración económica C_j, i.e.: $P(C_j|E \cap fallo)$, con $\sum_j P(C_j|E \cap fallo) = 1$. Se determina qué consecuencia ocurre generando un nuevo número aleatorio $u_C \sim \mathcal{U}(0,1)$. Este proceso se repite para todos los estados que componen al tránsito. El tránsito puede finalizar bien porque el buque llega a destino o bien porque las consecuencias de un fallo impiden que continúe. Una vez finalizado el tránsito se tienen las consecuencias ocurridas en cada estado $\{C_1, \ldots, C_{N_E}\}$, y se calcula la consecuencia total del tránsito C_T. La consecuencia total no tiene porqué ser lineal respecto a las consecuencias en cada estado. Las funciones más simples serían $C_T = \sum_{E=1}^{N_E}$ o $C_T = max\{C_1, \ldots, C_{N_E}\}$.

En la simulación de una vida útil este esquema se repite para cada tránsito simulado, y el valor del riesgo asociado al toque de fondo se estima como la suma de las consecuencias de cada tránsito, el costo de las esperas, y la inversión inicial

$$C_{VU} = \sum_{T=1}^{N_T} (C_T + E_{Esp,T} C_{Esp}) + C_{Ini} \tag{5.8}$$

Al realizar M simulaciones de la vida útil se estima el riesgo esperado como $E[C_{VU}] = M^{-1} \sum_{VU=1}^{M} C_{VU}$.

La aplicación de este esquema requiere disponer de un volumen importante de información respecto a la probabilidad de ocurrencia de distintas consecuencias bajo distintas condiciones, así como respecto a su valoración económica. Esto podría incluir modelos de daños del casco, etc. A continuación se discuten las hipótesis tomadas para simplificar este procedimiento.

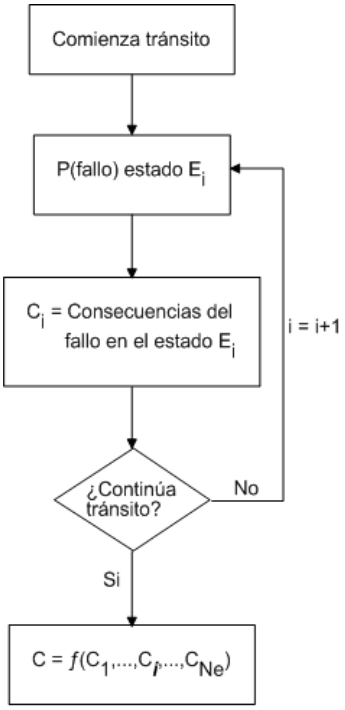


Figure 5.11: Esquema de cálculo para el riesgo en cada estado y en cada tránsito.

Procedimiento simplificado

El procedimiento simplificado parte de las siguientes hipótesis: (a) la probabilidad de fallo en cada estado de tránsito que compone un tránsito es independiente de los demás estados del tránsito; (b) las posibles consecuencias, su probabilidad condicionada al toque de fondo, y su valoración son las mismas para todos los tramos y para todos los estados; (c) el valor de las consecuencias de un tránsito es igual al de las consecuencias de mayor valor ocurridas durante el tránsito.

Bajo estas hipótesis el procedimiento de cálculo del riesgo en cada tránsito se reduce a lo siguiente. Primero se estima un valor esperado de las consecuencias válido para todos los estados de tránsito en todos los tramos

$$E[C] = \sum_j C_j P(C_j|fallo) \tag{5.9}$$

Luego se calcula la probabilidad de fallo en cada tránsito $P_{F,T}$, y se calcula el riesgo por toque de fondo en el tránsito como

$$R_T = P_{F,T} E[C] \tag{5.10}$$

Conociendo el riesgo en cada tránsito, el riesgo total en la vida útil se calcula sumando los riesgos por toque de fondo, los costos asociados a la espera y la inversión inicial

$$R_{VU} = \sum_{T=1}^{N_T} (R_T + T_{Esp,T} C_{Esp}) + C_{Ini} \tag{5.11}$$

Si bien 5.8 y 5.11 tienen la misma estructura, ambas son conceptualmente diferentes. 5.8 calcula para cada tránsito un costo simulado, mientras que 5.11 calcula para cada tránsito un costo esperado.

Es importante resaltar que el riesgo en la vida útil no puede estimarse directamente a partir de la probabilidad de fallo en la vida útil $\sum R_T \neq P_{F,VU} E[C]$, ya que en una vida útil puede ocurrir más de un fallo por toque de fondo y por tanto $E[C]$ no es el valor esperado de las consecuencias en la vida útil.

Una vez realizadas M simulaciones de la vida útil del sistema se calcula el riesgo esperable $\bar{R_{VU}}$ (media del riesgo calculado en los M experimentos), y el límite de confianza de 90% para R_{VU}, calculado como el cuantil empírico de R_{VU} correspondiente al 90%.

Valoración de las consecuencias

Dar un valor monetario a las consecuencias del fallo es un requisito básico para obtener un cálculo de riesgo. En este trabajo se utilizan la valoración de las consecuencias realizada en *Abdelouarit* (2010); sin embargo se recomienda profundizar en la valoración económica de las consecuencias realizando un estudio específico para cada caso de estudio.

Al tiempo de espera se le asigna un valor económico estimado de 10.000Ă/6hrs. Solo se tienen en cuenta las esperas producidas por la no operatividad del canal como consecuencias de un calado insuficiente. A las esperas producidas por otras causas se les asigna costo nulo.

Para el toque de fondo *Abdelouarit* (2010) define cinco escenarios de consecuencias, cada uno con un rango de costos asociados. Aquí se calcula el costo esperable de cada escenario asumiendo que la probabilidad de dichos costos, condicionada al escenario, es uniforme en el rango de valores definido en *Abdelouarit* (2010). Los cinco escenarios son:

C_1: El barco toca fondo con poco o ningún daño. Las consecuencias son inspección del casco y pequeñas reparaciones. $E[C_1] = 1$ mill. Euros.

C_2: El barco toca fondo con daño al casco. Las consecuencias son inspección y reparación del casco, posible pérdida de carga y afectaciones a otros tránsitos. $E[C_2] = 11$ mill. Euros.

C_3: El barco encalla pero puede desencallarse con marea alta. Las consecuencias son inspección del casco y pequeñas reparaciones, y posibles afectaciones a otros tránsitos. $E[C_3] = 1$ mill. Euros.

C_4: El barco encalla y necesita rescate. Las consecuencias son inspección del casco y pequeñas reparaciones, y afectación de las operaciones en el puerto. $E[C_4] = 5$ mill. Euros.

C_5: El barco encalla y naufraga. Las consecuencias son la pérdida total del barco y la carga, y afectaciones severas a las operaciones en el puerto. $E[C_5] = 50$ mill. Euros.

La probabilidades absolutas asignadas a estos escenarios por *Van de Kaa* (1984) (citado en *Abdelouarit* (2010)) se listan en la tabla 5.1 (columna 2). Usando estas probabilidades se calcula la probabilidad de cada escenario condicionada a la ocurrencia del fallo por toque de fondo (tabla 5.1 columna 3). Con estas últimas se estima el costo esperable de las consecuencias del fallo por toque de fondo usando (5.9), obteniéndose 4.35 mill. Euros.

Consecuencias	$P(C_j)$ (*Abdelouarit*, 2010)	$P(C_j\|fallo)$
C_1	5×10^{-4}	0.3267
C_2	5×10^{-4}	0.3267
C_3	5×10^{-4}	0.3267
C_4	5×10^{-5}	1.96×10^{-2}
C_5	5×10^{-7}	1.63×10^{-4}

Table 5.1: Probabilidad de los escenarios de consecuencias del fallo por toque de fondo.

Inversión inicial

Para el cálculo de la inversión inicial se calcula el volumen a dragar en función del nuevo calado, asumiendo que el canal originalmente tiene un calado de 13m, con un área de aproximadamente 1.850.000m2. El precio de dragado se asume 3 Euros/m^3.

5.5.5 Metodología de simulación

Para tener una muestra representativa de las variables de interés se deben simular M vidas útiles de la obra (ver figura 5.2), siendo M un número de orden O(10^3) (*Losada*, 2002; *Kottegoda and Rosso*, 2008), lo que implica simular O(10^4) años meteorológicos.

Teniendo en cuenta que el modelo planteado en este trabajo no tiene dependencia interanual (la simulación de agentes climáticos no incluye tendencias ni ciclos interanuales; la simulación de agentes de uso y explotación es estacionaria), se opta por acelerar el proceso de simulación mediante el uso de técnicas bootstrap *Kottegoda and Rosso* (2008). Se procede de la siguiente forma: (i) se simulan M^* años meteorológicos, (ii) se construyen M combinaciones distintas de N años cada una, siendo N la vida útil de la obra, (iii) se calculan las variables objetivo para cada una de estas M combinaciones.

5.6 Aplicación

La metodología y el modelo descritos se aplican para diseñar el calado del canal de acceso del puerto de la Bahía de Cádiz minimizando el riesgo en la vida útil del mismo.

5.6.1 Criterios generales de proyecto

Los criterios generales de proyecto para el canal de acceso y la zona de maniobras se establecen siguiendo la metodología ROM 0.0 (*Losada*, 2002; *Grupo de Puertos y Costas*, 2008). Los mismos se listan en la tabla 5.2.

El calado de diseño debe ser tal que minimice el riesgo en la vida útil de la obra (25 años) cumpliendo con los criterios generales de proyecto de la tabla 2.

Criterio	Valor
Vida Útil	25 años
Máxima Prob. de fallo en la VU	0.1
Pperatividad mínima	0.95

Table 5.2: Criterios generales de proyecto.

5.6.2 Política de uso del canal

La política de uso del canal se establece para limitar la probabilidad de fallo por toque de fondo y salida de márgenes del canal.

Las condiciones en que el buque sale de márgenes del canal se establecieron en la sección 5.5.2 de forma determinista. Para dichas condiciones el canal se considera no operativo.

En *Solari et al.* (2010) se aplicó una versión preliminar del modelo descrito, obteniéndose una política de uso tentativa para limitar la probabilidad de fallo por toque de fondo. La política de uso obtenida es

$$Op = \begin{cases} 1 & \text{si} H_{m0} \geq H_{umb} \text{y} NM \geq \alpha(H_{m0} - H_{umb}) \\ 1 & \text{si} H_{m0} < H_{umb} \\ 0 & \text{otros casos} \end{cases} \tag{5.12}$$

donde $Op = 1$ indica que el canal está operativo.

5.6.3 Procedimiento

Se simula el comportamiento del sistema para calados de entre 14 y 14.5 m, con políticas de uso del canal establecidas usando los resultados de las sección 5.5.2 y (5.12), con $\alpha = 1$ y H_{umb} entre 1.4 y 2.8 m.

Para cada combinación de calado y H_{umb} se simulan 1000 años. Con estos 1000 años se construyen, de forma aleatoria, 1000 vidas útiles de 25 años cada una (ver sección 5.5.5). Para cada combinación de 25 años se calcula la probabilidad de fallo en la vida útil, la operatividad, el riesgo y los tiempos de espera.

Con la muestra de 1000 datos se calcula la distribución de probabilidad empírica de cada una de las variables de interés. Para definir el calado óptimo se utilizan el riesgo y la probabilidad de fallo que son superados el 10% del tiempo (cuantil de 90% de sus

correspondientes distribuciones empíricas) y la operatividad que es superada el 90% del tiempo (cuantil de 10% de su distribución empírica).

5.6.4 Resultados

La figura 5.12 presenta las curvas de iso-riesgo (en log(Euros), correspondientes al límite superior de 90%) para distintos calados y valores de H_{umb}. En gris se señala la zona en la cual no se cumple la operatividad mínima (con un 90% de certeza), y en azul se señala la zona en que la probabilidad de fallo es superior a la máxima permitida (también con un 90% de certeza).

La alternativa óptima - i.e. la de riesgo mínimo -, de entre las simuladas se señala con un círculo: calado $14.1\,m$ y $H_{umb} = 1.6\,m$. La línea punteada marca la alternativa optima que se encuentra fijando el calado o H_{umb}.

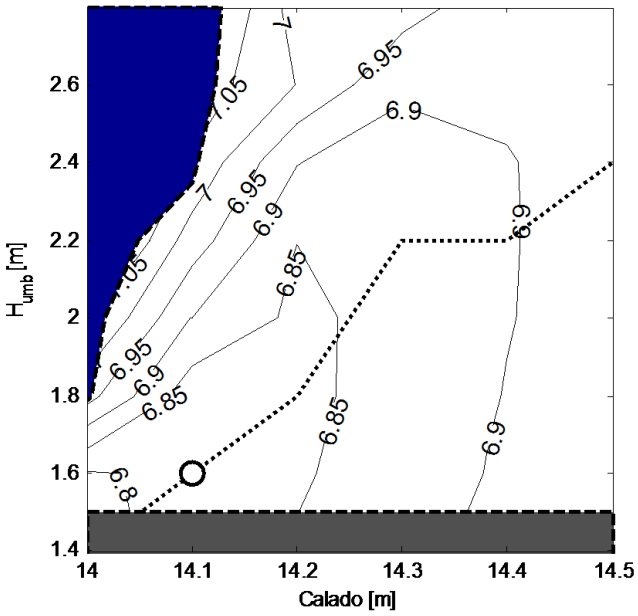


Figure 5.12: Curvas de iso-riesgo en función de la profundidad del canal y del valor de Hu usado para definir la política de operación. Area gris: operatividad en la VU menor al 95%. Area azul: Pfallo en la VU mayor a 90%.

La figura 5.13 presenta las frecuencias esperables de la duración de las esperas de entrada, junto con sus límites de confianza. Se observa que aproximadamente el 92% de los barcos que llegan a puerto no tienen que esperar para entrar.

En la figura 5.14 se analiza el número de paradas operativas del canal asociadas a la política de uso (5.12), establecida para limitar la probabilidad de fallo por toque de fondo. Se observa que las paradas más frecuentes son de entre 2 y 3 horas de duración, pero que existe una variabilidad importante en el número de paradas operativas por año.

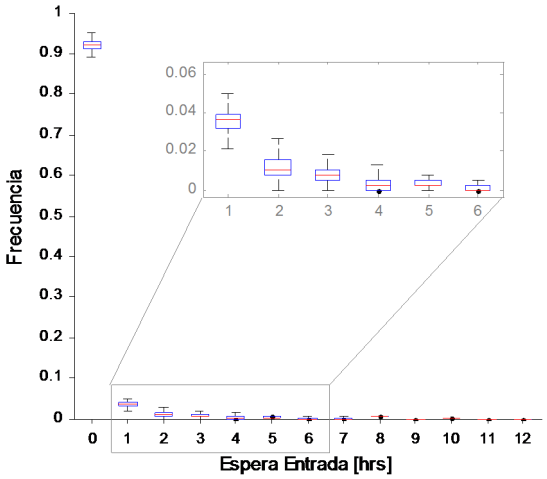


Figure 5.13: Box-Plot de la frecuencia de la duración de las esperas de entrada.

5.7 Discusión y conclusiones

Durante el diseño de las áreas de navegación se busca maximizar la seguridad y el servicio del sistema, cumpliendo con los requisitos de proyectos establecidos por la normativa vigente, sin que esto implique una inversión inicial excesiva.

Realizando el diseño con base en la minimización del riesgo total en la vida útil (incluyendo en el riesgo la inversión inicial) se obtiene un diseño óptimo teniendo en cuenta los tres aspectos antes mencionados: seguridad, servicio e inversión inicial.

La metodología y el modelo presentados calculan el riesgo en la vida útil mediante técnicas de simulación de Monte Carlo. Mediante esta aproximación se tienen en cuenta la interacción existente entre la seguridad y el servicio del sistema, y se cuantifica la incertidumbre existente en el cálculo de las distintas variables de interés (riesgo total, probabilidad de fallo en la vida útil, operatividad, etc.).

Este último aspecto es particularmente importante cuando se trabaja con sistemas complejos como el puerto, en donde las variables climáticas son determinantes en la ocurrencia de los distintos modos de fallo y parada. En el caso de estudio planteado se utilizó esta metodología para el diseño del calado y la política de uso de las áreas navegables del puerto de la Bahía de Cádiz. Para el diseño de estos dos elementos el principal modo de fallo tenido en cuenta ha sido el toque de fondo.

El modo de fallo salida de márgenes se analizó de modo determinista y se tuvo en cuenta en la definición de las políticas de uso. Sin embargo la metodología propuesta permite tratar este modo de fallo de forma probabilista, y de este modo diseñar el ancho del canal con base en la minimización del riesgo.

Como continuación de este trabajo se abren varias líneas, tanto de investigación como de desarrollo tecnológico. La más importante de ellas es incluir en el modelo de simulación un módulo de piloto automático que permita tratar sus parámetros de forma aleatoria. Esto permitirá incluir el error humano en los tránsitos (respecto a la

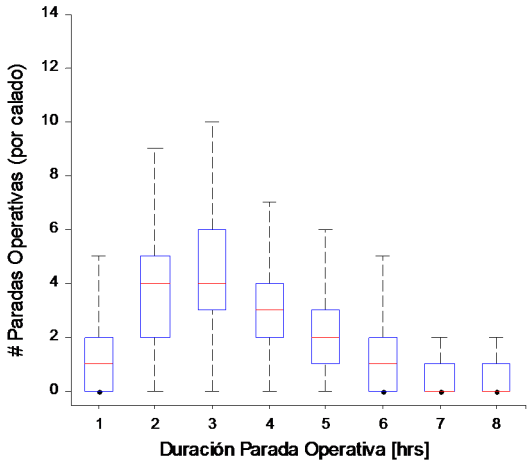


Figure 5.14: Número de paradas operativas por año, producidas por la política dada en (5.12), y su duración en horas.

importancia del error humano ver *Faber* (2003)), y aplicar el diseño en base a riesgo al ancho del canal. Por otro lado, para generalizar el uso de esta metodología es necesario trabajar en la creación de bases de datos de buques, costos, etc., y en la generación de un entorno de trabajo que facilite su aplicación por parte de las empresas proyectistas.

Si bien las líneas de trabajo mencionadas propician una mejora en la funcionalidad del modelo, es importante destacar que las mismas deben ser enmarcadas dentro de la metodología de trabajo descrita. El uso de esta metodología permite diferenciar de forma clara las distintas escalas temporales y espaciales involucradas en los distintos modos de fallo y parada, para luego integrarlas de forma correcta en el cálculo de las probabilidades en la vida útil; esto es, de hecho, el principal aporte de este trabajo.

Agradecimientos

Esta investigación fue financiada por el Ministerio de Educación de España, a través del programa FPU de becas doctorales (AP2009-03235). Parte del trabajo fue realizado durante la estancia de Sebastián Solari en la Universidad Técnica de Delft, financiada por la Junta de Andalucía a través de las ayudas para estancias en centros de excelencia (convocatoria 2/2009). La investigación ha sido parcialmente financiada a través de los proyectos de investigación y desarrollo "Optimización de la Operatividad Portuaria Mediante Técnicas de Simulación" (CIT-460000-2009-021) y "Gestión Integral de la Seguridad Portuaria" (P-53/08), financiados por el Ministerio de Educación y Ciencia, y a través del proyecto "Estudio de Alternativas para la Ampliación de la Dársena Portuaria de Cádiz", financiado por la Autoridad Portuaria de la Bahía de Cádiz.

Se agradece a Puertos del Estado los datos de oleaje suministrados, a la Autoridad del Puerto de la Bahía de Cádiz por su colaboración, y al Dr. Pieter van Gelder de

la Universidad Técnica de Delft por sus comentarios en cuanto a simulación de series climáticas y cálculo de riesgos.

Referencias

Abdelouarit, Y. (2010), Probabilistiche diepte modellering binnenhavengebied haven van Rotterdam, Master Thesis (in dutch), Technische Universiteit Delft.

Briggs, M. J. (2006), Ship Squat Predictions for Ship/Tow Simulator.

Briggs, M. J., L. E. Borgman, and E. Bratteland (2003), Probability assessment for deep-draft navigation channel design, *Coastal Engineering*, *48*(1), 29–50, doi: 10.1016/S0378-3839(02)00159-X.

DNV (2006), WASIM Wave Loads on Vessels with Forward Speed.

Faber, M. (2003), Risk assessment for civil engineering facilities: critical overview and discussion, *Reliability Engineering & System Safety*, *80*(2), 173–184, doi: 10.1016/S0951-8320(03)00027-9.

Gonzalez, M., R. Medina, J. Gonzalez-Ondina, A. Osorio, F. Mendez, and E. Garcia (2007), An integrated coastal modeling system for analyzing beach processes and beach restoration projects, SMC, *Computers & Geosciences*, *33*(7), 916–931, doi: 10.1016/j.cageo.2006.12.005.

Grupo de Puertos y Costas (2008), Estudio de Alternativas para la Ampliación de la Dársena Portuaria de Cádiz, *Tech. rep.*, Universidad de Granada.

Journée, J. (2001), Theoretical Manual of SEAWAY, *Tech. rep.*, Delft University of Technology.

Kirby, J. T., and H. Özkan (1994), Combined refraction/diffraction model for spectral wave conditions. Ref/Dif S version 1.1. Documentation and userâĂŹs manual. Report No. CACR- 94-04, *Tech. rep.*, Center Applied Coastal Research, University of Delaware.

Kottegoda, N., and R. Rosso (2008), *Applied Statistics for Civil and Environmental Engineers*, 2nd editio ed., Wiley-Blackwell.

Llorca, J. (2009), Tema B3. Terminales Portuarias: Diseño de Infraestructura, *Tech. rep.*

Losada, M. A. (2002), *ROM 0.0 General procedure and requirements in the design of harbor and maritime structures. PART I.*, 1st ed., Puertos del Estado, Madrid.

Losada, M. A. (2009), *ROM 1.0-09 Climatic Agents Description & Breakwaters Design Criteria*, Puertos del Estado, Madrid.

Losada, M. A., A. Baquerizo, M. Ortega-Sanchez, J. M. Santiago, and E. Sanchez-Badorrey (2010), Socioeconomic and Environmental Risk in Coastal and Ocean Engineering, in *Handbook of Coastal and Ocean Engineering*, edited by Y. C. Kim, chap. 33, pp. 923–952, World Scientific Publishing Co. Pte. Ltd.

Massel, S. R. (1996), *Ocean Surface Waves: Their Physics and Prediction*, World Scientific Publishing Co. Pte. Ltd.

Newland, D. (2005), *An Introduction to Random Vibrations, Spectral & Wavelet Analysis*, third edit ed., 512 pp., Dover Publications.

Ng, W.-C., and C.-S. Wong (2006), Evaluating the Impact of Vessel-Traffic Interference on Container Terminal Capacity, *Journal of Waterway, Port, Coastal, and Ocean Engineering*, *132*(2), 76, doi:10.1061/(ASCE)0733-950X(2006)132:2(76).

Noritake, M., and S. Kimura (1990), Optimum Allocation and Size of Seaports, *Journal of Waterway, Port, Coastal, and Ocean Engineering*, *116*(2), 287–299.

Pachakis, D., and A. S. Kiremidjian (2003), Ship Traffic Modeling Methodology for Ports, *Journal of Waterway, Port, Coastal, and Ocean Engineering*, *129*(5), 193, doi:10.1061/(ASCE)0733-950X(2003)129:5(193).

Puertos del Estado (Ed.) (1999), *ROM 3.1-99 Recommendations for the Design of the Maritime Configuration of Ports, Approach Channels and Harbour Basins*, Puertos del Estado, Madrid, Spain.

Quy, N., J. Vrijling, and P. van Gelder (2008), Risk- and Simulation-Based Optimization of Channel Depths: Entrance Channel of Cam Pha Coal Port, *Simulation*, *84*(1), 41–55, doi:10.1177/0037549708088958.

Reeve, D. (2009), *Risk and Reliability: Coastal and Hydraulic Engineering*, 1st ed., Spon Press.

Sacone, S., and S. Siri (2009), An integrated simulation-optimization framework for the operational planning of seaport container terminals, *Mathematical and Computer Modelling of Dynamical Systems*, *15*(3), 275–293, doi:10.1080/13873950902808636.

Sgouridis, S. P., D. Makris, and D. C. Angelides (2003), Simulation Analysis for Midterm Yard Planning in Container Terminal, *Journal of Waterway, Port, Coastal, and Ocean Engineering*, *129*(4), 178, doi:10.1061/(ASCE)0733-950X(2003)129:4(178).

Shabayek, A., and W. Yeung (2001), Effect of seasonal factors on performance of container terminals, *Journal of Waterway, Port, Coastal, and Ocean Engineering*, *127*(June 2001), 135.

Shabayek, A., and W. Yeung (2002), A simulation model for the Kwai Chung container terminals in Hong Kong, *European Journal of Operational Research*, *140*(1), 1–11, doi:10.1016/S0377-2217(01)00216-8.

Shabayek, A., and W. W. Yeung (2000), A queuing model analysis of the performance of the Hong Kong container terminals, *Transportation Planning and Technology*, *23*(4), 323–351, doi:10.1080/03081060008717656.

Solari, S., and M. Losada (2011a), Non-stationary wave height climate modeling and simulation, *Journal of Geophysical Research*, *116*(C09032), 1–18, doi: 10.1029/2011JC007101.

Solari, S., and M. A. Losada (2011b), A unified statistical model for hydrological variables including the selection of threshold for the POT method, *submitted to Water Resource Research*.

Solari, S., and M. A. Losada (2011c), Unified distribution models for met-ocean variables: application to series of significant wave height., *submitted to Coastal Engineering*.

Solari, S., and P. van Gelder (2011), On the use of Vector Autoregressive (VAR) and Regime Switching VAR models for the simulation of sea and wind state parameters, in *on CENTEC Anniversary Book. Centre for Marine Technology and Engineering (CENTEC), Technical University of Lisbon, Portugal, in pressing.*, edited by C. Guedes Soares, Taylor & Francis.

Solari, S., A. Moñino, A. Baquerizo, and M. Losada (2010), Simulation Model for Harbor Verification and Managment, in *Proceedings of 32nd Conference on Coastal Engineering, Shanghai, China, 2010*, edited by J. McKee Smith and P. Lynett, p. 11, Coastal Engineering Research Council, Shanghai.

Spencer, J., E. Bowers, and G. Lean (1990), Safe Underkeel Allowances for Vessels in Navigation Channels, in *Proceedings of 22nd Conference on Coastal Engineering*, pp. 3127–3139, Delft, The Netherlands.

Van de Kaa, E. (1984), Safety criteria for channel depth design.

van Dorsser, J., R. Haskoning, M. Wolters (2010), A Very Long Term Forecast for the development of the Cargo Flows in the le-Havre-Hamburg range, in *Port Infrastructure Seminar 2010*, Delft, The Netherlands.

Vossers, I. G. (1962), *Behavior of Ships in Waves. Vol C of Resistance, Propulsion, and Steering of Ships*, Haarlem.

Conclusiones

Conclusiones generales

A fin de profundizar en los conocimientos necesarios para la verificación, optimización y gestión de sistemas costeros y portuarios, la introducción de esta tesis plantea dos objetivos generales: (a) investigar y desarrollar metodologías de simulación de series temporales de variables geofísicas, y (b) definir e implementar una metodología de verificación y optimización basada en técnicas de Nivel III y en cálculo de riesgo de un sistema formado por las áreas navegables de un puerto.

En lo referente al primero de estos objetivos, se propuso una metodología genérica para la simulación de series temporales multivariadas de variables geofísicas y se aplicó en la simulación de variables oceanográficas y atmosféricas. Partiendo de las series de datos y siguiendo los pasos señalados en esta tesis –ajuste de distribuciones mixtas estacionarias y no estacionarias, ajuste de modelos de dependencia temporal y de dependencia cruzada entre variables–, se está en condiciones de simular nuevas series aleatorias multivariadas que conservan las principales características estadísticas de las series originales.

La metodología propuesta presenta desarrollos novedosos en el área de la ingeniería de puertos y costas. Entre ellos destacan la propuesta de funciones de distribución paramétricas mixtas, no estacionarias, capaces de representar correctamente todo el rango de valores de una variable y los ciclos de variabilidad de la misma, y la propuesta de una metodología de simulación de series temporales univariadas basada en el uso de copulas.

Es evidente que la aplicabilidad de esta metodología no se limita al caso de estudio elegido en esta tesis –la optimización de los canales de navegación–. Por el contrario la misma tiene un amplio rango de aplicación en la verificación de sistemas de diversas características, como ser la realización de estudios morfológicos a mediano plazo, el análisis de riesgo de inundaciones en zonas costeras, o la optimización de obras mediante la simulación de su desempeño a lo largo de la vida útil, entre otras.

Si el primer objetivo se enfoca en entender y modelar los forzantes predominantes en los sistema costeros y portuarios, el segundo centra la atención en cómo modelar la respuesta del sistema frente a las acciones ejercidas por los mismos. Para esto se propuso un marco teórico y una metodología de trabajo genérica que, mediante simulaciones de Monte Carlo, permite calcular los distintos índices que caracterizan la respuesta del sistema: probabilidad de fallo en la vida útil, operatividad media anual, etc.

La metodología desarrollada es novedosa en cuanto propone una forma de trabajo que permite conjugar aspectos de seguridad y de operatividad en un mismo modelo del sistema, cuando en la bibliografía estos dos aspectos suelen trabajarse disociados. Esto permite el cálculo del riesgo total en la vida útil de la obra, lo que en definitiva permite realizar una verificación, u optimización en su caso, integral.

Por último, las dos metodologías desarrolladas –simulación de series temporales y cálculo de riesgo en áreas navegables– se aplicaron en la verificación y optimización con base en riesgo de la profundidad de dragado de los canales de navegación del puerto de la Bahía de Cádiz.

En resumen esta tesis presenta una metodología de trabajo para la verificación y optimización integral de sistemas costeros y portuarios en los que las acciones ejercidas por los agentes climáticos marinos y atmosféricos son predominantes, haciendo especial hincapié en los sistemas formados por las áreas navegables de los puertos.

Conclusiones específicas

La distribución mixta propuesta, compuesta de una distribución central (truncada o no, según el caso) y de dos distribuciones Pareto generalizadas en las colas, es una herramienta eficaz para representar la distribución de probabilidad de una variable aleatoria geofísica en todo su rango de valores, tanto en condiciones estacionarias, asumiendo una distribución media anual, como en condiciones no estacionarias en la que es posible acomodar ciclos climáticos, tendencias y covariables.

A su vez, el modelo mixto proporciona de forma automática el umbral superior necesario para aplicar el método de picos sobre el umbral, así como la incertidumbre de dicho umbral. Se propuso una metodología simple para incluir esta incertidumbre en la estimación de los intervalos de confianza de los cuantiles de alto período de retorno.

El modelo fue probado en seis series diferentes: dos de precipitación diaria, dos de caudal medio diario y dos de altura de ola significante. En todos los casos se obtuvo que el modelo mixto mejora el ajuste de forma significativa respecto al que se obtiene con las distribuciones paramétricas comúnmente usadas, y que el umbral superior estimado por el modelo es adecuado para el cálculo de extremos mediante el método de picos sobre el umbral. Se observa que al usar el umbral identificado por el modelo mixto, los intervalos de confianza de los cuantiles de alto período de retorno obtenidos son más estrechos que los que se obtienen con el umbral seleccionando mediante otros métodos disponibles. También se observa que el efecto de incluir la incertidumbre del umbral en el cálculo de los intervalos de confianza es despreciable desde el punto de vista práctico, al menos para los datos analizados.

Se propuso un modelo basado en el uso de copulas para modelar la dependencia temporal de series temporales univariadas. Comparado con modelos autorregresivos éste es un modelo más versátil, dado que existen diversidad de familias de copulas que se pueden utilizar para modelar la dependencia temporal. La aplicación del modelo basado en copulas a una serie temporal de alturas de ola significante resulta en que las simulaciones obtenidas con el mismo son mejores que las obtenidas usando modelos

autorregresivos.

Para la simulación de series temporales multivariadas se introdujo una metodología basada en el uso de distribuciones mixtas no estacionarias y distintos modelos autorregresivos vectoriales (un modelo estándar y dos modelos con cambio de régimen). Esta metodología fue utilizada con una serie pentavariada compuesta de altura de ola significante, período de pico y dirección media de oleaje y velocidad y dirección de viento.

De los resultados obtenidos se concluye que, aunque con ciertas limitaciones en cuanto a la representación del régimen de persistencias, los modelos autorregresivos vectoriales son adecuados para la simulación de series temporales multivariadas de variables oceanográficas y atmosféricas. Si bien los modelos autorregresivos vectoriales con cambio de régimen son capaces de reproducir estructuras de dependencia más complejas que las reproducidas por los modelos estándar, proporcionando una mejor representación de las distribuciones marginales bivariadas, también producen estructuras de autocorrelación que difieren de la de las variables originales, por lo que se recomienda precaución a la hora de su utilización en procesos de diseño, verificación u optimización.

El problema de la verificación y optimización de una sistema costero o portuario se abordó primero desde el punto de vista conceptual y teórico y luego desde el punto de vista práctico, trabajando sobre un caso de estudio.

En lo conceptual se definió la estructura genérica que debe tener un modelo de verificación y optimización de sistemas portuarios, y cómo se debe aplicar el mismo en los procesos de verificación y optimización del sistema. Luego, partiendo de trabajo previos y centrando la atención en las áreas navegables, se definieron los aspectos teóricos necesarios para el cálculo de la probabilidad de fallo y parada del sistema mediante técnicas de Nivel III, así como para el cálculo del riesgo en la vida útil.

La metodología desarrollada deja en evidencia que el cálculo del riesgo en la vida útil, en un sistema complejo como el analizado, no es una tarea inmediata. En primer lugar de no procederse con el debido rigor es fácil arribar a resultados erróneos, lo que hace que la metodología propuesta sea valiosa para futuras aplicaciones prácticas. En segundo lugar es de destacar el número y complejidad de estudios previos necesarios para implementar un modelo como el propuesto. Entre estos destaca la simulación aleatoria de series temporales estudiada en esta tesis, sin embargo otros aspectos quedan abiertos a futuros trabajos de investigación. Algunos de estos se discuten a continuación.

Líneas de trabajo

Finalizada esta tesis quedan abiertas diversas líneas de trabajo, tanto de investigación como de desarrollo tecnológico e innovación, las cuales se listan y discuten a continuación.

Simulación de series temporales

Distribuciones circulares

En este trabajo las variables dirección de viento y dirección de oleaje se trataron como si las mismas fuesen variables lineales. Sin embargo existen modelos específicos

para representar la probabilidad de ocurrencia de variables direccionales, denominados distribuciones circulares, el uso de los cuales debe ser investigado.

Modelos no estacionarios

En los modelos no estacionarios se han incluido ciclos de variación estacional y, en menor medida, ciclos de variación plurianual. Sin embargo no se han incluido covariables representativas del estado del sistema climático, lo que limita la capacidad de los modelos para representar la variabilidad interanual observada en las series temporales originales. Se debe trabajar en incluir esto en las funciones de distribución mixtas no estacionarias desarrolladas en esta tesis, así como en las distribuciones circulares antes mencionadas.

En cuanto a los ciclos de variación sí incluidos en los modelos, los mismas fueron representados mediante series de Fourier. Esta aproximación puede requerir de un gran número de componentes cuando los ciclos no se asemejan a funciones sinusoidales, lo que repercute en un elevado número de parámetros. Es interesante por tanto buscar métodos alternativos para representar aquellos ciclos que se apartan notoriamente de una sinusoide. Un método que resulta prometedor por su alta versatilidad en relación al número de parámetros que utiliza es el uso de polilíneas.

Simulación multivariada basada en copulas

En el caso de series univariadas, la metodología de simulación de series temporales basada en copulas produjo mejores resultados que la basada en los modelos AR y ARMA, con la ventaja adicional que implica la existencia de diversidad de familias de copulas entre las que seleccionar la de mejor ajuste. Sin embargo, extender esta metodología a series multivariadas no resulta trivial, ya que existe un gran número de distribuciones bivariadas que se deben respetar en el proceso de construcción de la copula multivariada.

Resulta por tanto una línea de trabajo interesante el desarrollo de un método de construcción de copulas, alternativo al aquí propuesto, que facilite la aplicación de esta herramienta en la simulación de series temporales multivariadas.

Dependencia temporal no estacionaria

Si bien en el capítulo 4 se utilizaron modelos estacionarios para modelar la dependencia temporal de las variables normalizadas, hay evidencia de que ésta podría no ser estacionaria, presentando al menos un ciclo anual.

Tener en cuenta la no estacionareidad de la dependencia temporal, ya sea investigando la forma de extender los modelos autorregresivos vectoriales a condiciones no estacionarias o desarrollando nuevos modelos, podría repercutir en una mejora de los resultados obtenidos, en particular en lo referente al régimen de persistencias y al comportamiento extremal de las series simuladas.

Herramientas de análisis y simulación de series de datos

Resulta evidente que para poder simular series temporales multivariadas el proyectista debe ajustar diversidad de modelos y decidir sobre cuáles son lo que mejor ajuste proporcionan en cada caso. De momento no se tiene conocimiento de una herramienta que permita realizar esta tarea de forma amigable.

Si bien es posible que a la fecha una herramienta de estas características no fuese de particular utilidad fuera de la academia, dado que también es limitada la oferta de

modelos amigables que permitan el estudio probabilista de sistemas costeros y portuarios –siendo quizás la principal excepción los modelos de propagación de oleaje y de evolución de línea de costa de tipo una línea–, sí es esperable que resulte de interés una vez combinada con modelos de análisis de sistemas costeros o portuarios con los que el ingeniero proyectista ya se encuentre familiarizado.

Verificación y optimización de sistemas

Desarrollo de herramientas generales

Para generalizar el uso de la metodología propuesta en el capítulo 5 es necesario trabajar en la creación de bases de datos de buques, costos, etc., y por sobre todo en la generación de un entorno de trabajo que facilite su aplicación por parte de las empresas proyectistas en distintos proyectos, sin que esto implique en cada caso la programación del modelo a medida y desde cero.

Modelos de piloto automático

En esta tesis se utilizó un modelo de piloto automático externo al modelo de verificación y optimización desarrollado. Sin embargo incluir un modelo de estas características dentro del modelo de verificación permitiría diseñar y optimizar con base en riesgo y de forma conjunta, tanto el ancho como el calado del canal de navegación.

Modelos de respuesta del buque

La calidad de los resultados obtenidos podría mejorarse al mejorar el modelo utilizado para calcular la respuesta dinámica del buque frente al oleaje, tanto durante el tránsito como durante las operaciones de carga y descarga. Para ello existen diversas líneas a explorar: (a) incluir en el modelo propuesto un modelo numérico de cálculo de la respuesta del buque de base física, (b) utilizar modelos numéricos de respuesta del buque no lineales, más adecuados que los lineales en condiciones de oleaje severo, y (c) utilizar modelos físicos a escala reducida para verificar y complementar los datos generados con los modelos numéricos.

Mantenimiento de los canales de navegación

En canales de navegación ubicados en zonas de fuerte transporte de sedimientos, como pueden ser estuarios o *coastal inlets*, uno de los principales costos en que se incurre en la vida útil es en el mantenimiento de los mismos (i.e. dragados periódicos). Para aplicar el modelo propuesto en estas condiciones es necesario incluir en el mismo la evolución morfológica del canal.

Conclusions

General conclusions

In order to deepens in the knowledge required for the verification, optimization and management of coastal and harbor systems, two general objectives were proposed in the introduction of this thesis: (a) to explore and develop methodologies for the simulation of time series of geophysical variables, and (b) to define and implement a methodology for the verification and optimization of the navigable areas of a port, based on Level III technics and risk analysis.

With regards to the first objective, a methodology was proposed for the simulation of multivariate time series that is of general application for geophysical variables. This methodology was applied to oceanographic and metheorologic variables. Starting with data series and following the steps described in this thesis –fitting stationary and non-stationary mixture distributions to the data, fitting models for the time dependance and interdependance of the variables–, one is able to simulate new random multivariate time series that retain the statistical characteristics of the original series.

The proposed methodology has several innovative aspects in the field of coastal and harbor engineering. Among them outstand the proposed parametric non-stationary mixture distributions, that are able to adequately represent all the range of values of the variables and its cycles of variation, and the proposal of a methodology for the simulation of time series that is based on the use of copulas.

Clearly the applicability of this methodology is not limited to the study case chosen for this thesis –the optimization of navigable areas of a port–. On the contrary, it has a wide range of possible applications for the verification of systems with different characteristics, such as conducting mid term morphology studies, flood risk analysis in coastal areas, or the optimización of maritime works by simulating its performance over its useful life, among others.

While the first objective is focused on understanding and modeling the predominant agents that force coastal and harbor systems, the second objective is focus on how to model the response of the system to the actions exerted by these agents. For this, a theoretical framework and a general working methodology were proposed that, by means of Monte Carlo simulations, is able to estimate the indexes that characterize the response of the system: its failure probability over the useful life, its mean anual operability, etc.

The developed methodology is innovative as it propose a framework that allows

to combine safety and operability aspects into a single calculation procedure, when in the literature these two aspects are worked separately. This allows to calculate the overall risk in the entire useful life of the system, which ultimately allows for an integral verification or optimization.

Lastly, the two methodologies developed –time series simulation and risk-based optimization of navigable areas– were applied to the Port of the Bay of Cádiz, for the verification and risk-based optimization of the depth of the entrance channel.

In summary this thesis presented a methodology for the integral verification and optimization of coastal and harbor systems on which the actions exerted by the oceanographic and atmospheric agents are predominant, with particular emphasis on the system formed by the navigable areas of a harbor.

Specific conclusions

The proposed mixture distribution, comprised by a central distribution (truncated or not, as appropriate) and two generalized Pareto distributions for the tails, is an efficient tool for modeling the probability distribution of a geophysical variable over its entire range of values, both in stationary conditions, assuming a mean annual distribution, as in non-stationary conditions, where it is possible to accommodate climate cycles, trends and covariates.

In turn, the mixture model automatically provides the high threshold required to implement the method of peaks over the threshold, as well as the uncertainty of the threshold. A simple methodology is proposed to include this uncertainty in the estimation of confidence intervals of high return period quantiles.

The model was tested against six data series: two of daily precipitation, two of mean daily flow, and two of significant wave height. In all cases the proposed model improved the fit of the data relative to the fit obtained with commonly used distributions, and the upper threshold estimated by the model was adequate to apply the peaks over the threshold method for extremes calculation. It is noted that using the threshold identified by the mixture model, the confidence intervals obtained for high return period quantiles are narrower than those obtained with the threshold selected by means of other methods. It was also observed that the effect of including the uncertainty of the threshold in the calculation of confidence intervals is negligible from a practical point of view, at least for the analyzed data.

A model was proposed, based on the use of copulas, for modeling the time dependence of univariate time series. Compared to autoregressive models this is a more versatile methodology, since there are various families of copula that can be used to model the time dependence. The application of the model to a data series of significant wave height indicated that the simulations obtained via the copula-based time-dependance model were better than those obtained using an autoregressive moving average model.

For the simulation of multivariate time series a methodology was introduces, that is based on the use of non-stationary mixture distributions and different vector autoregressive models (one standard model and two regime switching models). This methodology

was tested with a 5-variate serie composed of significant wave height, mean wave direction, peak period and mean wind speed and direction.

From the obtained results it is concluded that, with some limitations on the representation of persistence regimes, vector autoregressive models are suitable for the simulation of multivariate time series of oceanographic and atmospheric variables. While regime switching vector autoregressive models are able to reproduce more complex dependency structures than reproduced by standard models, providing a better representation of the bivariate marginal distributions, they also produce autocorrelation structures that differ from that of the original variables, so it is recommended to be cautious in its use during design, verification or optimization processes.

The problem of verification and optimization of a coastal and harbor system was first addressed conceptually and theoretically, and then from the practical point of view, working on a case study.

On a conceptual level a generic structure for the model was defined, as well as the way it should be applied during verification and optimization processes. Then, based on previous works and focusing on navigable areas, a theoretical framework was defined, specifying how to calculate the failure probability and the operability of a system by means of Level III techniques, and how to calculate the overall risk on the useful life of the system.

The proposed methodology shows that the estimation of the overall risk on the useful life of a complex system, as the one analyzed in this thesis, is not trivial. First, if not proceeding with due rigor, it is easy to achieve erroneous results, which makes the proposed methodology valuable for future practical applications. Second, it is important to highlight the number and complexity of previous studies required to implement a model like the one proposed. Among these are the methodologies for time series simulation studied in this thesis, but other aspects are open to future research. Some of these aspects are discussed below.

Open lines

Finished this thesis, there are several open lines to work in both research, and technological development and innovation, which are listed and discussed below.

Time series simulation

Circular distributions

Throughout this work directional variables, like wind or wave direction, were treated as linear variables. However there exist distribution models that are specific designed for circular variables, which use should be explored.

Non-stationary distributions

Seasonal cycles and, to a lesser extent, pluriannual cycles, were included in the non-stationary distributions. However, no covariates were included representing the state of the global or regional climate, which limits the ability of simulations to reproduce the interannual variability observed in the original series. It is required to explore the

inclusion of these covariates in the non-stationary mixture distributions developed in this thesis, as well as in the circular distributions mentioned above.

As to the cycles that are included in the model, they were represented by Fourier series. This approach may require a large number of components when the cycles are not similar to sinusoidal functions, which results in a large number of parameters. It is interesting therefore to seek for alternative methods to represent those cycles that deviate markedly from a sinusoid. One method, that is promising for its high versatility in terms of the number of parameters that uses, is the use of polylines.

Multivariate time series simulation based on copulas

For univariate time series, the simulation methodology based on the use of copulas produced better results than that based on AR and ARMA models, with the additional advantage that implies the existence of several copulas families, among which select the one with the best fit. However, to extend this methodology to multivariate time series is not trivial, since there is a large number of bivariate distributions that must be preserved in the process of building the multivariate copula.

It is therefore interesting to explore the development of a copulas construction method, alternative to the one proposed here, that may facilitate the use of copulas for the simulation of multivariate time series.

Non-stationary time dependance models

While in Chapter 4 stationary models were used to model the time dependence of the normalized variables, there is evidence that the time dependance may not be stationary, presenting at least an annual or seasonal cycle.

Take into account the non stationarity of the time dependence, either by extending the vector autoregressive models to non-stationary conditions or by developing new models, may improve the results achieved, particularly in regard with the persistence regimes and the extremal behavior of the simulated series.

Tools for the analysis and simulation of time series

It is clear that in order to simulate multivariate time series the analyst must fit a variety of models, and decide which are the ones that gives the best in each case. To the knowledge of the author there is no tool that allows to perform this analysis within a friendly environment.

While it is possible that at the moment a tool with these characteristics may not be particularly useful outside academia, as there is a limited offer of user friendly models that allows the probabilistic study of coastal and harbor systems –perhaps being the main exception the wave propagation and one-line shoreline evolution models–, it is expected that such tool becomes of interest if combined with models for coastal and harbor analysis with which the engineer is already familiarized.

Systems verification and optimization

Development of user-friendly tools and databases

To generalize the use of the methodology proposed in Chapter 5 it is necessary to work in the creation of different databases: of ships, costs, etc., and above all in the development of a user-friendly work environment that facilitate the application of the

methodology in different projects, without having to program the model from zero for each one.

Autopilot (fast-time) simulation models

In this thesis an autopilot (fast-time) model was used that is external to the proposed verification and optimization model. However, the incorporation of an autopilot model within the verification and optimization model will allow for the risk-based verification and optimization of both, width and depth of the navigable ares, at the same time.

Ship response models

The quality of the results could be improved by improving the model used to calculate the dynamic response of the ship to the waves, both for ships in transit and during loading and unloading. For this there are several lines to explore: (a) to include within the proposed model, a physical based numerical model for calculating the response of the ship, (b) using non-linear numerical models to calculate the response of the ship, most appropriate that linear models in severe wave conditions, and (c) use small-scale physical models to verify and supplement the data generated by numerical models.

Maintenance of navigation channels

For navigation channels located in areas of high sediment transport, such as estuaries and coastal inlets, one of the major costs incurred during the useful life of the channels is on maintenance (i.e. periodic dredging). To apply the proposed model in these environments it is necessary to include into the model the morphological evolution of the channels.