Non-stationary wave height climate modelling and simulation

S. Solari,[12] and M.A. Losada,[1]

S. Solari, Grupo de Dinámica de Flujos Ambientales, Universidad de Granada, Av. del Mediterráneo s/n, Granada, 18006, Spain. (ssolari@ugr.es)

M. A. Losada, Grupo de Dinámica de Flujos Ambientales, Universidad de Granada, Av. del Mediterráneo s/n, Granada, 18006, Spain. (mlosada@ugr.es)

[1]Grupo de Dinámica de Flujos
Ambientales, Universidad de Granada,
Spain.

[2]Centro Interdisciplinario para el Manejo
Costero Integrado del Cono Sur,
Universidad de la República, Uruguay.

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.

1. 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.

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 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 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.

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 2, the proposed model is presented together with the procedure for simulating new time series. In Section 3, 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 4, 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 G).

2. Methodology

The non-stationary model (Section 2.1) 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 2.2) 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 2.3) and the structure of the ARMA models (Section 2.4), which are used to compare the results obtained with those obtained using the copula-based time-dependence model.

2.1. Non-stationary distribution function

Solari and Losada [2011a, b] present a mixture model

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

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

$$f_m(x|x < u_1) = \frac{1}{\sigma_1}\left(1 - \frac{\xi_1}{\sigma_1}(x - u_1)\right)^{-\frac{1}{\xi_1}-1} \quad \xi_1 \neq 0 \quad (2a)$$

$$f_M(x|x > u_2) = \frac{1}{\sigma_2}\left(1 + \frac{\xi_2}{\sigma_2}(x - u_2)\right)^{-\frac{1}{\xi_2}-1} \quad \xi_2 \neq 0 \quad (2)$$

with

$$\sigma_1 = -\xi_1 u_1 \qquad \xi_1 = -\frac{F_c(u_1)}{u_1 f_c(u_1)} \qquad \sigma_2 = \frac{1 - F_c(u_2)}{f_c(u_2)} \quad (3)$$

This model is similar to that proposed by *Cai et al.* [2007] for ARMA models with the exception that in (1), the continuity of the probability density function is assured by the conditions presented in (3). Furthermore, *Cai et al.* [2007] do not provide a method of threshold estimation, whereas *Solari and Losada* [2011a, b] 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:

$$\theta(t) = \theta_{a0} + \sum_{k=1}^{N} \left(\theta_{ak}\cos(2\pi k t) + \theta_{bk}\sin(2\pi k t)\right) \quad (4)$$

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 and Losada*, 2011a]. 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,

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

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

$$P(H \leq H^*) = \frac{1}{T}\int_t^{t+T} P(H \leq H^*|t)dt \tag{5}$$

where $P(H \leq H^*|t)$ is the non-stationary LN-GPD model (1) (NS-LN-GPD):

$$P(x_t|t) = \begin{cases} F_m(x_t|t)\Phi(Z_1) & x_t < u_1(t) \\ F_c(x_t|t) & u_1(t) \leq x_t \leq u_2(t) \\ \Phi(Z_2) + F_M(x_t|t)\,(1-\Phi(Z_2)) & x_t > u_2(t) \end{cases} \tag{6}$$

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

2.2. Temporal dependence

The NS-LN-GPD model (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

$$C_{12}(u, v) = Prob[P_t \leq u, P_{t-1} \leq v] \tag{7}$$

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:

$$C_{1|2}(u, v) = Prob[P_t \leq u \,|\, P_{t-1} = v] = \frac{\partial C_{12}}{\partial v}(u, v) \tag{8}$$

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

Given copula $C_{1\ldots k}$ (which defines the joint probability of k successive states) and, consequently, given the Markov model of order $k-1$, variables $F_{1|2\ldots k} = Prob[P_t|P_{t-1}, \ldots, P_{t-k+1}]$ and $F_{k+1|2\ldots k} = Prob[P_{t-k}|P_{t-1}, \ldots, P_{t-k+1}]$ are constructed. The dependence between two variables is measured using Kendall's τ_k or Spearman's ρ_s statistic (see Appendix C). 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 $C_{1\ldots k+1}$

$$
\begin{aligned}
&C_{1...k+1}(u_1, ..., u_{k+1}) \\
&\quad = Prob[P_t \leq u_1, ..., P_{t-k} \leq u_{k+1}] \\
&\quad = \int_{-\infty}^{u_2} ... \int_{-\infty}^{u_k} C_{1k+1}(F_{1|2...k}, F_{k+1|2...k}) \\
&\qquad C_{2...k}(dx_2, ..., dx_k)
\end{aligned} \tag{9}
$$

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 D 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 $C_{1...k}$.

2.2.1. 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 E). 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.

2.3. Simulation methodology

The simulation process consists of two parts. First, the time-dependence model of copulas (9) is used to obtain the series of probabilities $\{P_t\}$; then, the non-stationary model (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

$$\begin{aligned} u_t =& \frac{\partial C_{1\ldots k}}{\partial u_2 \ldots \partial u_k}(P_t, \ldots, P_{t-k+1}) \\ =& \frac{\partial C_{1k}}{\partial F_{k|2\ldots k-1}}\Big(F_{1|2\ldots k-1}(P_t, \ldots, P_{t-k+2}), \\ & F_{k|2\ldots k-1}(P_{t-1}, \ldots, P_{t-k+1})\Big) \end{aligned} \tag{10}$$

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 (9) to perform the simulations because (10) can be resolved using the bivariate copulas. To obtain P_t, equation (10) can be numerically solved using the bisection method. The simulation process for a third-order Markov model is described in Appendix F.

2.4. ARMA models

An ARMA(p,q) model is given by

$$Z_t = \phi_1 Z_{t-1} + \ldots + \phi_p Z_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} \tag{11}$$

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 (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 (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 (6).

3. 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 and Losada* [2011b].

3.1. 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 (4)); interannual variation, covariables and trends were not considered.

Fourier series are evaluated (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 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 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 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 2 and 3 show the estimated NS-LN-GPD and NS-LN parameters, respectively.

Figure 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. Generally 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 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 5 (the bottom graph) shows the annual PDF. The NS-LN-GPD model fits the mode better than the NS-LN model.

Finally, Figure 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.

3.2. 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].

When the moving average of the data is displayed on a graph (Figure 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:

$$\mu_{LN,anual} = a_{i1}\cos(2\pi t/5) + b_{i1}\sin(2\pi t/5) + a_{i2}\cos(2\pi t/26) + b_{i2}\sin(2\pi t/26) \quad (12)$$

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} (12).

Figure 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 5 and 6 and therefore not shown here).

3.3. 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 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 (E1) 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 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 $F_{3|2}$. 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 (D2), 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 (E8) 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 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. 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.

3.5. 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 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.

3.5.1. 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 11 presents the results only for model SM-C. By comparing Figure 11 with Figure 5 , it is clear that the distribution of the simulated data series (Figure 11)

is equal to the theoretical distribution (Figure 5). This finding is because the simulated series is very long (500 years).

Table 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 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 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 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 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.

3.5.2. 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 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 and Losada* [2011b]; 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 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 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

[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.

models tended to overestimate the frequency of short durations (approx. 5 hours), and underestimate frequency of long durations (approx. 30 hours).

Finally, Figure 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 and Losada* [2011b] 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.

3.6. 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.

4. 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.

Appendix A: Data standardization

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

$$Z_e = \begin{cases} Z1 - Z_{min} & H(t) < u_1(t) \\ Z_{LN} & u_1(t) \leq H(t) \leq u_2(t) \\ Z_2 + Z_{max} & H(t) > u_2(t) \end{cases} \tag{A1}$$

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

$$Z_{LN} = \frac{\log(H(t)) - \mu_{LN}(t)}{\sigma_{LN}(t)} \tag{A2}$$

$$Z_{min} = \frac{1}{\xi_1(t)} \log\left(1 - \frac{\xi_1(t)}{\sigma_1(t)} \left(H(t) - u_1(t)\right)\right) \tag{A3}$$

$$Z_{max} = \frac{1}{\xi_2(t)} \log\left(1 + \frac{\xi_2(t)}{\sigma_2(t)} \left(H(t) - u_2(t)\right)\right) \tag{A4}$$

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

$$Z_m(P_e) = \begin{cases} Z_1 + \log(P_e/\Phi(Z_1)) & P_e < \Phi(Z_1) \\ \Phi^{-1}(P_e) & \Phi(Z_1) \leq P_e \leq \Phi(Z_2) \\ Z_2 - \log(1 - \frac{P_e - \Phi(Z_2)}{1 - \Phi(Z_2)}) & P_e > \Phi(Z_2) \end{cases} \tag{A5}$$

$$P_m(Z_e) = \begin{cases} \Phi(Z_1)\exp(Z_e - Z_1) & Z_e < Z_1 \\ \Phi(Z_e) & Z_1 \leq Z_e \leq Z_2 \\ \Phi(Z_2) + (1 - \Phi(Z_2))(1 - \exp(Z_2 - Z_e)) & Z_e > Z_2 \end{cases} \tag{A6}$$

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

Appendix B: 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 $F_{XY} = Prob[X \leq x, Y \leq y] = C(F_X(x), F_Y(y))$.

Appendix C: 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

$$\tau_k = \frac{c - d}{c + d} \tag{C1}$$

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n}(R_i - S_i)^2}{n^3 - n} \tag{C2}$$

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.

Appendix D: 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:

$$F_{1|2}(u, v) = Prob[P_t \leq u \,|\, P_{t-1} = v] = \frac{\partial C_{12}}{\partial v}(u, v) \tag{D1a}$$

$$F_{3|2}(v, w) = Prob[P_{t-2} \leq w \,|\, P_{t-1} = v] = \frac{\partial C_{23}}{\partial v}(v, w) \tag{D1b}$$

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]

$$C_{123}(u, v, w) = \int_{-\infty}^{v} C_{13}(F_{1|2}(u, x), F_{3|2}(x, w)) F_2(dx) \tag{D2}$$

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}

$$\begin{aligned} F_{1|23}(u,v,w) =& Prob[P_t \leq u \,|\, P_{t-1}=v, P_{t-2}=w] \\ =& \frac{\partial^2 C_{123}}{\partial v \partial w} \Big/ \frac{\partial^2 C_{23}}{\partial v \partial w} \\ =& \frac{\partial C_{13}}{\partial F_{3|2}}(F_{1|2}(u,v), F_{3|2}(v,w)) \end{aligned} \tag{D3a}$$

$$\begin{aligned} F_{4|23}(v,w,y) =& Prob[P_{t-3} \leq y \,|\, P_{t-1}=v, P_{t-2}=w] \\ =& \frac{\partial^2 C_{123}}{\partial u \partial v} \Big/ \frac{\partial^2 C_{12}}{\partial u \partial v} \\ =& \frac{\partial C_{24}}{\partial F_{2|3}}(F_{2|3}(v,w), F_{4|3}(w,y)) \end{aligned} \tag{D3b}$$

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

$$\begin{aligned} & C_{1234}(u,v,w,y) \\ & = Prob[P_t \leq u, P_{t+1} \leq v, P_{t+2} \leq w, P_{t+3} \leq y] \\ & = \int_{-\infty}^{w} \int_{-\infty}^{v} C_{14}(F_{1|23}(u,x_1,x_2), F_{4|23}(x_1,x_2,y)) \\ & \quad C_{23}(dx_1, dx_2) \end{aligned} \tag{D4}$$

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 (D4)

$$
\begin{aligned}
&C_{1|234}(u,v,w,y) \\
&= Prob[P_t \le u \,|\, P_{t-1}=v, P_{t-2}=w, P_{t-3}=y] \\
&= \frac{\partial^3 C_{1234}}{\partial v \partial w \partial y} \Big/ \frac{\partial^3 C_{234}}{\partial v \partial w \partial y} \\
&= \frac{\partial C_{14}}{\partial F_{4|23}}(F_{1|23}(u,v,w), F_{4|23}(v,w,y))
\end{aligned} \tag{D5}
$$

Appendix E: 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].

$$
C_{12}(u,v) = Prob[x \le u, y \le v] = \exp\{-V(u,v)\} \tag{E1}
$$

with

$$
V(u,v) = (1-\theta_1)\hat{u} + (1-\theta_2)\hat{v} + \left[(\theta_1\hat{u})^\theta + (\theta_2\hat{v})^\theta\right]^{1/\theta} \tag{E2}
$$

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

The conditioned distributions are given by

$$
\begin{aligned}
C_{1|2}(u,v) =& Prob[x \le u \,|\, y = v] = \frac{\partial C}{\partial v}(u,v) \\
=& \frac{C(u,v)}{v}\left[1 - \theta_2 + \theta_2\left(1 + \frac{(\theta_1\hat{u})^\theta}{(\theta_2\hat{v})^\theta}\right)^{\frac{1}{\theta}-1}\right]
\end{aligned} \tag{E3}
$$

$$
\begin{aligned}
C_{2|1}(u,v) =& Prob[y \le v \,|\, x = u] = \frac{\partial C}{\partial u}(u,v) \\
=& \frac{C(u,v)}{u}\left[1 - \theta_1 + \theta_1\left(1 + \frac{(\theta_2\hat{v})^\theta}{(\theta_1\hat{u})^\theta}\right)^{\frac{1}{\theta}-1}\right]
\end{aligned} \tag{E4}
$$

whereas the density is

$$\begin{aligned} c_{12}(u,v) = Prob[x=u, y=v] &= \frac{\partial^2 C_{12}}{\partial u \partial v}(u,v) \\ &= \frac{C(u,v)}{uv}\Big\{ \left[C_{1|2}(u,v)\right]\left[C_{2|1}(u,v)\right] + \\ &\quad \theta_1\theta_2(\theta-1)\left((\theta_1\hat{u})^\theta + (\theta_2\hat{v})^\theta\right)^{\frac{1}{\theta}-2}(\theta_1\theta_2\hat{u}\hat{v})^{\theta-1}\Big\} \end{aligned} \tag{E5}$$

The parameters of this copula are estimated by means of maximum likelihood using (E5).

The Fréchet copula family is given by

$$C_{12}(u,v) = \alpha M_2(u,v) + (1-\alpha-\beta)\Pi_2(u,v) + \beta W_2(u,v) \tag{E6}$$

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]

$$\tau_K(\alpha,\beta) = \frac{(\alpha-\beta)(\alpha+\beta+2)}{3} \tag{E7}$$

$$\rho_S(\alpha,\beta) = \alpha - \beta \tag{E8}$$

Appendix F: 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

$$u_2 = C_{2|1}(P_1, P_2) \tag{F1}$$

(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

$$u_3 = C_{3|1}\Big(C_{1|2}(P_1, P_2), C_{3|2}(P_2, P_3)\Big) \tag{F2}$$

(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

$$\begin{aligned} u_t = & C_{4|1}\Bigg(C_{1|23}\Big(C_{1|2}(P_{t-3}, P_{t-2}), C_{3|2}(P_{t-2}, P_{t-1})\Big), \\ & C_{4|23}\Big(C_{2|3}(P_{t-2}, P_{t-1}), C_{4|3}(P_{t-1}, P_t)\Big)\Bigg) \end{aligned} \tag{F3}$$

(v) Once the series $\{P_t\}$is simulated, the series $\{H_t\}$ is constructed, using the inverse of the NS-LN-GPD (6).

In steps ***(ii)*** to ***(iv)***, the expressions of the conditioned copulas are analytically resolved, whereas equations (F1), (F2) and (F3) are numerically solved with the bisection method.

Appendix G: List of abbreviations

Table 6 lists the abbreviations used throughout the article.

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.

Solari, S., and M. A. Losada (2011a), Unified distribution models for climate variables. Part I: Description, *Submitted to Coastal Engineering*.

Solari, S., and M. A. Losada (2011b), Unified distribution models for climate variables. Part II: Application to a series of significant wave height, *Submitted to Coastal Engineering*.

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.

Figure 1. Physical phenomena evolving in different time scales, and statistical models for the appropriate modelling of the sea-state variables.

Figure 2. Minimum Bayesian Information Criterion obtained for different numbers of parameters in the NS-LN-GPD model, with maximum approximation of the fourth order ($\bigcirc$), 6th order ($\triangle$) and 9th order ($\square$).

Figure 3. Time evolution of μ_{LN}, σ_{LN} and ξ_2 for the NS-LN-GPD [4,2,2] model.

Figure 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).

Figure 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 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).

Figure 7. Ninety-day Moving Average of H_s and the $\mu_{LN}(t)$ parameter of interannual model.

Figure 8. Mean and standard deviation of P_t, estimated on an annual scale for each state, and their moving average smooth curves.

Figure 9. Empirical copula $C(P_t, P_{t-1})$ (thick line) and asymmetric Gumbel-Hougaard copula (thin line).

Figure 10. Five years of measured significant wave heights (top) and simulated significant wave heights (bottom).

Figure 11. Accumulated probability on log-normal paper (top graph) and probability density (bottom graph). Original (dots) and simulated (green line) data series.

Figure 12. Autocorrelation function (ACF) for the four dependence models used and for a simulation run using an ARMA(3,3) model.

Figure 13. Persistence over thresholds 0.5, 1, 1.5, 2, 2.5 and 3m.

Figure 14. Annual maxima H_s: empirical data (dots), data from the copula models (green lines) and data from the ARMA models (blue lines).

Figure 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 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 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).

Table 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

Table 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

Table 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%)			

Table 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.

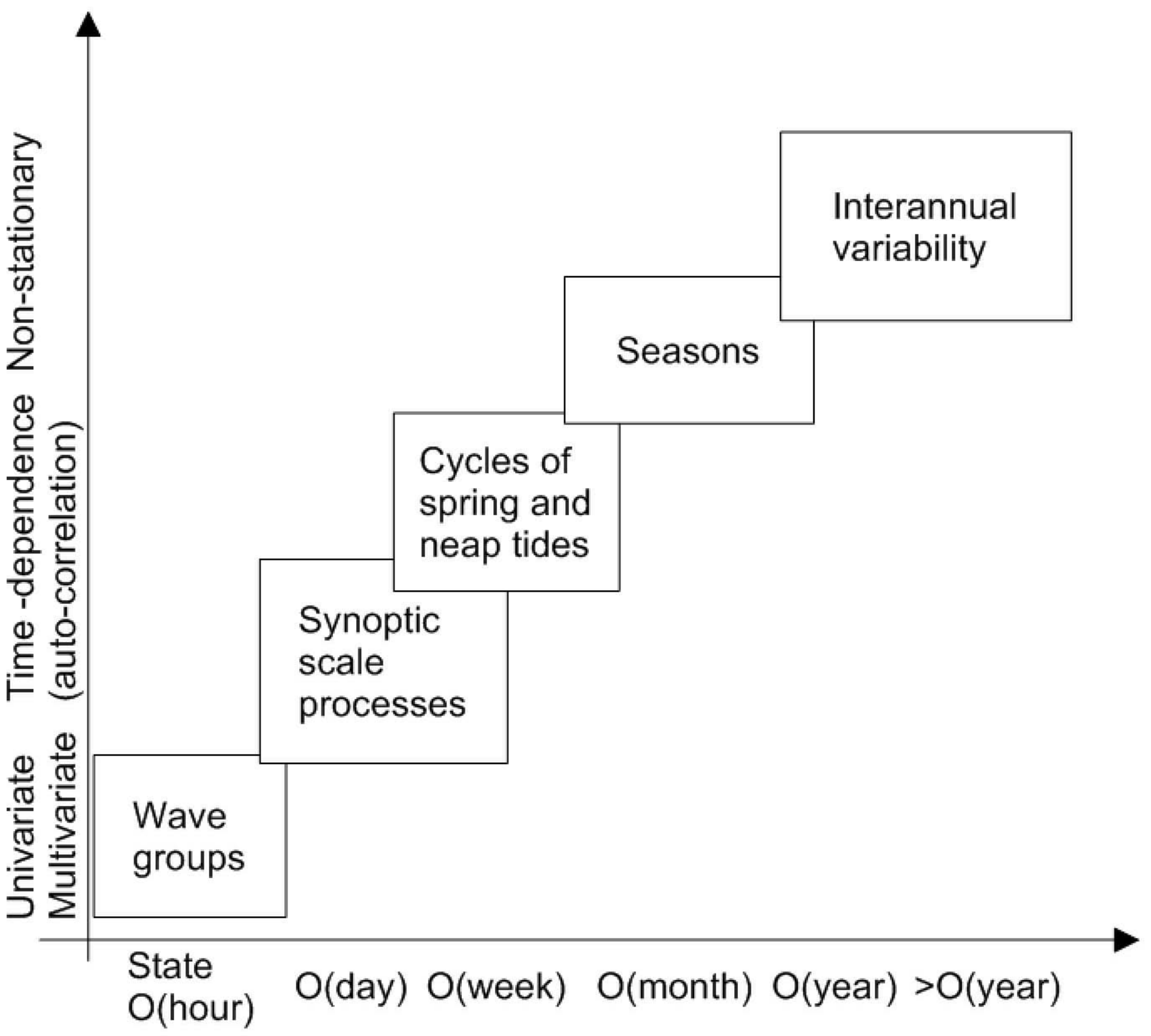
	C_{12} G-H Asim.			C_{13} Fréchet		C_{14} Fréchet	
	θ	θ_1	θ_2	α	β	α	β
SM	5.697	0.995	0.971	0	0.194	0.005	0
IM	5.462	0.994	0.969	0	0.192	–	–

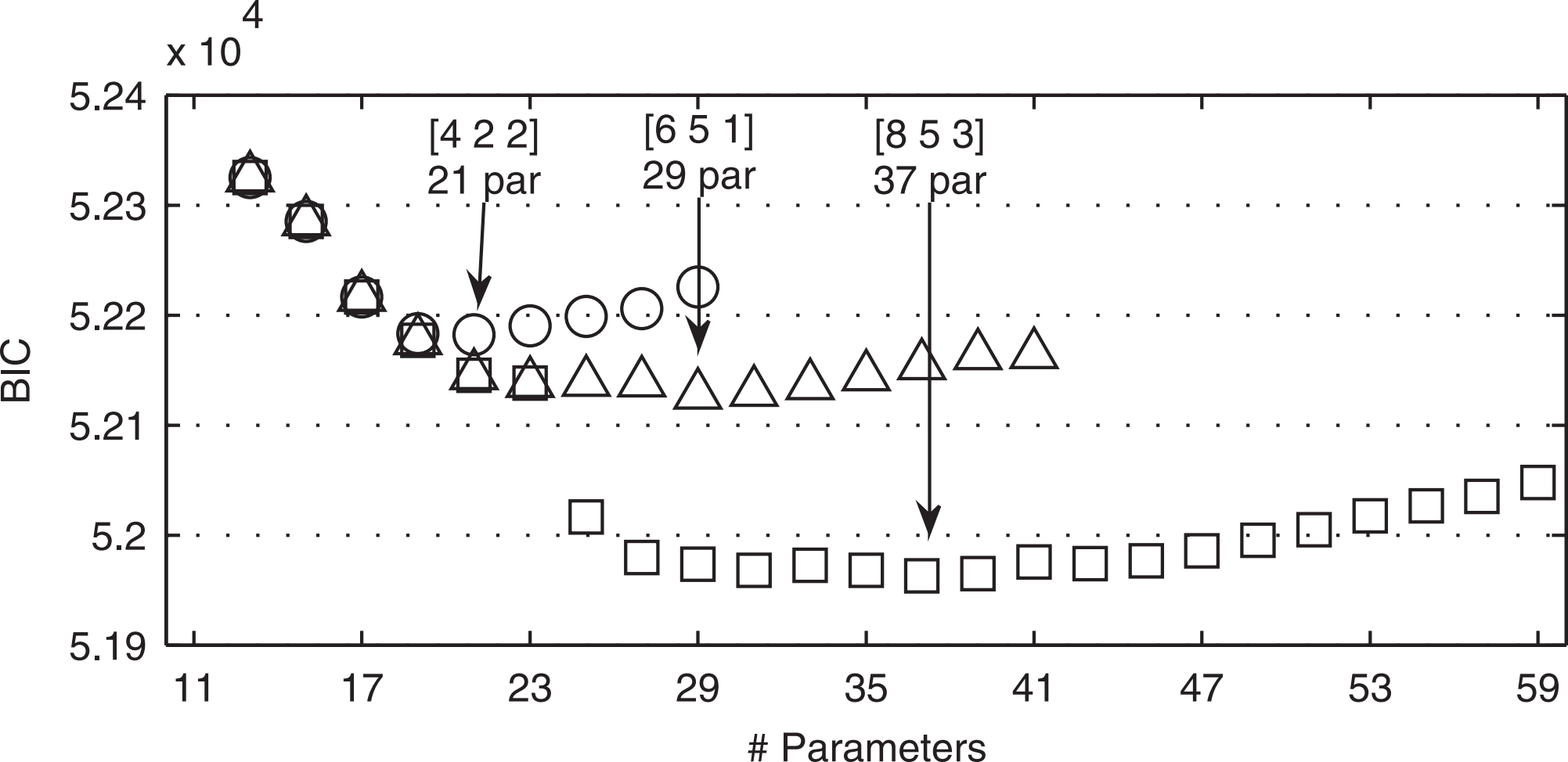
Table 5. Statistics obtained from the first four central moments.

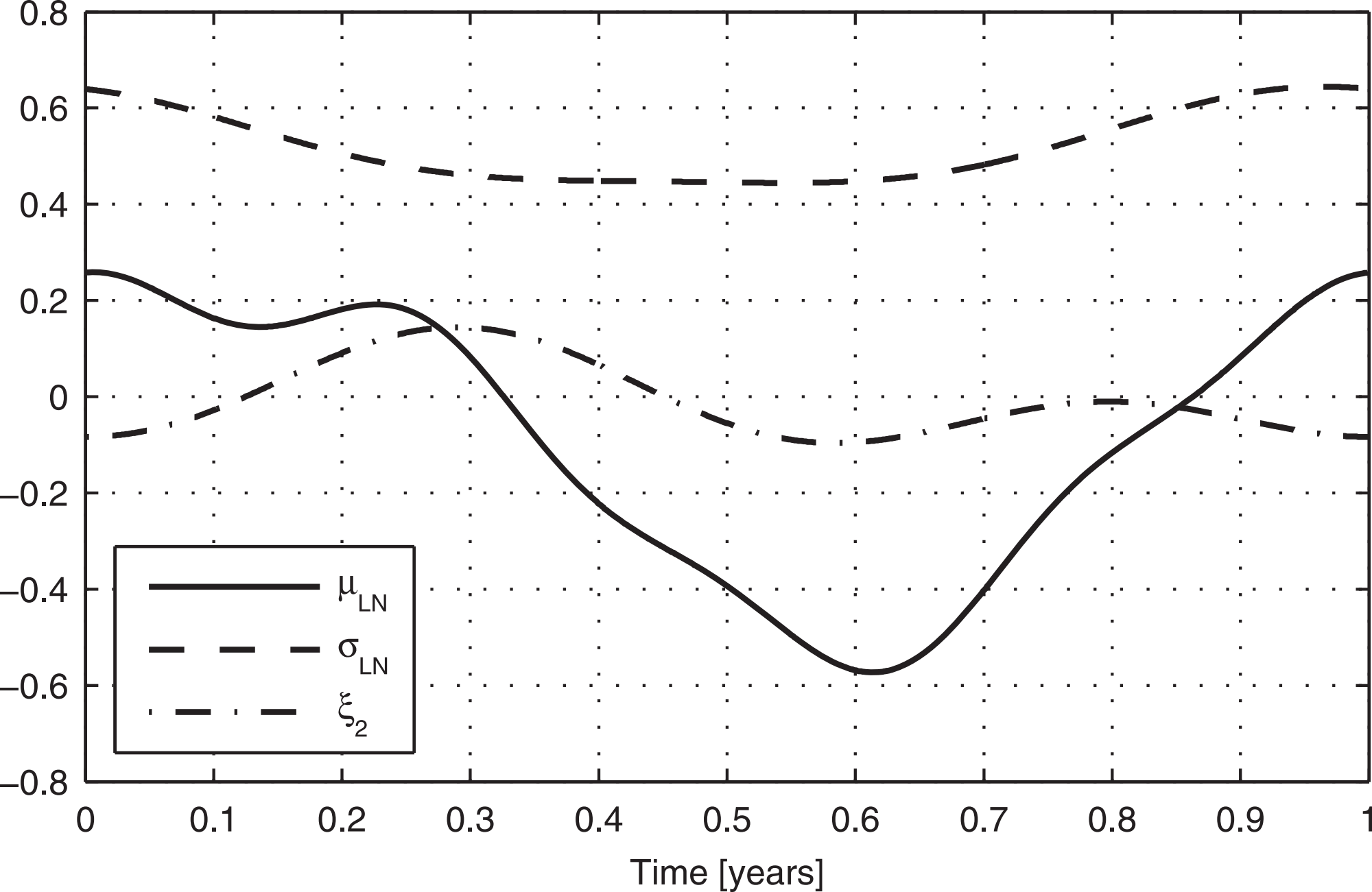
	Data	SM-C	IM-C	SM-A	IM-A
Mean	1.088	1.077	1.086	1.090	1.093
Variance	0.548	0.521	0.538	0.539	0.556
Skewness	2.127	2.106	2.275	2.159	2.410
Kurtosis	10.006	10.468	12.290	10.846	14.326

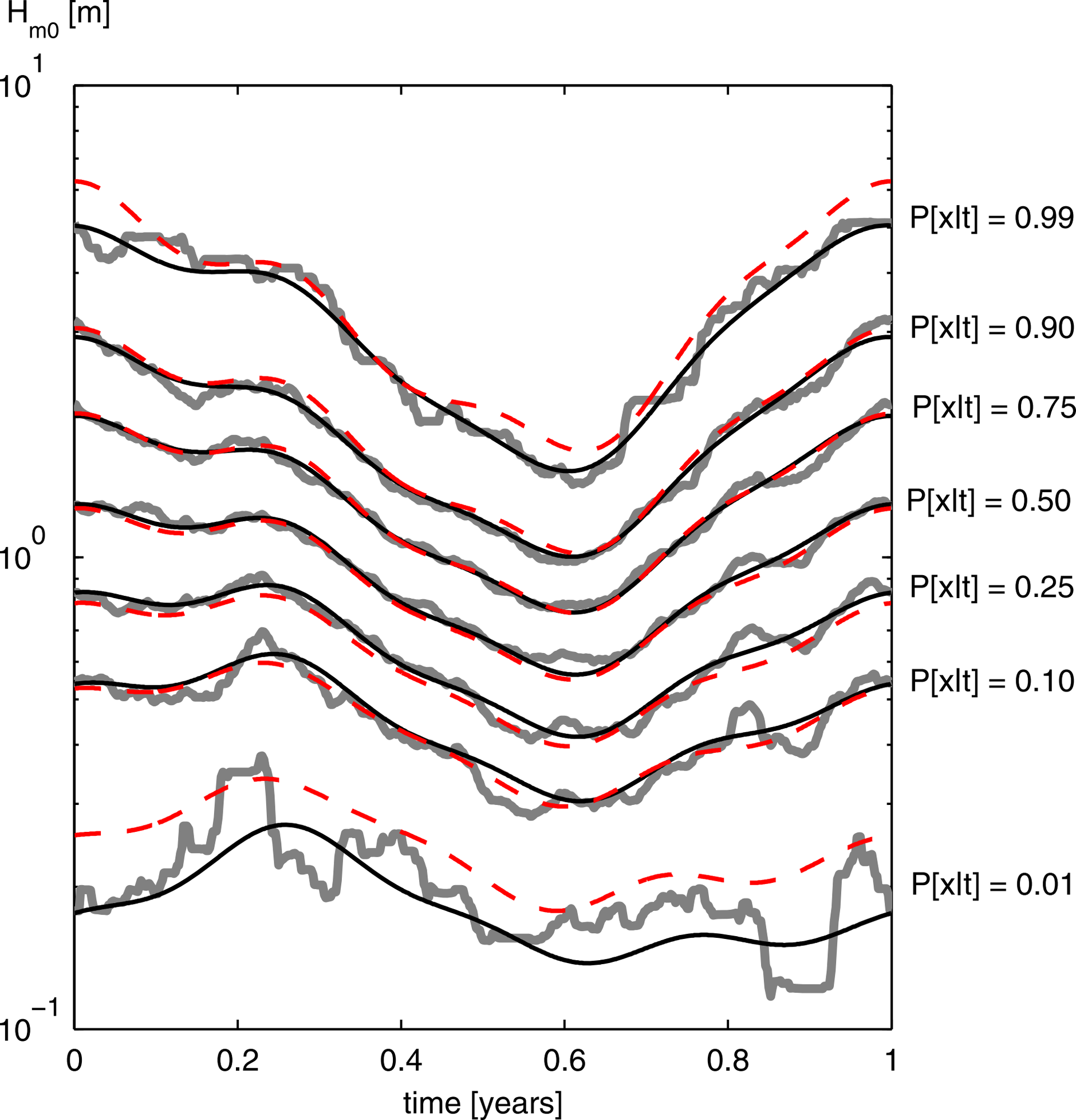
Table 6. List of abbreviations.

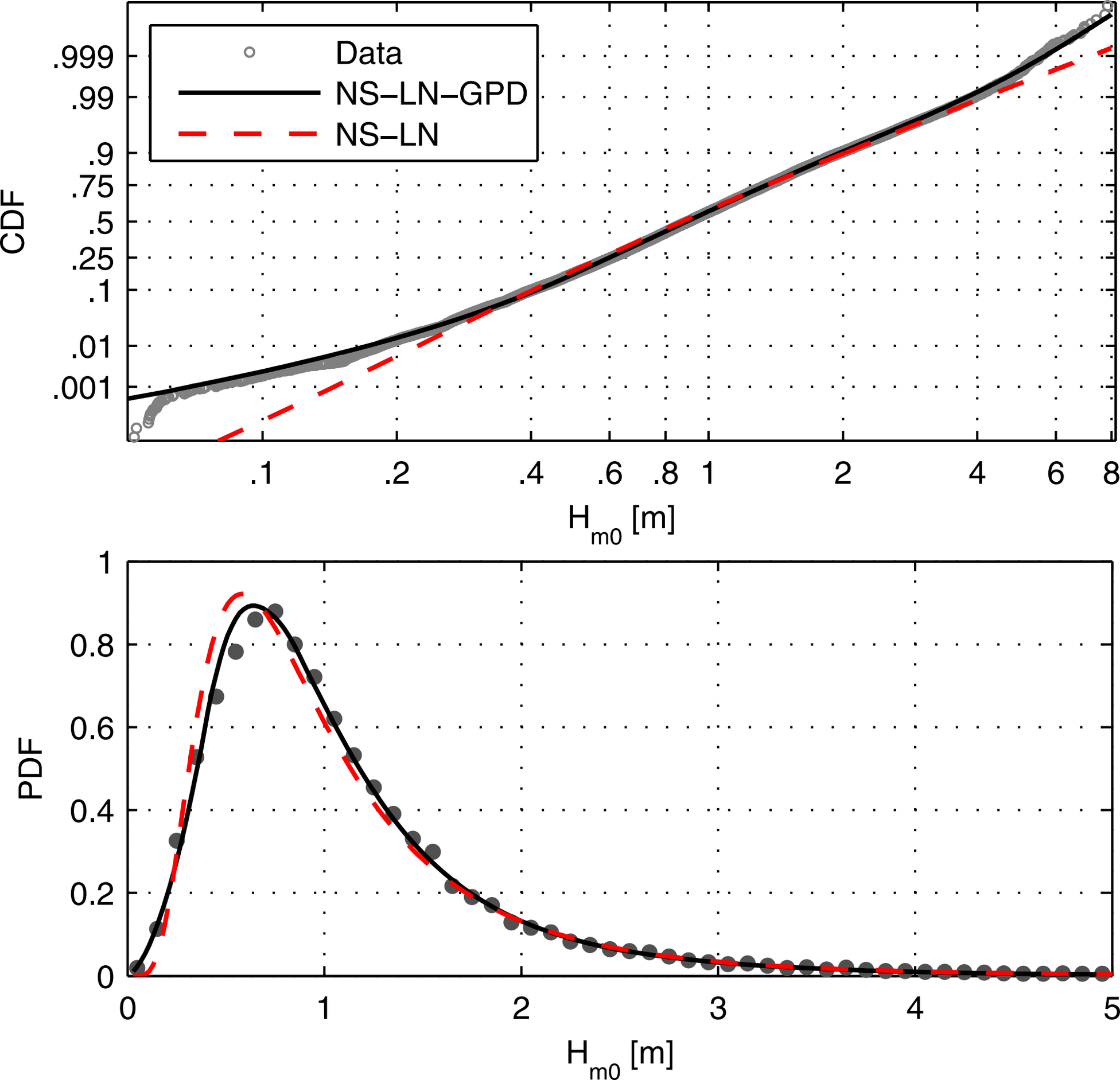
Abbreviation	Description
BIC	Bayesian Information Criterion
GPD	Generalised Pareto distribution
IM	NS-LN-GPD model fitted to the data allowing the parameters to have interannual variations
IM-A	Combination of IM model for marginal distribution and ARMA model for time dependency
IM-C	Combination of IM 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

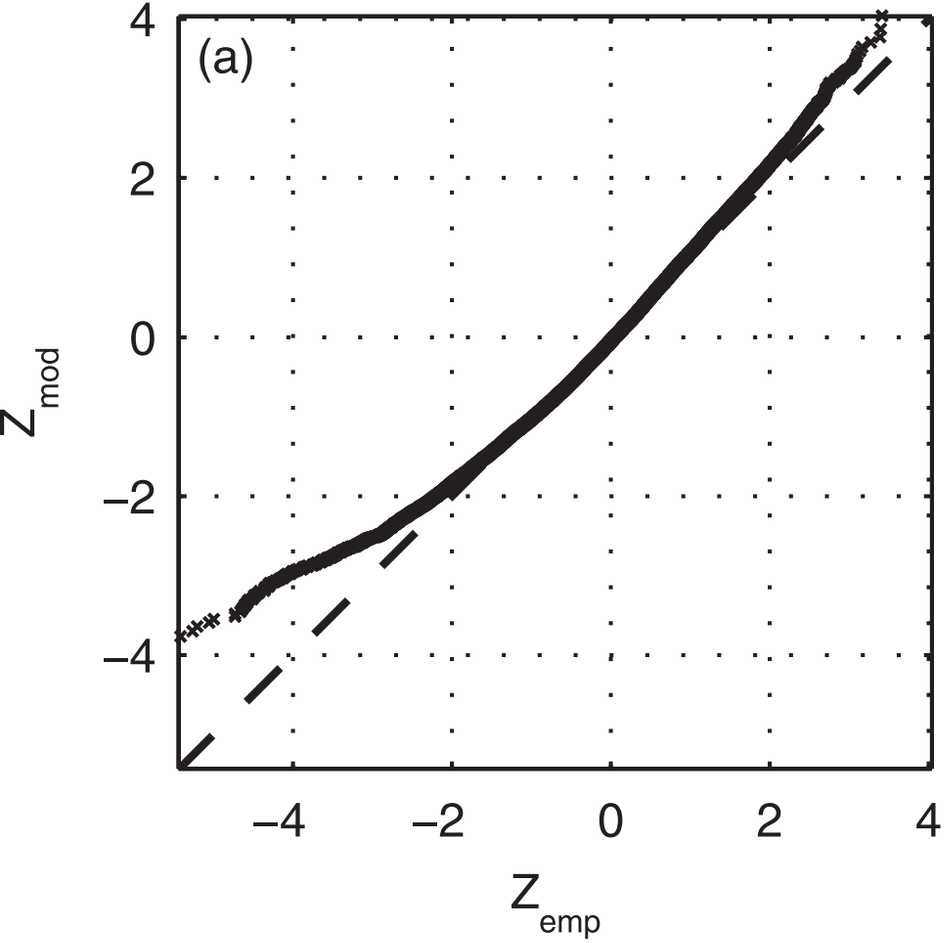

Univariate Multivariate
Time -dependence (auto-correlation)
Non-stationary
Wave groups
Synoptic scale processes
Cycles of spring and neap tides
Seasons
Interannual variability
State O(hour)
O(day)
O(week)
O(month)
O(year)
>O(year)

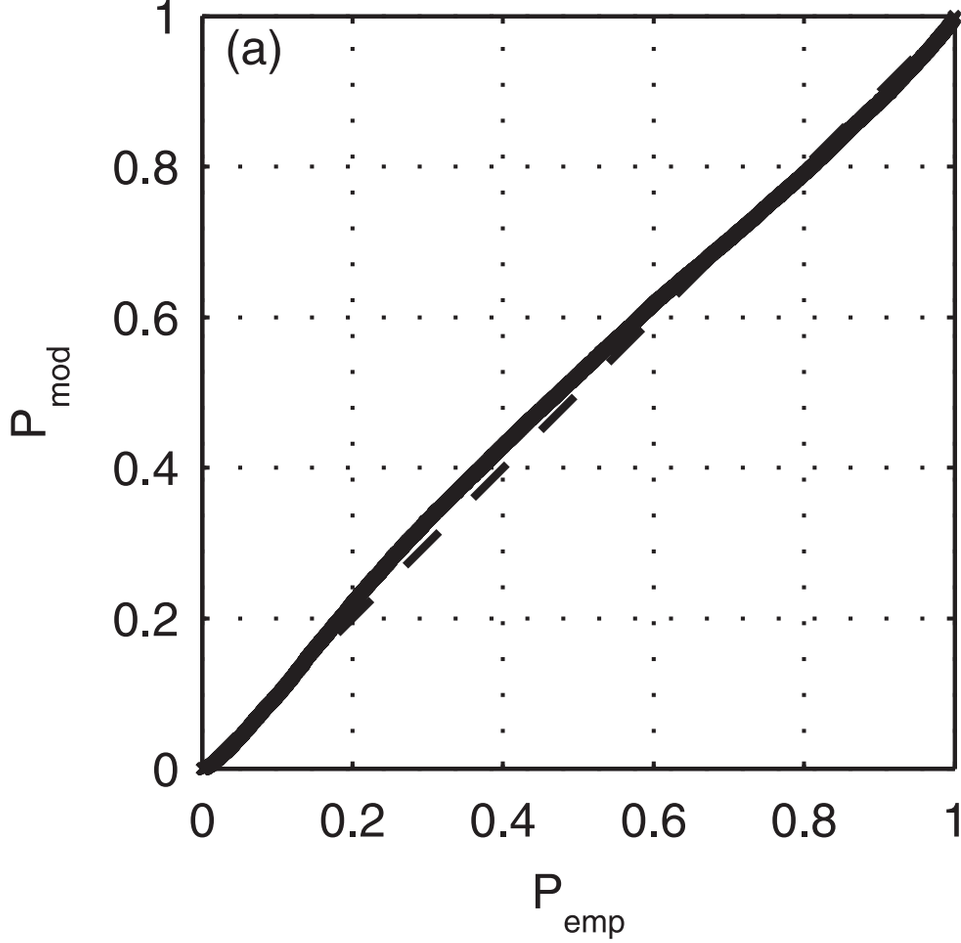

x 10^4
5.24
5.23
5.22
5.21
5.2
5.19
BIC
[4 2 2]
21 par
[6 5 1]
29 par
[8 5 3]
37 par
11
17
23
29
35
41
47
53
59
Parameters

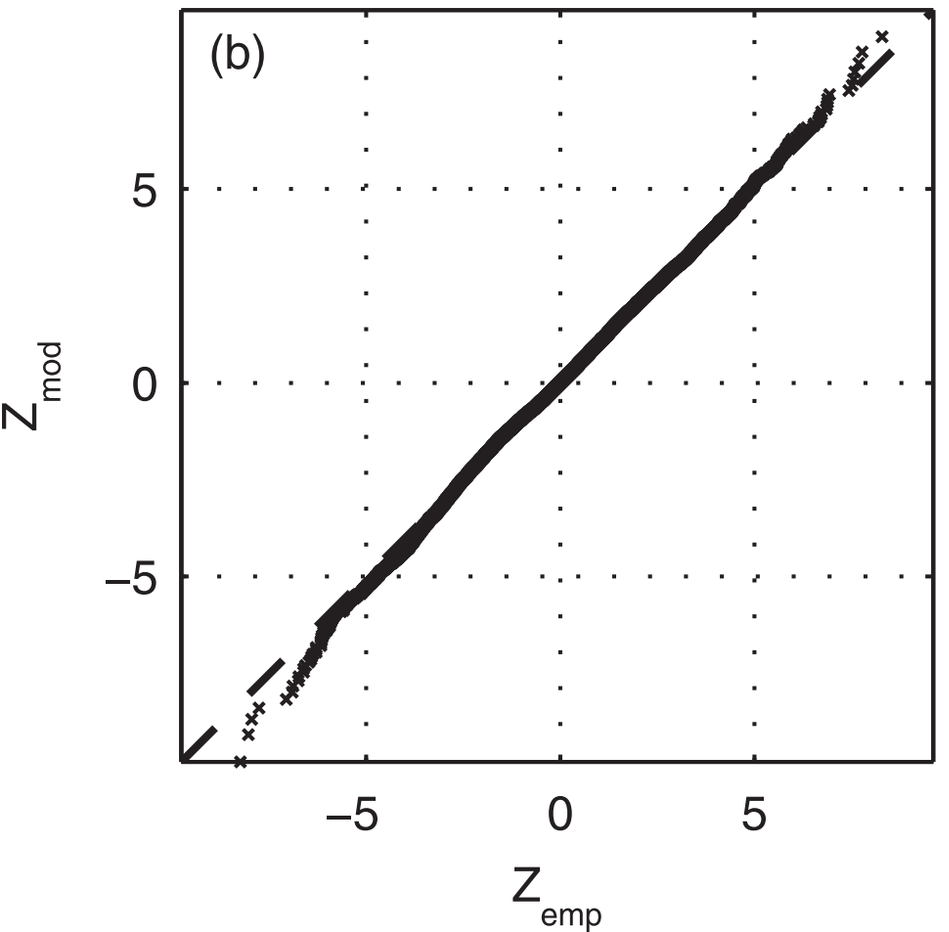

0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Time [years]
μLN
σLN
ξ2

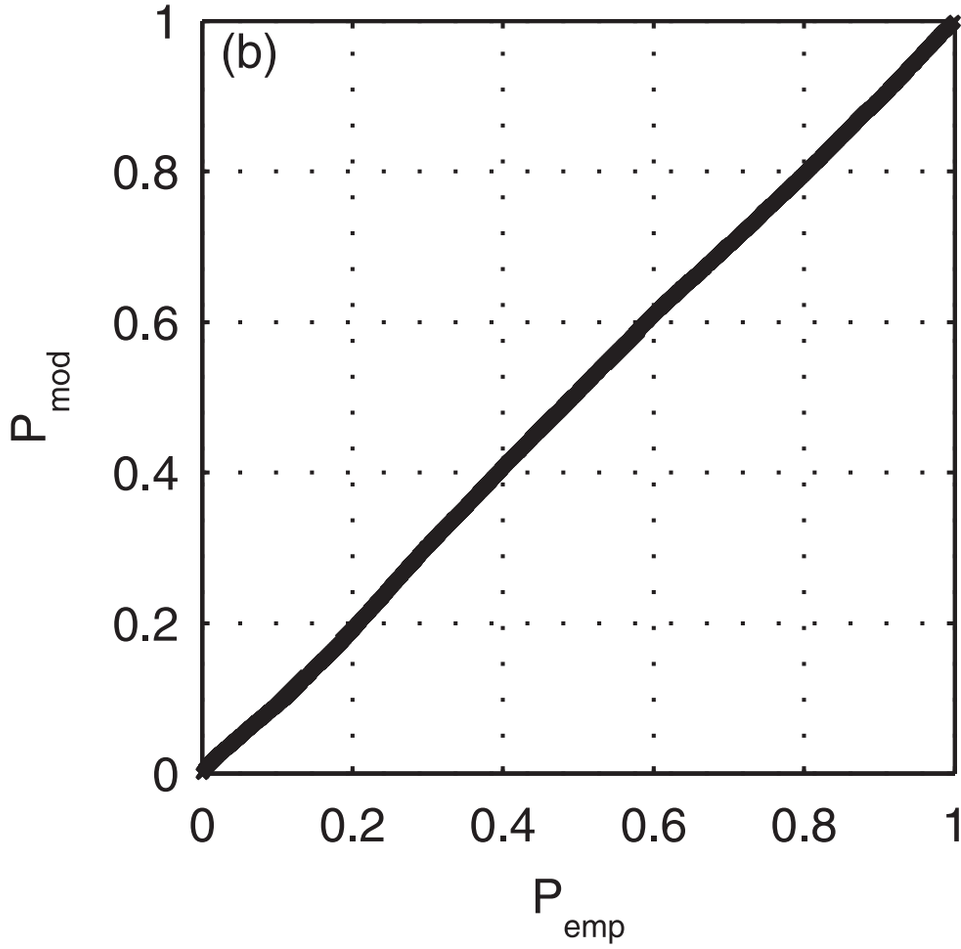

H_m0 [m]
10^1
10^0
10^−1
0
0.2
0.4
0.6
0.8
1
time [years]
P[xlt] = 0.99
P[xlt] = 0.90
P[xlt] = 0.75
P[xlt] = 0.50
P[xlt] = 0.25
P[xlt] = 0.10
P[xlt] = 0.01

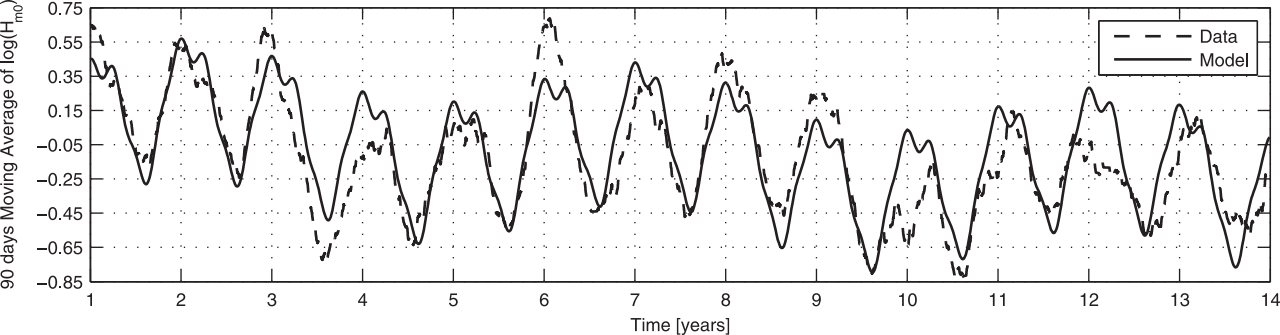

Data
NS–LN–GPD
NS–LN
CDF
.999
.99
.9
.75
.5
.25
.1
.01
.001
.1
.2
.4
.6
.8
1
2
4
6
8
H m0 [m]
PDF
1
0.8
0.6
0.4
0.2
0
0
1
2
3
4
5
H m0 [m]

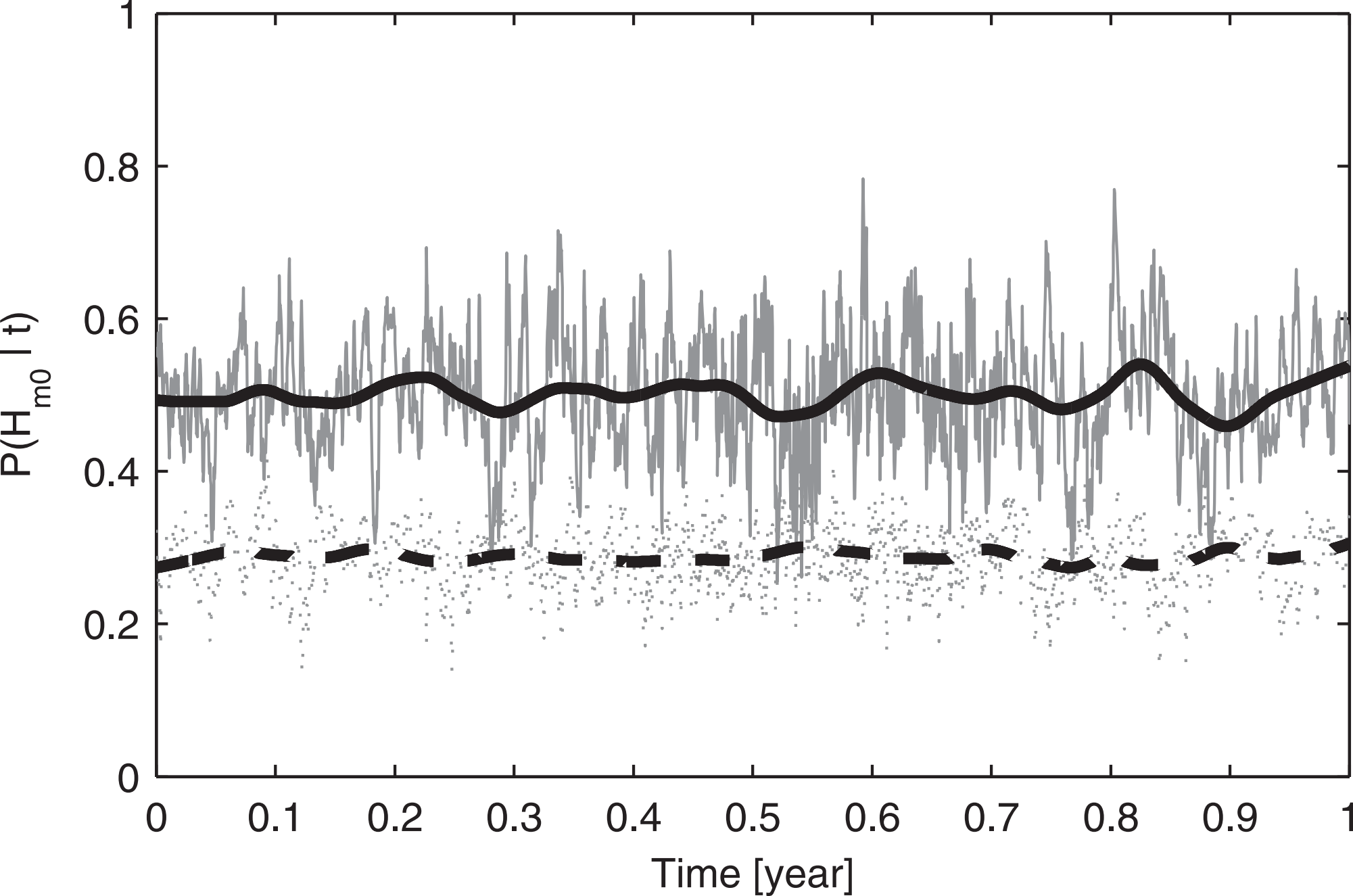

(a)
4
2
0
−2
−4
Z_mod
−4
−2
0
2
4
Z_emp

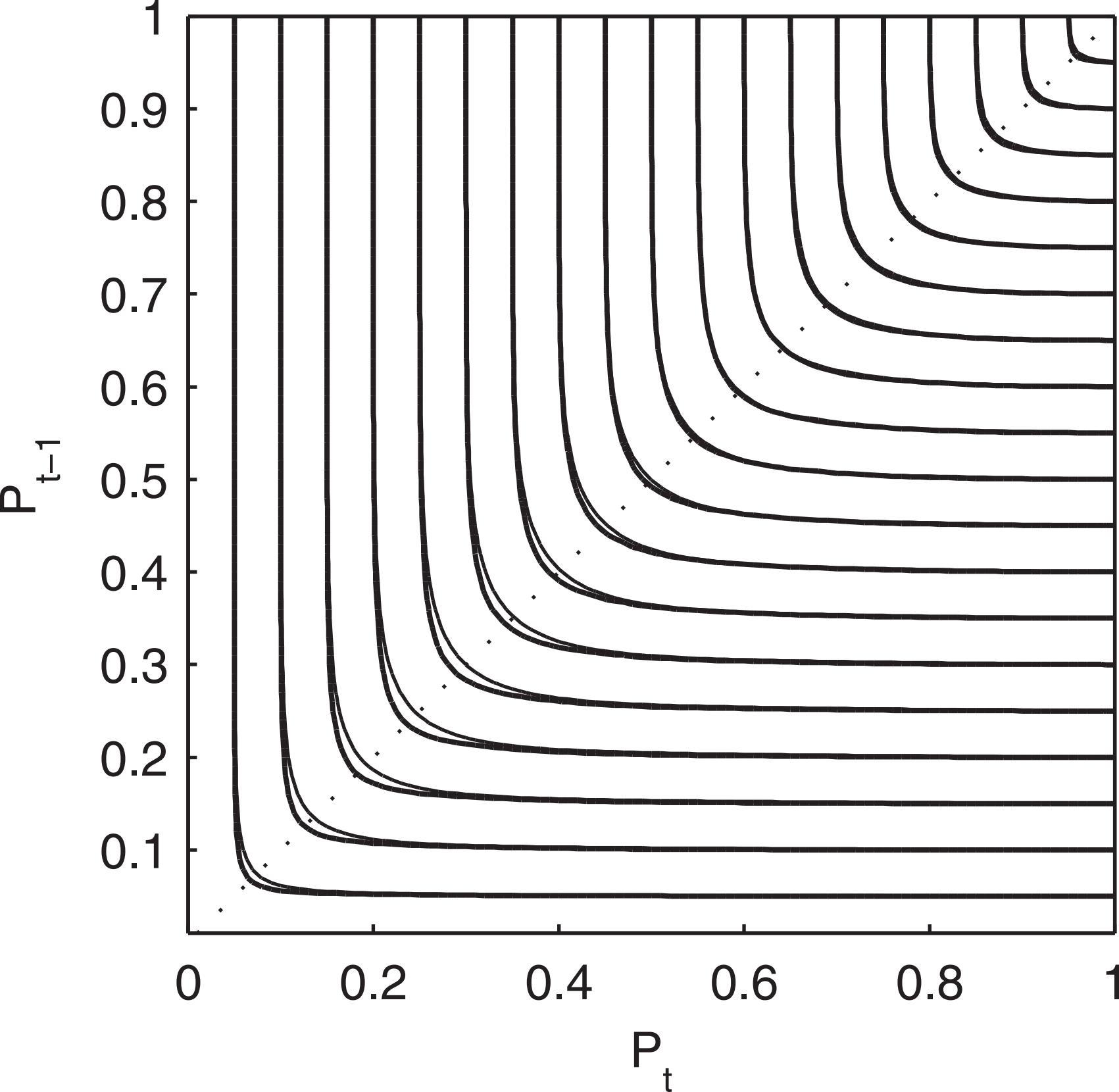

(a)
1
0.8
0.6
0.4
0.2
0
P_mod
0
0.2
0.4
0.6
0.8
1
P_emp

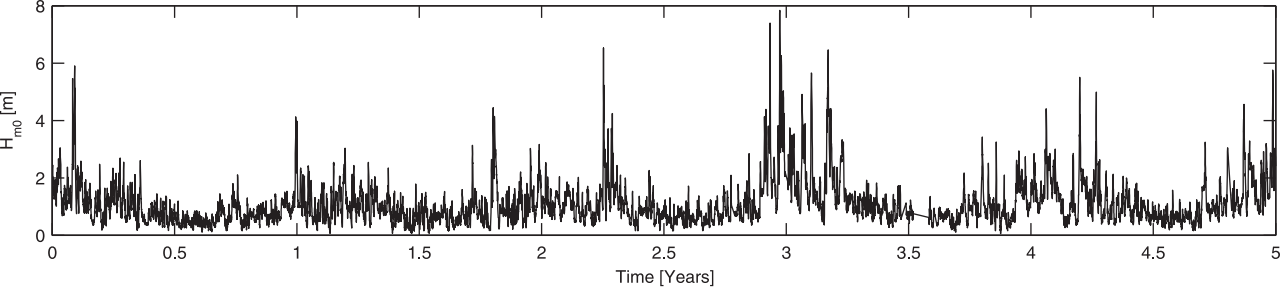

(b)
5
0
−5
Z_mod
−5
0
5
Z_emp

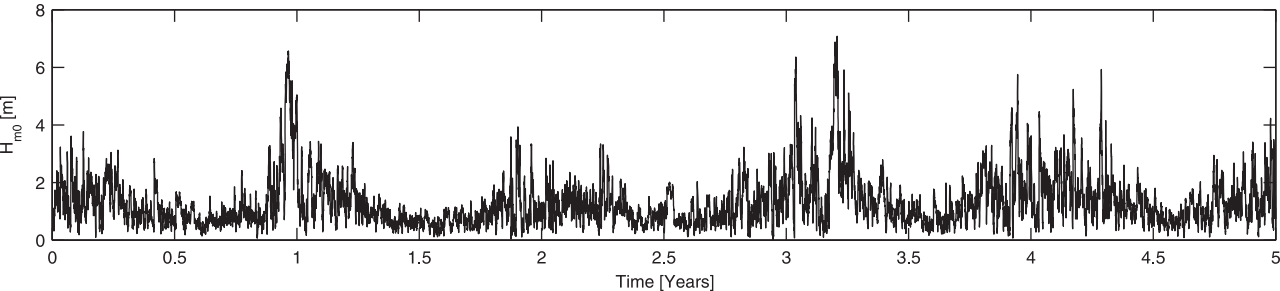

(b)
1
0.8
0.6
0.4
0.2
0
P_mod
0
0.2
0.4
0.6
0.8
1
P_emp

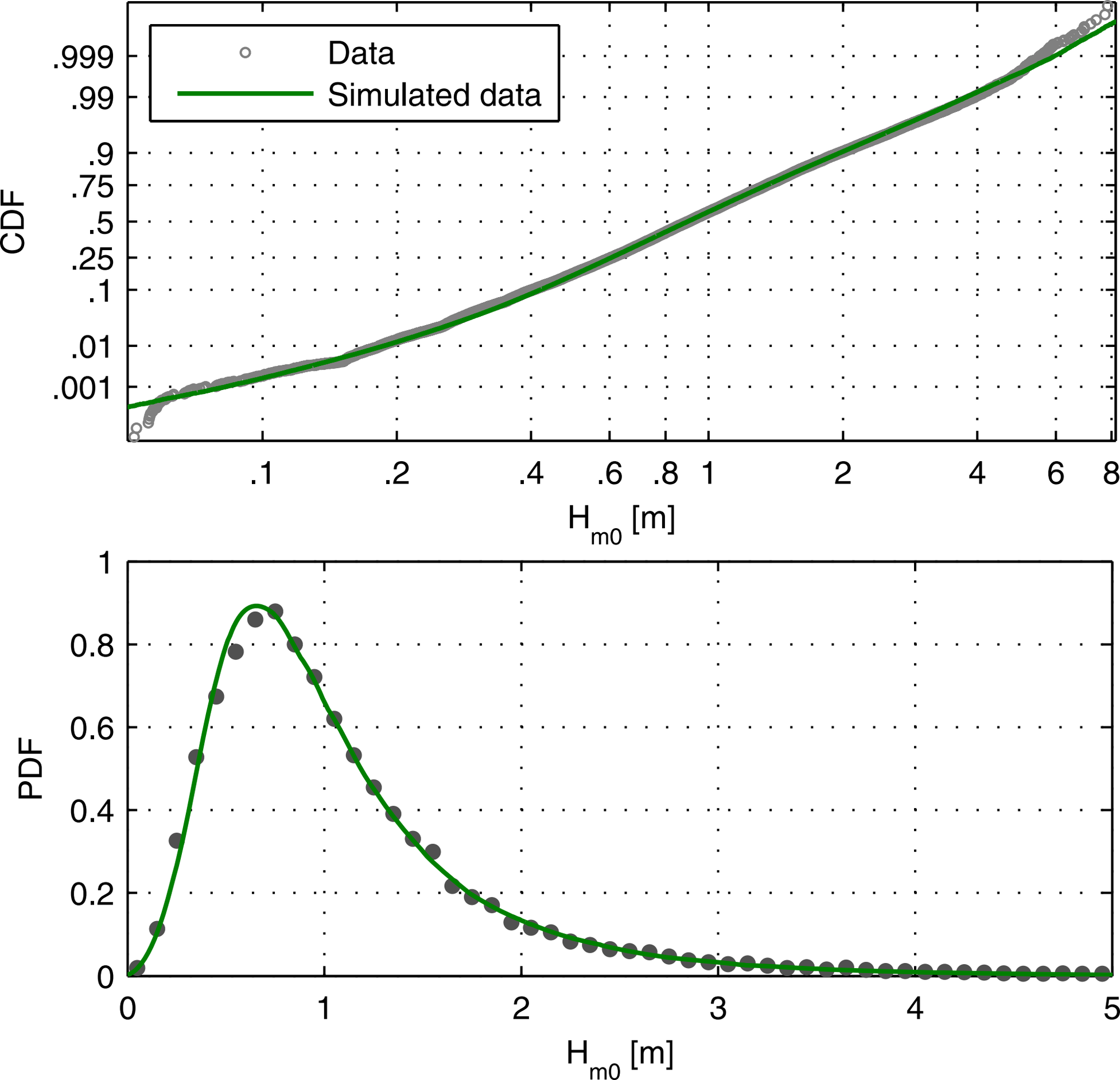

90 days Moving Average of log(H_m0)
0.75
0.55
0.35
0.15
−0.05
−0.25
−0.45
−0.65
−0.85
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Time [years]
Data
Model

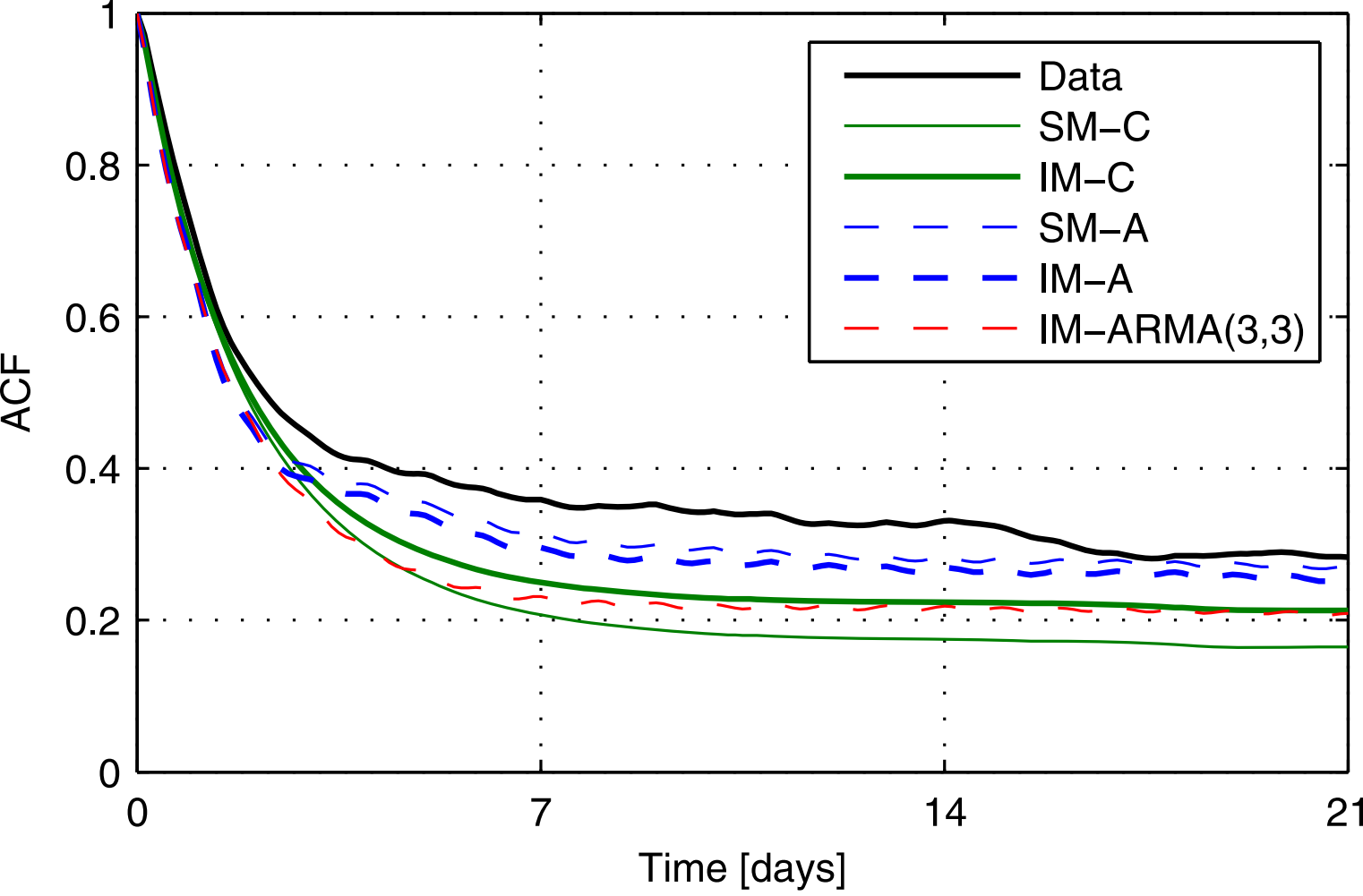

1
0.8
0.6
0.4
0.2
0
P(H_m0 | t)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Time [year]

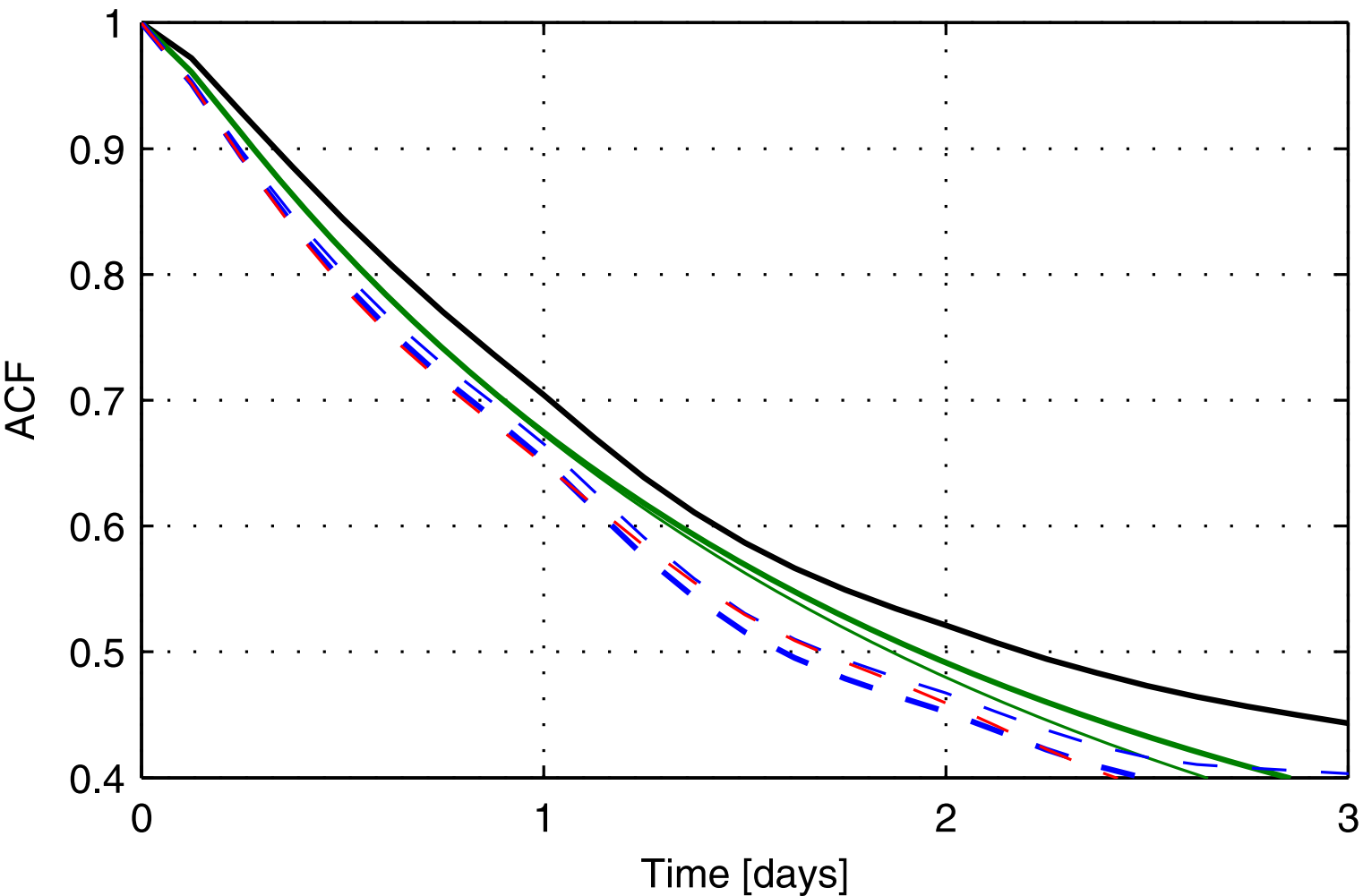

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
P_{t-1}
0
0.2
0.4
0.6
0.8
1
P_t

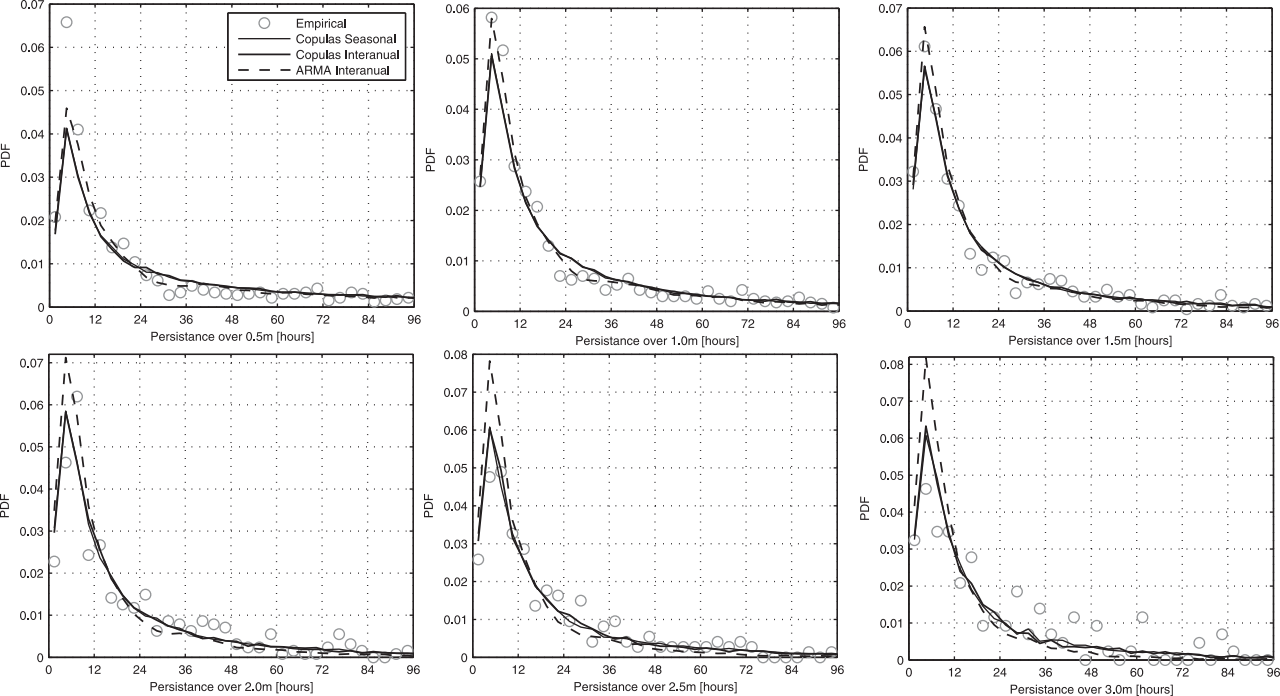

8
6
4
2
0
H_{m0} [m]
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Time [Years]

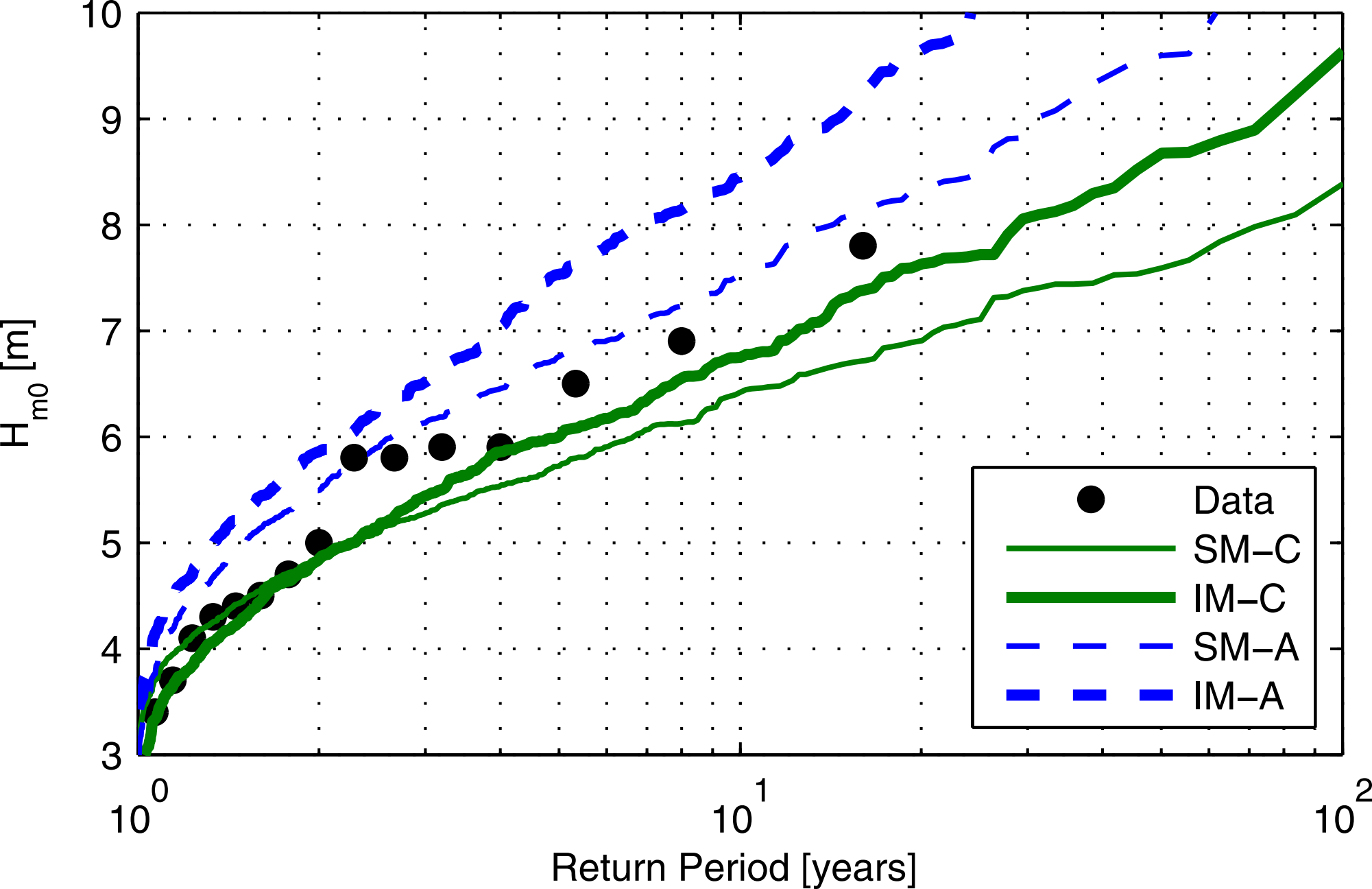

8
6
4
2
0
H_{m0} [m]
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Time [Years]

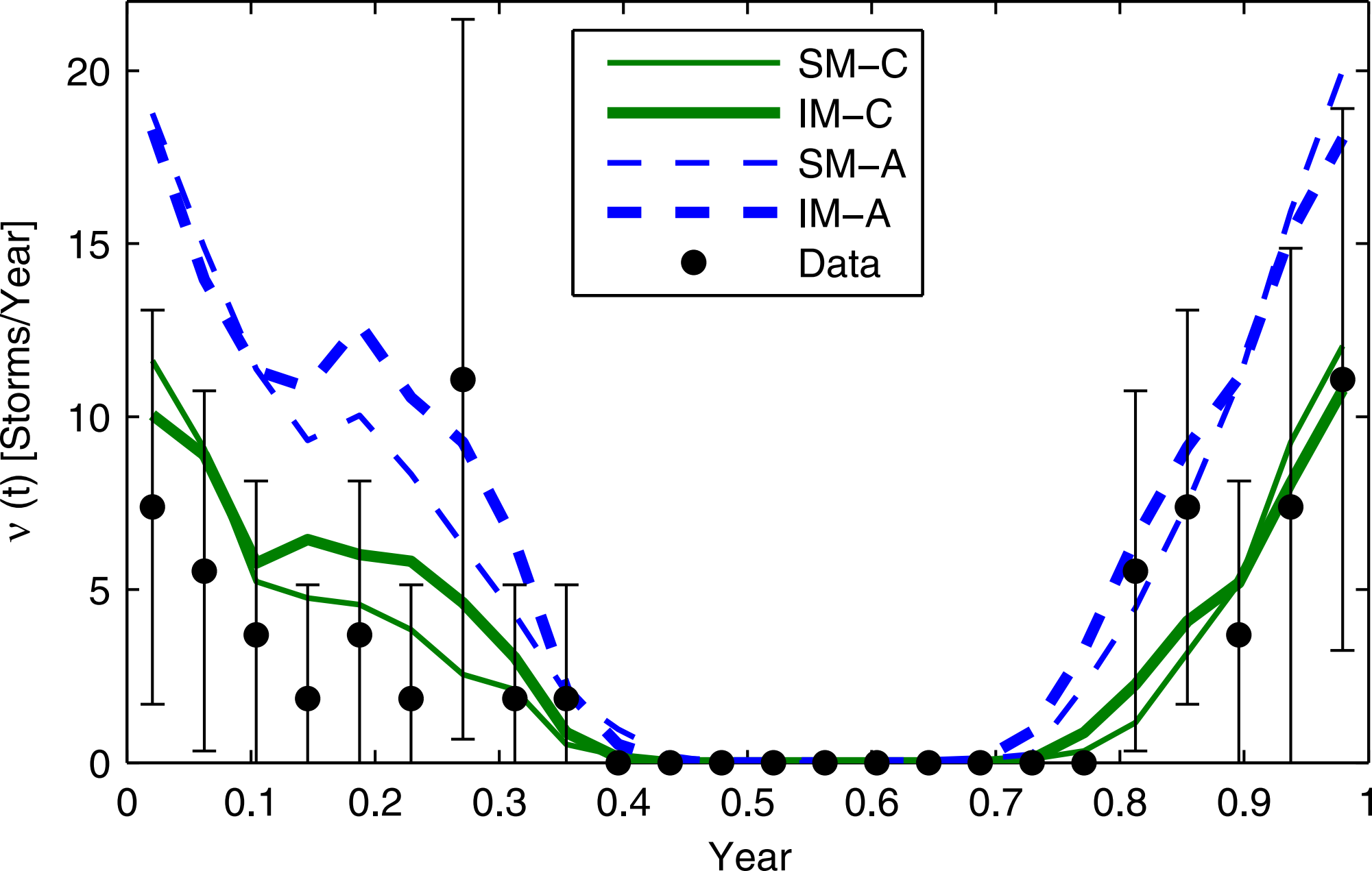

Data
Simulated data
CDF
.999
.99
.9
.75
.5
.25
.1
.01
.001
.1
.2
.4
.6
.8
1
2
4
6
8
H_{m0} [m]
PDF
1
0.8
0.6
0.4
0.2
0
0
1
2
3
4
5
H_{m0} [m]

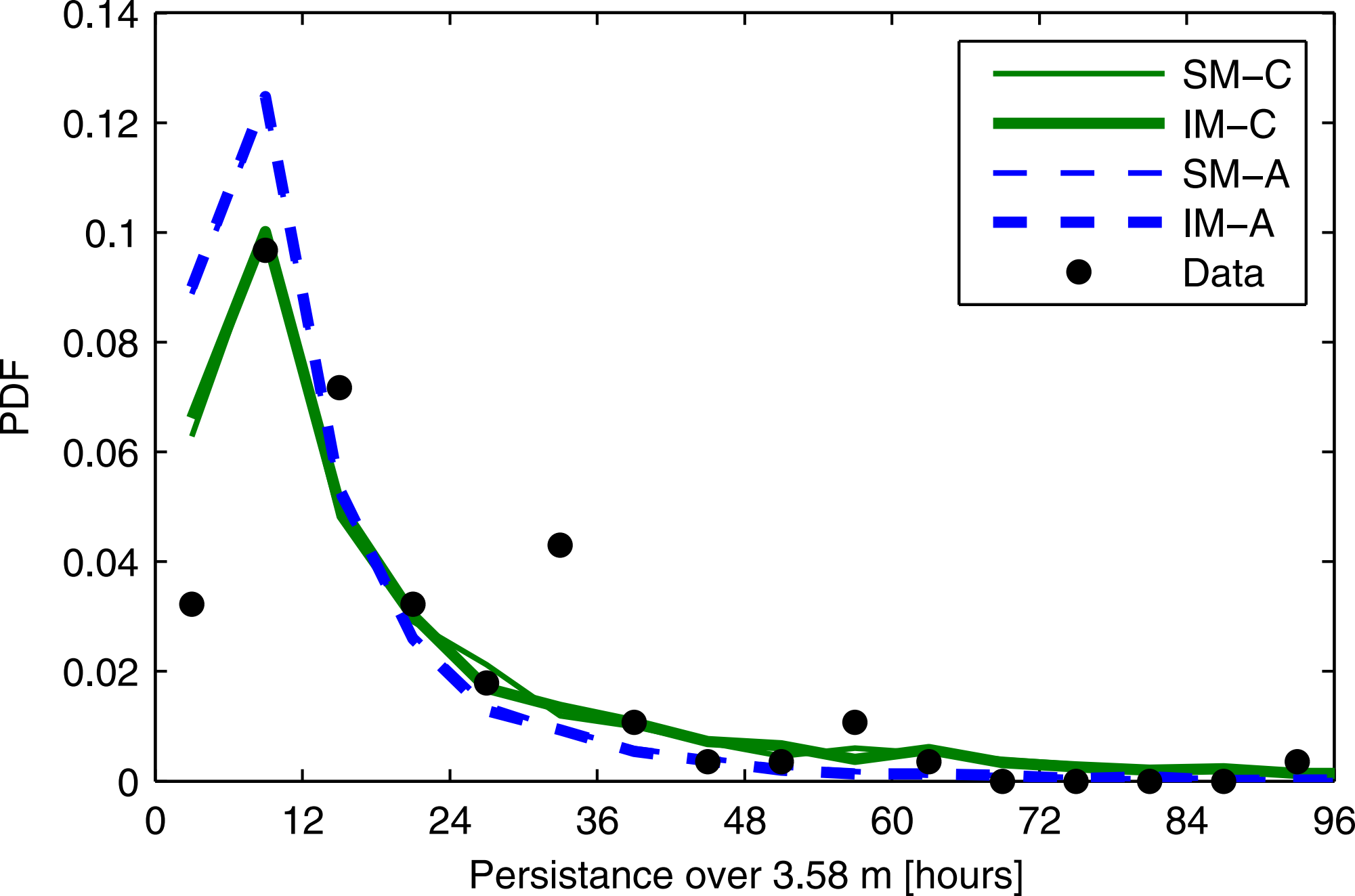

Data
SM–C
IM–C
SM–A
IM–A
IM–ARMA(3,3)
ACF
Time [days]
0
0.2
0.4
0.6
0.8
1
0
7
14
21

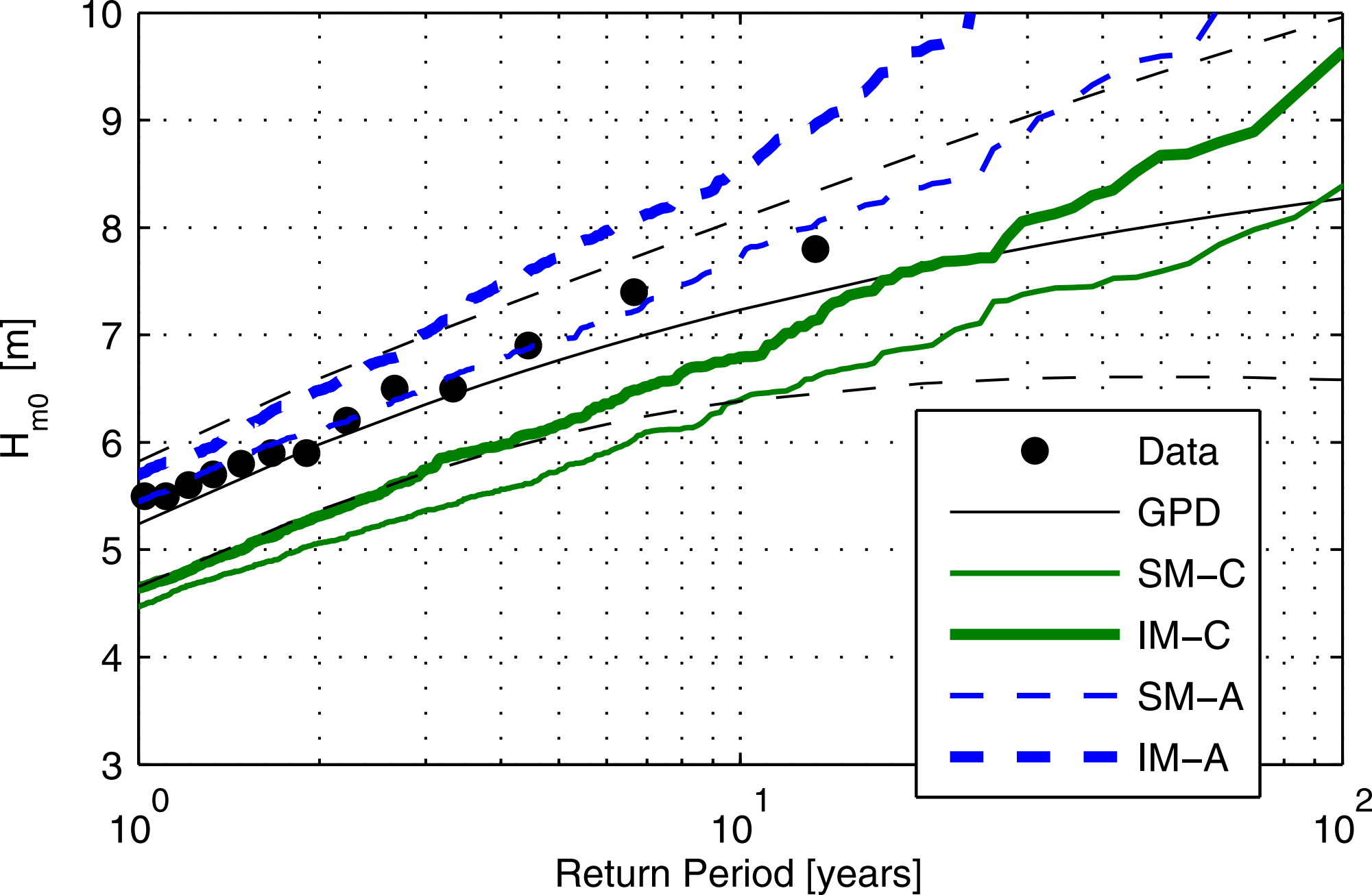

ACF
Time [days]
0.4
0.5
0.6
0.7
0.8
0.9
1
0
1
2
3


Empirical
Copulas Seasonal
Copulas Interanual
ARMA Interanual
PDF
Persistance over 0.5m [hours]
Persistance over 1.0m [hours]
Persistance over 1.5m [hours]
Persistance over 2.0m [hours]
Persistance over 2.5m [hours]
Persistance over 3.0m [hours]


10
9
8
7
6
5
4
3
H_{m0} [m]
10^0
10^1
10^2
Return Period [years]
Data
SM–C
IM–C
SM–A
IM–A


SM–C
IM–C
SM–A
IM–A
Data
ν (t) [Storms/Year]
Year
0
5
10
15
20
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


SM–C
IM–C
SM–A
IM–A
Data
PDF
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
12
24
36
48
60
72
84
96
Persistance over 3.58 m [hours]


10
9
8
7
6
5
4
3
H_{m0} [m]
10^0
10^1
10^2
Return Period [years]
Data
GPD
SM–C
IM–C
SM–A
IM–A