Marine Technology and Engineering – Guedes Soares et al. (Eds)
© 2011 Taylor & Francis Group, London, ISBN 978-0-415-69808-5

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

Sebastián Solari
Universidad de Granada, Grupo de Dinámica de Flujos Ambientales, Granada, Spain
Universidad de la República, IMFIA, Montevideo, Uruguay

Pieter H.A.J.M. van Gelder
Delft University of Technology, Faculty of Civil Engineering and Geosciences, Delft, The Netherlands

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.

1 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 (36.5ºN, 6.5ºW), 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 2 presents a brief revision of previous work on the use of autoregressive models for the simulation of multivariate met-ocean variables. In sections 3 to 5 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 6 to 8. Once parameters are estimated new time series are simulated. The comparison of the new series with the original ones is done at section 9. Finally, a discussion of the results is done at section 10, while main conclusions of this work are summarized at section 11.

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.

2 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.

3 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 $V_i(t) \sim F_i(V_i \mid t)$. Using this function, variables are normalized by means of

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

where $\Phi(x)$ is the standard normal univariate distribution. Non-stationary functions used in this work are introduced in section 4, the fitting of the functions to the data is presented on section 6, and the analysis of the normalized variables is shown in section 7.

(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 5, while the estimation of its parameters is shown in section 8.

(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 $V_i = F_i^{-1}(\Phi(Z_i) \mid t)$. The simulated time series obtained with the different models are analyzed on section 9.

4 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.

4.1 *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

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

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 (2) to be continuous and to have lower bound equal to zero, the following relations must be fulfilled (Solari and Losada 2011a; Solari and Losada 2011b)

$$\sigma_1 = -\xi_1 u_1 \quad \xi_1 = -\frac{F_c(u_1)}{u_1 f_c(u_1)} \quad \sigma^2 = \frac{1-F_c(u_2)}{f_c(u_2)}$$

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 (see Solari and Losada 2011c). 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 $(\mu_{LN}, \sigma_{LN}, \xi_2, Z_1, Z_2)$, while those of the WB-GPD distribution are $(\alpha_{LN}, \beta_{LN}, \xi_2, Z_1, Z_2)$. Parameters μ_{LN}, σ_{LN}, α_{WB}, β_{WB} and ξ_2 are time varying, and are expressed as Fourier series

$$\theta = \theta_{a0} + \sum_{k=1}^{N_k} \left(\theta_{ak} \cos(2\pi k t) + \theta_{bk} \sin(2\pi k t)\right) \qquad (3)$$

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).

4.2 *Bi log-normal model*

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

$$f(x) = \alpha f_{LN_1}(x) + (1-\alpha) f_{LN_2}(x) \qquad (4)$$

The parameters of the distribution are $(\mu_1, \sigma_1, \mu_2, \sigma_2, \alpha)$, with $\sigma_1, \sigma_2 > 0$ and $0 \leq \alpha \leq 1$. All parameters are time varying and are expressed as Fourier series using (3).

4.3 *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

$$f(x) = \sum_{i=1}^{4} \alpha_i f_{Ni}(x) \left[F_{Ni}(360) - F_{Ni}(0)\right]^{-1} \qquad (5)$$

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.

5 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 isused, and as a consequence these two models are piecewise linear.

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

5.1 *VAR(p) model*

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

$$y_t = \nu + \sum_{i=1}^{p} A_i y_{t-i} + u_t \qquad (6)$$

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

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

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

5.2 *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

$$y_t = \nu^{(j)} + \sum_{i=1}^{p} A_i^{(j)} y_{t-i} + u_t^{(j)} \quad \text{if} \quad r_{j-1} < z_{t-d} \leq r_j \tag{7}$$

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 repeatedwith 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.

5.3 *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

$$y_t = \nu^{(j)} + \sum_{i=1}^{p} A_i^{(j)} y_{t-i} + u_t^{(j)} \quad \text{if} \quad s_t = j \tag{8}$$

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

$$f(Y_n, s_1, \ldots, s_n, \lambda) = f(Y_n \mid S_n, \lambda) P(s_1) \prod_{t=2}^{n} P(s_t \mid s_{t-1}) \tag{9}$$

where

$$f(Y_n \mid S_n, \lambda) = f(Y_r \mid S_r, \lambda) \prod_{t=r+1}^{n} f(y_t \mid Y_{t-1}, s_t, \lambda) \tag{10}$$

with $f(Y_r \mid S_r, \lambda)$ being the likelihood of the first r observations and $f(y_t \mid Y_{t-1}, s_t, \lambda)$ 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 $P(s_t \mid s_{t-1})$ 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 (9) on all the possible realizations of s_t.

6 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 1 and 2 show the model obtained for H_{m0}. In figure 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 2, where agreement between both is noticeable.

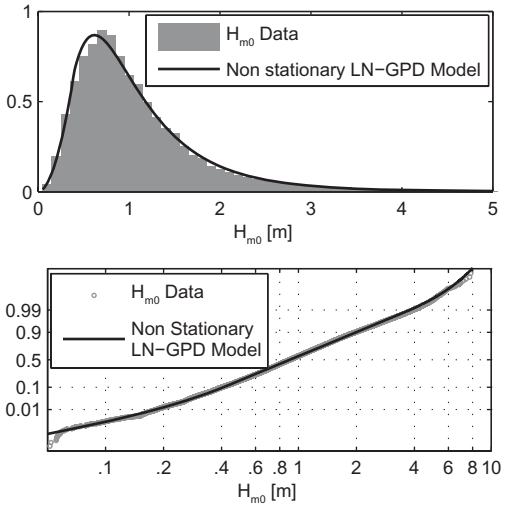

Figure 1. Annual probability density functions PDF (top) and cumulative distribution function CDF (bottom) for H_{m0}.

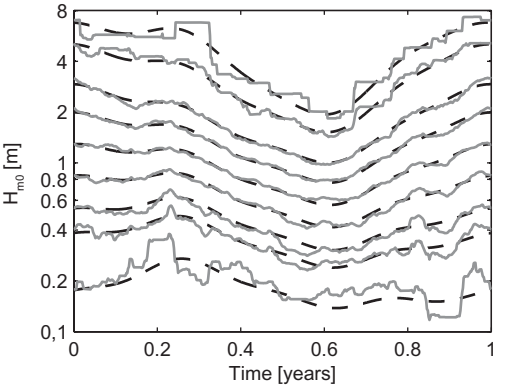

Figure 2. Empirical (gray) and modeled (black) quantiles of 1, 5, 10, 25, 50, 75, 90, 99 and 99.9% for H_{m0} as a function of time.

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

7 NORMALIZED DATA SERIES

By means of (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 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 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 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_{0M} 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 (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.

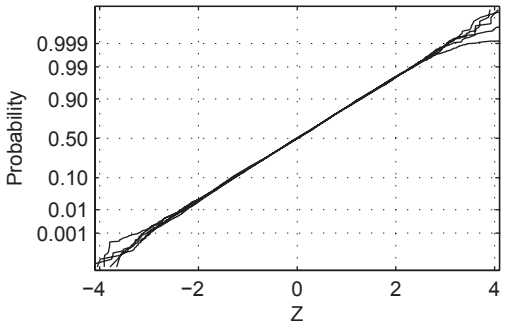

Figure 3. CDF of the normalized variables in normal probability paper.

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.

8 AUTOREGRESSIVE MODELS PARAMETERS ESTIMATION

8.1 *VAR model*

The parameters of the VAR(p) model are estimated through the least square method described on section 5, for order p between 1 and 8. For each model the BIC is calculated, and the lower BIC is obtained with $p = 7$.

8.2 *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

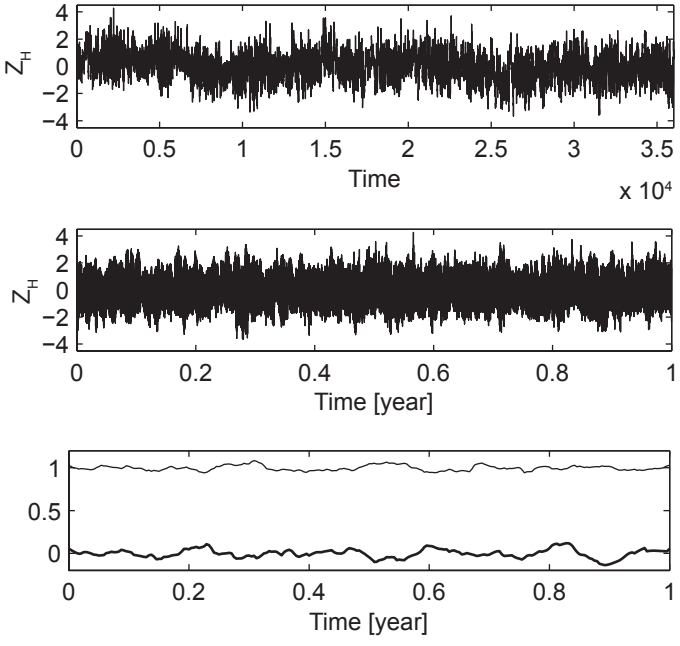

Figure 4. Time series of the normalized variable Z_H (top), normalized variable on annual scale (middle), and 90 days moving average of the mean and the standard deviation (bottom).

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 $(r_1, r_2) = (3.8\ m/s, 7.1\ m/s)$.

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.

8.3 *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 (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 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 lowerthan 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 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

8.4 *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 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).

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

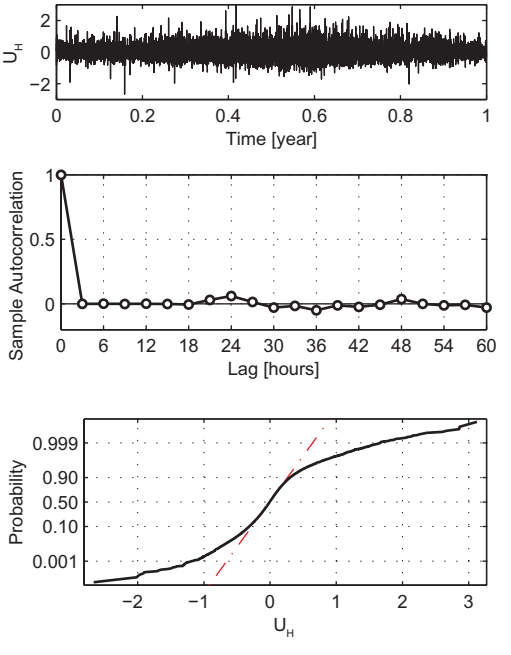

Figure 5. Error $\hat{U}_H$ obtained with VAR(7) model (top), autocorrelation of $\hat{U}_H$ (middle), CDF of $\hat{U}_H$ in normal probability paper (bottom).

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.

9 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.

9.1 *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 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.

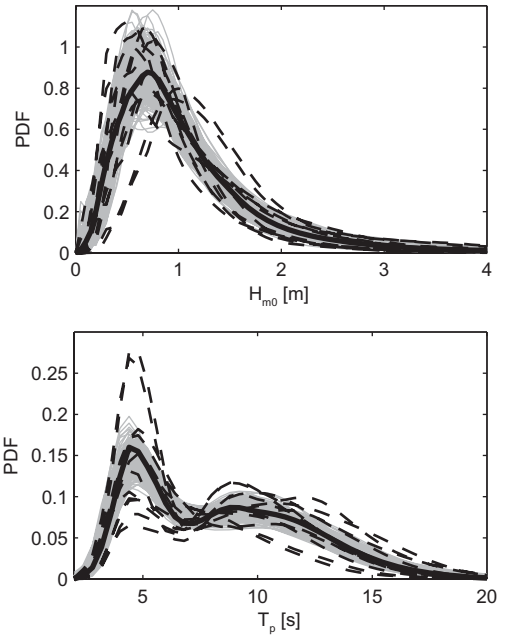

Figure 6. Interanual variability of the anual PDF for H_{m0} (top) and T_p (bottom). Grey: annual PDF for each simulated year; Broken Black: anual PDF for each measured year; Continuous Black: mean anual measured PDF.

9.2 *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 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 7 and 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 dis-

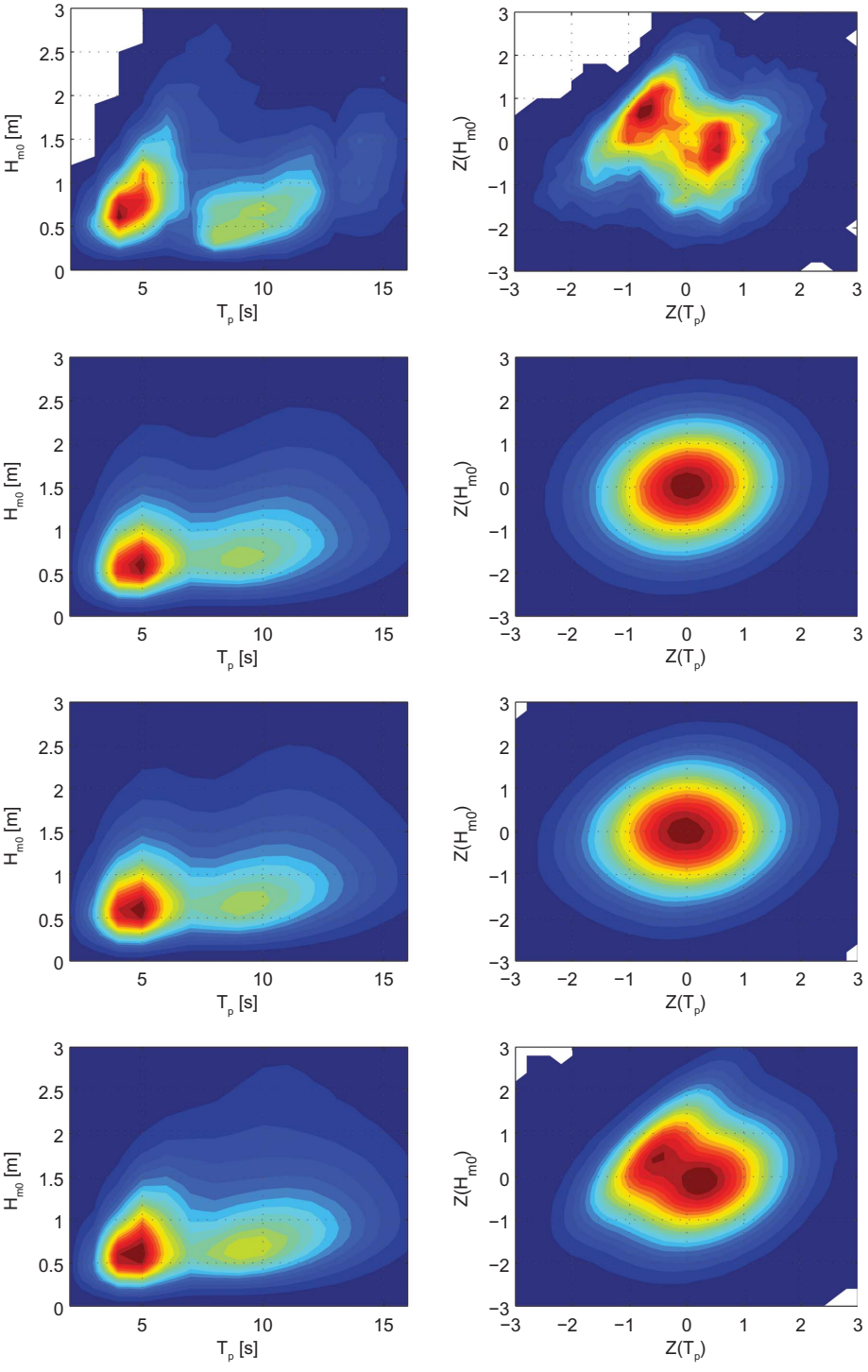

Figure 7. Bivariate distribution of $H_{m0} - T_p$ (left) and of $Z_H - Z_T$ (right), obtained with the measured data (top) and with the data simulated with the VAR (second from top), the TVAR (third from top) and the MSVAR (bottom) models.

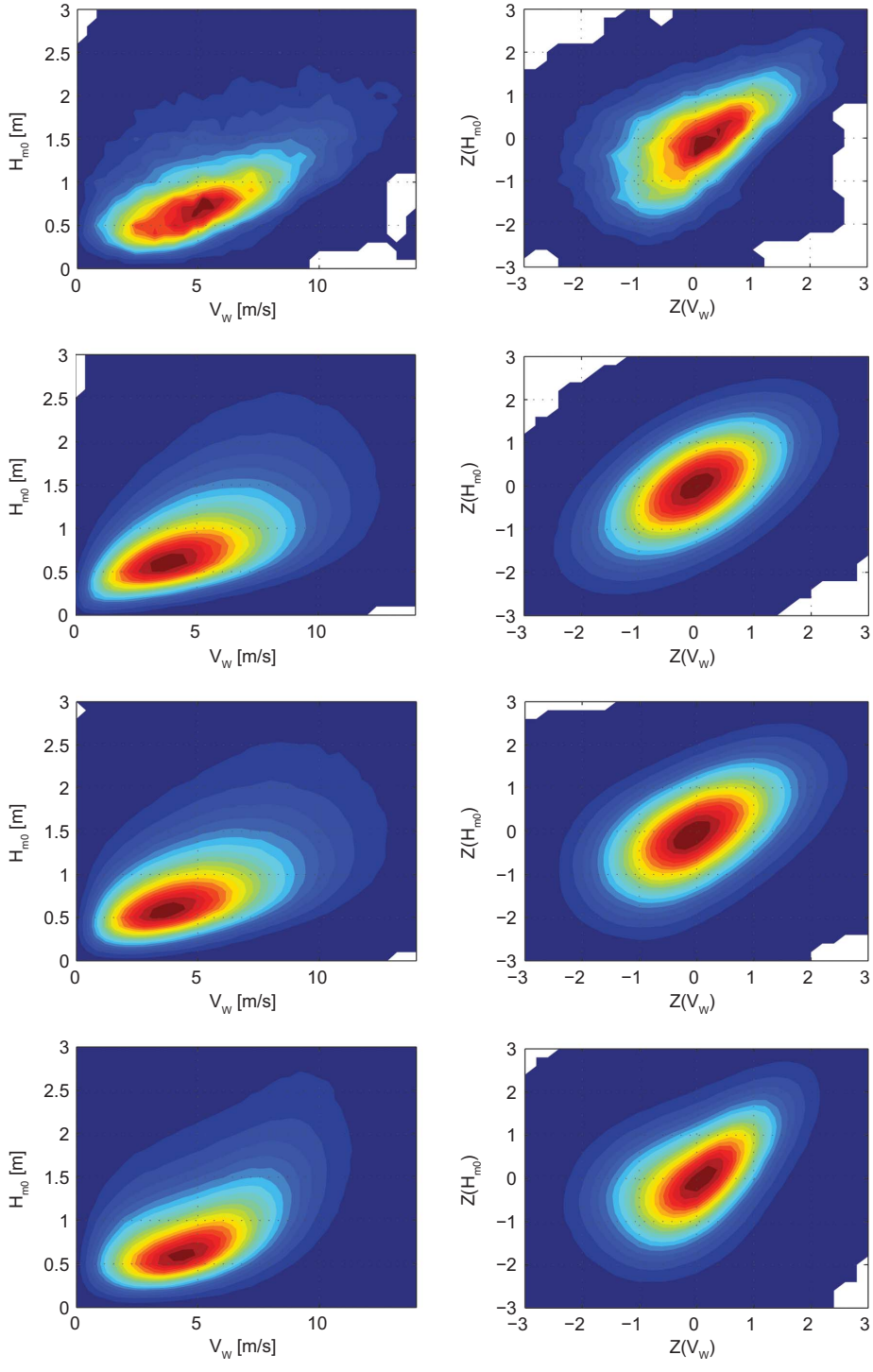

Figure 8. Bivariate distribution of $H_{m0} - V_W$ (left) and of $Z_H - Z_V$ (right), obtained with the measured data (top) and with the data simulated with the VAR (second from top), the TVAR (third from top) and the MSVAR (bottom) models.

tribution 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).

9.3 *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 9 shows persistence regimes of H_{m0} and V_W over thresholds corresponding to nonexceedance probabilities 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.

9.4 *Auto- and cross-correlation*

Figure 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

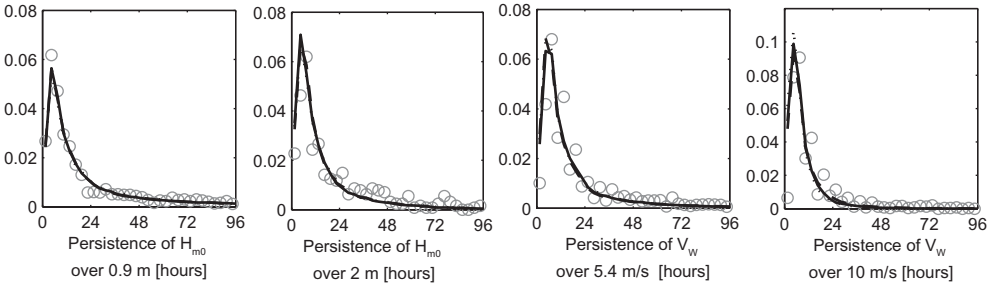

Figure 9. Persistence of H_{m0} and V_W over thresholds corresponding to mean annual probability 0.5 and 0.9.

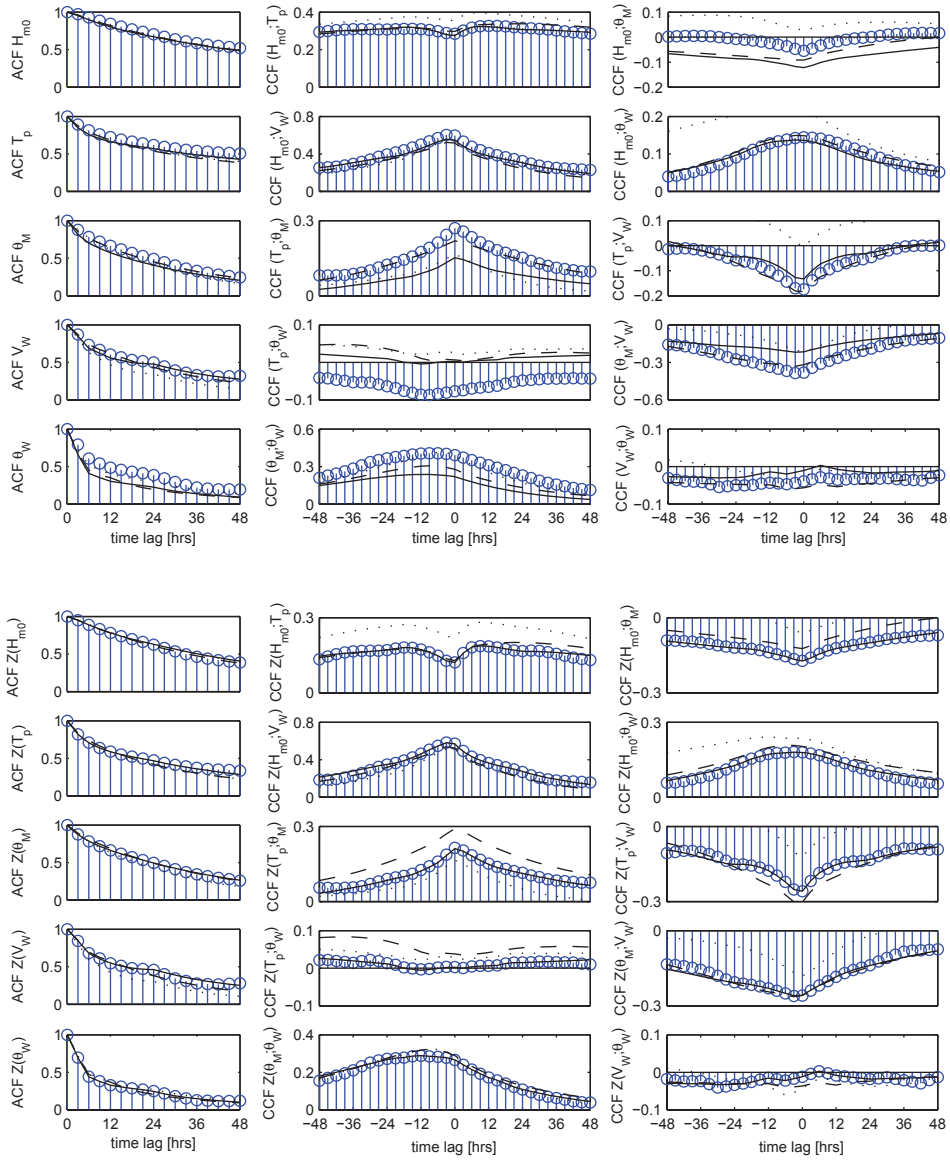

Figure 10. Auto- and cross correlation functions of the original (top) and normalized (bottom) variables. Lags expressed in hours. Blue circles: measured data; continuos line: VAR model simulated series; dashed line: TVAR model simulated series; dotted line: MSVAR model simulated series.

measured data shows negative cross-correlation (with higher winds sea condition is observed, while with lower winds swell prevails).

10 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 normalization and stationarization of the data series. However, the proposed method has two limitations. First, it does not account for inter-annual 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 (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 cross-correlation 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.

11 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.

ACKNOWLEDGEMENTS

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 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.

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.

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). Identification of minimum, mean and maximum wave regimes. part i: Model description. *To be submitted to Coastal Engineering*.

Solari, S. and M.A. Losada (2011b). Identification of minimum, mean and maximum wave regimes. part ii: Application. *To be submitted to Coastal Engineering*.

Solari, S. and M.A. Losada (2011c). Non-stationary wave height climatie modeling and simulation. *To be submitted to Coastal Engineering*.

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.