Probabilistic Engineering Mechanics
Volume 22, Issue 2, April 2007, Pages 113-126
doi:10.1016/j.probengmech.2006.08.003

Survey of stochastic models for wind and sea state time series

V. Monbet[a, b], , P. Ailliot[a], and M. Prevosto[b]

[a] Department of Applied Statistics, University of South Brittany, BP 573, 56017 Vannes cedex, France

[b] Hydrodynamics and Metocean, IFREMER/DCB/ERT/HO, BP 70, 29280 Plouzané, France

Abstract:

The knowledge of sea state and wind conditions is of central importance for many offshore and nearshore operations. In this paper, we make a complete survey of stochastic models for sea state and wind time series. We begin with methods based on Gaussian processes, then non-parametric resampling methods for time series are introduced followed by various parametric models. We also propose an original statistical method, based on Monte Carlo goodness-of-fit tests, for model validation and comparison and this method is illustrated on an example of multivariate sea state time series.

Keywords: Sea state; Wind; Nonlinear time series; Simulation; Model validation

1 Introduction

The knowledge of sea state and wind conditions is of central importance for many offshore and nearshore operations. For instance, wind time series permit to evaluate the power values produced by wind turbine, or to investigate load matching and storage requirements (Brown et al., 1984), (Castino et al., 1998). The evolution of sea state and wind conditions is also determinant in coastal erosion (Waeles et al., 2004). And several questions concerning safety,

reliability and feasibility of offshore activities (O'Caroll, 1984), (Monbet *et al.*, 2001a) as well as maritime transport (Baxevani *et al.*, 2004), (Ailliot *et al.*, 2003) and the drift of floating objects (Ailliot *et al.*, 2006b)) or oil spills are directly related to wave and wind conditions.

Even if there is a huge amount of data collected on physical quantities related to wind and waves, they remain sparse relatively to the size of oceans (simply one can not observe everywhere at the same time). Hence estimates of risks for undesired scenarios for ocean operations are usually computed by means of stochastic models. Often scenarios are defined as excursions above critical values of some responses of complicated systems and Monte Carlo can be the only way to derive probabilities of interest.

In this paper, we make a survey of stochastic models for wind and sea state time series and we focus mainly on simulation. We have chosen to consider only time series at the scale of the sea state (i.e. with time step from 1 to 6 hours) and at a given geographical point. As a consequence, events at the scale of waves are not modeled and no spatial information is taken into account. Several authors have proposed spatio-temporal models, see for instance (Vanhoff *et al*, 1997), (Baxevani *et al.*, 2004), (Ailliot *et al.*, 2006a) and references therein.

In section 2, after a short description of the various sources of data, we discuss what a good model should be and then we briefly present various methods to deal with the non-stationarities (interannual variability, seasonal and daily components,...). Then the most famous tools for time series modeling, i.e. the Box and Jenkins method (Box *et al.*, 1976) and some extensions of it are introduced in section 3. Non parametric resampling methods are described in section 4. These methods was rarely proposed in the literature for sea state processes but we think that they could be of high interest for many applications. In section 5, various parametric models are briefly presented. Then, in section 6, a validation method is proposed to measure the ability of a model to simulate realistic artificial sequences. Finally, in section 7 some of these methods are illustrated on a multivariate time series describing sea state conditions along a ferry line in Aegean Sea.

2 Generalities

2.1 Data

The data which describe wind and sea state conditions can be gathered in two categories: gt;

In situ observations obtained from ships, buoys, satellites, ...

Outputs of meteorological models, i.e. hindcast, nowcast or forecast data.

Some authors have compared the quality of these different kind of data, see for instance (Woolf *et al.*, 2002), (Caires *et al.*, 2004), (Izquierdo *et al.*, 2005) and references therein. The main drawbacks of buoys data is that they often include missing and noisy data and the longer buoy time series are typically only about 10 years long. Satellite data are recorded with very small time steps along the track but it generates sparse data in time for a given location. Generally, outputs of numerical models are easier to use because these data are available on long period (up to 50 years) and there is no missing data. But their quality is not always good, in particular, they are known to be smoother as in situ data and also to underestimate extreme events.

Let us now introduce some notations for the synthetic parameters which are used throughout this paper. For more precise definitions, see (IAHR, 1989). gt;

H_s: significant wave height (m). For H_s the most common definitions are the average of the highest one-third of wave heights or 4 times the standard deviation of the sea surface elevation process. These definitions are equivalent for seas with narrow band spectra. There is a third definition of H_s, namely $H_s = 4\sqrt{m_0}$, where m_0 is the zeroth-order moment of the sea state spectrum.

T: wave period (s). The most often used definitions are the spectral peak wave period, which is the inverse of the frequency at which the spectral density function of the elevation time series is maximum, and the zero crossing period which is the average time between successive zero downcrossing waves.

Θ_m: mean direction to which waves are traveling (degree).

U: wind intensity (ms^{-1}). It is the mean of the speed of the air particles at 10 meters over a fixed time period (20 minutes in general).

Φ: wind direction (degree). It is the mean direction from which the wind is blowing at 10 meters over a fixed time period (20 minutes in general).

In practice, for each sea state, the sea state parameters H_s, T and Θ_m can be deduced from the wave directional spectra, and this spectra may exhibit several peaks. It is the case for instance when several systems coexist, such as wind sea generated by local winds and swell radiated from distant storms. For some applications, it can be useful to separate the wave energy into different components (see Wang *et al.*, 2001), but such separation is not considered in this paper.

2.2 What is a good model?

First of all, the response to this question depends on what the model is built for. The contexts of use considered in this paper are: explanation of a physical phenomena, time series simulation, forecasting, time series reconstruction. However, a particular attention is paid to simulation and reconstruction. The word *reconstruction* can refer to two different problems: *missing data reconstruction* which uses the observed time series itself for missing values completion (Stefanakos *et al.*, 2001) and *cross-reconstruction* where a time series is reconstructed given another one. For instance a wave time series can be approximately reconstructed given a wind time series (Ailliot *et al.*, 2003) or wave time series at other locations (Arena *et al.*, 2004)

Once the context of use is specified, the choice of the model relies on a compromise between its ease of use and its accuracy in describing various features of the physical phenomenon.

The **ease of use** can be measured qualitatively according to several criteria: gt;

model robustness to the data source (ship, buoy, satellite, meteorological model). As mentioned above, there may exist missing or aberrant data, the data may not be recorded at a constant time step or only descriptive statistics may be available, such as marginal distributions or persistence durations, for instance (Vik, 1981), (Hogben, 1987).

model robustness to the nature of the process. Each sea state parameter has specific characteristics, its evolution may also depend on the geographical area and on the season. Furthermore, depending on the considered application, the process may be univariate or multivariate, and in the second case a strong dependence may exist between the components as for H_s and T or H_s and U. Finally, the state space of the considered parameters can be either finite, positive or the torus $\mathbb{R}/2\pi\mathbb{Z}$ for the circular processes Θ_m and Φ.

mathematical properties: amount of data needed to accurately estimate the model, asymptotic properties of the estimators...

necessary time for implementation of the algorithms, and running estimation, simulation or reconstruction.

The **accuracy of a model** can be evaluated by comparing statistics of the observed time series with those computed from artificial realizations of the model. Generally, only graphical comparison are performed, and it does not permit to measure the goodness of fit. We propose in section 6 a formal method, based on Monte Carlo tests, to compare and validate the models. It quantifies, in

particular, the ability to restore chosen features of the observed time series such as the marginal cumulative distribution functions (cdf), the distribution of annual extremes, the distribution of storms duration, the autocovariance functions, etc.

2.3 Modeling non-stationarity

Generally, several types of non-stationary components can be identified in meteorological time series. In particular, there may exist interannual components induced by natural cycles (ENSO, IPO, NAO...) or human activities, seasonal components and also daily components.

Athanassoulis and Stefanakos (1995) identified year-to-year variability in sea state time series by comparing mean annual values. Different methods have been proposed to describe trends in time series (see Brockwell *et al.*(1991)). For example, they can be approximated by a polynomial function and then eliminated in order to obtain a trend-free time series. However, such approach is difficult to implement here because of the few amount of data with respect to the temporal scale of the events.

Seasonal components are generally easy to observe on meteorological time series and several methods have been proposed to describe these components (see Cunha *et al.*, 1999, for a recent review). Two methods are commonly used: gt;

Let $\{Y_t\}$ denote the process under consideration. It can be assumed, in a first approximation, that the following decomposition holds (Walton *et al.* (1990), Medina *et al.* (1991), Stefanakos *et al.* (2005)):

$$Y_t = m(t) + \sigma(t) Y_t^{stat} \tag{1}$$

where m and σ are deterministic periodic functions with period one year which represent respectively the seasonal mean and standard deviation of the process. And $\{Y_t^{stat}\}$ is assumed to be a stationary process. Methods for estimating the deterministic components $m(t)$ and $\sigma(t)$ and for checking the stationarity of Y_t^{stat} have been discussed by Athanassoulis *et al.* (1995). When $\{Y_t\}$ is an univariate process, m and σ can be approximated by a low-order trigonometric polynomial. An other common approach consists in computing the estimates of the monthly mean and the monthly standard deviation as seasonal pattern. The periodic functions m and σ are then deduced by repeating the same estimated pattern over successive years (Stefanakos *et al.*, 2005), (Monbet *et al.*, 2001a). This last method is easily generalized to multivariate

processes. In this case, σ_t is a matrix which must describe interactions between the different components.

In (Boukhanovski *et al*, 1999) and (Stefanakos *et al.*, 2002), the assumption that $m(t)$ and $\sigma(t)$ are deterministic is relaxed in order to introduce random variation, and it is assumed that

$$Y_t = m(t) + \sigma(t)(Y_t^{stat} + \epsilon_t) \tag{2}$$

where ϵ_t is a white noise process evolving at a monthly time scale.

An other approach consists in supposing that the process is piecewise stationary and to fit separate models for each month or for each season of the year (Brown *et al.*, 1984), (Borgman *et al.*, 1991).

In the first method, it is assumed that the standardized process Y_t^{stat} is stationary, and it may be a too strong assumption in some cases. However, its main advantage is that the estimation of m and σ does not require a lot of data, typically 3 or 4 years of data provide accurate estimates. On the contrary, in order to apply the second method, more data i generally required since a different model is fitted each month. Furthermore, artificial ruptures of the model are induced between successive months. But a great advantage of this method over the first one is that the stationarity assumptions seem less restrictive.

Some wind and sea state time series also exhibit daily components. The most common method to remove this components is to use the decomposition (1), with m and σ periodic functions with period one day (see (Brown *et al.*, 1984), (Daniel et al., 1991)).

In the sequel, we suppose that all the studied processes are stationary.

3 Models based on Gaussian processes

In general, wind and sea state time series cannot be assumed to be Gaussian. For instance, the marginal distribution of these processes are often asymmetric with positive support and positive skewness. However, when they have a continuous state space, it is possible to transform these time series into time series with Gaussian marginal distributions. If the transformed time series is supposed to be Gaussian, we can then use one of the existing techniques to simulate Gaussian processes (ARMA models, exact simulation methods,...). Let us describe more precisely the simulation method in a general framework.

Let $\{Y_t\}$ be a stationary process with values in $\mathbf{R}^d$. We assume that there exists a transformation $f : \mathbf{R}^d \to \mathbf{R}^d$ and a stationary Gaussian process $\{X_t\}$ such that $Y_t = f(X_t)$. The procedure consists in three main steps: gt;

Model calibration, which consists in determining the function f and the second order structure of the process $\{X_t\}$. In practice, f is chosen such that the marginal distributions and the second order structures of the processes $\{f(X_t)\}$ and $\{Y_t\}$ match. Transformation of different nature can be applied : Box-Cox transformation which directly operates on the process (Cunha *et al.*, 1999), Rozenblatt transformation which is based on the marginal distribution of the process (Monbet *et al.*, 2001a) and the $g-$transformation described in (Rychlik *et al.*, 1997) which preserves the crossing level of the process.

Sample generation in which realizations of the process $\{X_t\}$ are generated given the second order structure estimated in the previous step.

Mapping. In this step, the generated samples of $\{X_t\}$ are transformed into samples of $\{Y_t\}$ using the transformation f.

This general method includes in particular the method of Box and Jenkins (1976) and the Translated Gaussian Process method which are described more precisely hereafter.

3.1 Box and Jenkins method

The method of Box and Jenkins (1976) is undoubtedly the most usual model for wind time series (Brown *et al.*, 1984), (Daniel *et al.*, 1991), (Nfaoui *et al.*, 1996) and sea state time series (O'Carroll, 1984), (Stephanakos, 1999), (Cunha *et al.*, 1999), (Yim *et al.*, 2002). It is also used in many other application fields.

Let us first consider the univariate case for simplicity. The transformation $g = f^{-1}$ is selected in the family of the Box-Cox transformations (Cunha *et al.*, 1999):

$$g(y) = (y^\lambda - 1)/\lambda \text{ for } 0 < \lambda \leq 1 \text{ and } g(y) = ln(y) \text{ for } \lambda = 0$$

Parameter λ is selected in such a way that the marginal distribution of $X_t = g(Y_t)$ is roughly Gaussian. Various methods can be used to estimate the parameter λ (Brown *et al.*, 1984), (Daniel *et al.*, 1991). Generally, for H_s, the transformation $g(y) = ln(y)$ is used (O'Carroll, 1984), (Stefanakos *et al.*, 1999), the marginal distribution of this process being generally well approximated by a lognormal distribution. For the wind intensity, one generally applies a Box-Cox transformation with $0.5 < \lambda < 1$ (Brown *et al.*, 1984), (Nfaoui *et al.*, 1996). When the process is multivariate, a Box-Cox transformation is usually

applied independently on each component.

Then, once the transformation f has been selected, an ARMA model is adjusted to the transformed time series. For the process H_s, the following models have been proposed: AR(1) (O'Carroll, 1984), ARMA(2,2) (Stefanakos *et al.*, 1999), AR(20) (Cunha, 1999) and ARMA(4,4) (Yim, 2002). For the bivariate time series (H_s, T), Guedes Soares *et al* (2000) select AR(4) or AR(5) models depending on the location. For the wind intensity U, models AR(1) (Toll, 1997), AR(2) (Daniel *et al.*, 1991), (Nfaoui *et al.*, 1996), (More *et al.*, 2003) or AR(4) (Poggi *et al.*, 2003) have been used, more complex models giving no improvement.

3.2 Translated Gaussian Process

For wind and sea state parameters, another approach, based on the same principle is also popular. It is a non parametric method, in which the function f is selected on the basis of the normal score transformation and the Gaussian process is simulated by using exact simulation algorithms. This approach was been initially proposed by (Walton *et al.*, (1990)) in order to simulate realizations of the process H_s and then extended by (Borgman *et al.*, 1991) to multivariate time series (H_s, T, Θ_m) and it was also applied, for example, by to simulate the wind pressure on buildings (see (Gioffre *et al.*, 2000) and references therein) and by (DelBalzo *et al.*, 2003) to simulate (H_s, T, Θ_m) using buoy and ship observations . In the sequel, it will be denoted TGP (Translated Gaussian Process).

The normal score transformation, which permits to transform a continuous random variable Y into a Gaussian variable X, is defined as

$$x = N^{-1} F_Y(y)$$

where F_Y is the cumulative distribution function (cdf) of Y and N is the standard normal cdf. For multivariate variables $Y = (Y_1, \cdots, Y_n)$, a first generalization consists in applying independently the transformation on the various components:

$$(x_1, \cdots, x_d) = g(y_1, \cdots, y_d) = (N^{-1} F_{Y^{(1)}}(y_1), \cdots, N^{-1} F_{Y^{(d)}}(y_d))$$

with $F_{Y^{(i)}}$ the cdf of $\{Y^{(i)}\}$. This method is used for example in (Borgman *et al.*, 1991) and in (Gioffre *et al.*, 2000). However, it was shown in (Monbet *et al.*, 2001a) that this transformation does not allow to restore the marginal joint distribution when a strong relation exists between the components of the process. The Rozenblatt transformation (3) can then be used:

$$g : (y_1, \cdots, y_d) \rightarrow \tag{3}$$

$$\left(N^{-1}F_{Y^{(1)}}(y_1), N^{-1}F_{Y^{(2)}|Y^{(1)}=y_1}(y_2), \cdots, N^{-1}F_{Y^{(d)}|Y^{(1)}=y_1,\cdots,Y^{(d-1)}=y_{d-1}}(y_d)\right)$$

where $F_{Y^{(1)}}$ denotes the cdf of $Y^{(1)}$ and $F_{Y^{(k)}|Y^{(1)}=y_1,\cdots,Y^{(k-1)}=y_{k-1}}$ the one of the conditional distribution $P(Y^{(k)}|Y^{(1)} = y_1, \cdots, Y^{(k-1)} = y_{k-1})$. In (Monbet *et al.*, 2001a), the case of (H_s, T) was given as example. In (Ailliot *et al.*, 2001) a transformation which takes into account the specificity of the circular parameters Θ_m is proposed for the process (H_s, Θ_m). Using the same idea, in section 7 the transformation (4) is used for (U, Φ):

$$g : (u, \phi) \rightarrow (R^{-1}F_{U|\Phi=\phi}(u), F_{\Phi}(\phi)) \tag{4}$$

where R denotes the cdf of a Rayleigh distribution. If we denote $(L, \Theta) = g(U, \Phi)$, with g defined by (4), then it can be shown that the bivariate marginal of $(Lcos(\Theta), Lsin(\Theta))$ is Gaussian.

Generally, the cdf used to defined the transformation f are estimated using non parametric methods (see (Athanassoulis *et al.*, 2002) and references therein). Parametric models have also been used, in particular to approximate the tails of the distribution (Walton *et al.*, 1990). Several authors have also proposed parametric models for the multivariate distribution of metocean parameters (see (Athanassoulis *et al.*, 1994) and references therein, for example). (Ferreira *et al.*, 2002) compare parametric models and kernel density estimates to for the joint probability of H_s and T. And some papers concern more general approaches. For instance (Fouques *et al.*, 2004) describe two general methods in order to derive an approximate joint distribution from the margins. The first one matches the correlation matrix only, whereas the second one, which is based on a multivariate Hermite polynomials expansion of the normal distribution, is able to match joint moments of orders higher than two.

Once the initial process is transformed into a process which is assumed to be Gaussian, the second order structure can be estimated as in (Borgman *et al.*, 1991) and (Monbet *et al.*, 2001a). It may occur that the autocorrelation functions of the generated time series are significantly different from those of the initial sequence (see section 7). Gioffre *et al.* (2000) have proposed a solution to this problem in the particular case where the normal score transformation is applied independently on the different components.

The final step consists in simulating realizations of the stationary Gaussian process $\{X_t\}$ whose marginal distribution is the standard Gaussian distribution and whose autocorrelation function is known. Various exact simulation methods have been proposed (Borgman *et al.*, 1991), (Grigoriu, 1995), (Popescu *et al.*, 1998)

3.3 General discussion

The Box and Jenkins method is convenient to make simulation, forecasting and reconstruction (see for instance Stefanakos (1999), Guedes Soares *et al.* (1996), Ho *et al.* (2005)). Until now, TGP method has been only used for simulation, however it could also be applied for reconstruction and forecast since the underlying Gaussian process is completely characterized.

Both, Box and Jenkins and TGP methods seem to robust to noisy data (Monbet *et al.*, 2001a). These methods are easy to adapt and they have been used for various metocean parameters, as it is shown in the references above. Their main limitation is the dimension of the time series: it is difficult to apply these methods to time series with several components, in particular when the relation between the different components is complex. Indeed, in this case it may be hard to transform the original time series into Gaussian ones.

For Box and Jenkins method, statistical criteria exist to help in the choice of the order of ARMA models and to validate the model (tests on the residuals..). Moreover statistical properties of these models are well known (see Brockwell *et al.*, 1991). As far as we know, convergence properties of TGP have not been studied. When the underlying models are parametric, they do not need a large amount of data for the estimation. In TGP, the autocorrelation function and eventually the marginal distributions are estimated using non parametric methods and it may require larger data sets than in the case of parametric models.

The methods discussed in this section are extensively used in many application fields, so that they are implemented in quite a lot of softwares. And they run quickly.

It has been found that these methods provide a good description of the marginal distribution and the second order structure of the time series. However, they cannot restore some non-linearities which exist in many natural phenomena. For example, for a monovariate stationary Gaussian process with 0 mean, the time durations of the sojourns above a threshold s_0 has the same distribution as the time durations of the sojourns below the threshold $(-s_0)$. The transformed time series has similar characteristics and thus can not succeed in reproducing both calm and storm durations if they have different characteristics, as it is generally the case for sea state time series.

4 Resampling methods

Resampling methods have only been seldom used for meteorological time series. The principle is simple since it consists in randomly sampling in the data. These methods are generally used for bootstrap estimation and it was proved that it allows to estimate the distribution of a wide range of estimators in the case of independent observations (Efron *et al.*, 1993) and also of time series (Härdle *et al.*, 2003).

We describe here briefly standard methods for time series resampling. More details can be found in (Härdle *et al.*, 2003) and references therein.

4.1 Block resampling

Resampling by block is a well known method for implementing bootstrap for time series. For a time series $\{y_t\}$, blocks are defined as follows:

$$B_i = \{y_{t_i}, y_{t_i+1}, \cdots, y_{t_i+l_i}\}$$

Times t_i and block lengths l_i are sampled randomly. The resampled time series is the sequence of blocks:

$$\{B_1, \cdots, B_N\}$$

The blocks may be overlapping or not. The length of the blocks can for instance be sampled from the geometric distribution. Refer to (Lahiri, 2003) for a survey and exhaustive references.

Block bootstrap has been used to derive the statistical properties of many estimators (mean, variance, spectral density, etc.) for time series under quite low assumptions. For instance, several authors have shown that the block bootstrap is an efficient method to estimate confidence intervals for any function of the empirical mean (see (Härdle *et al.*, 2003) and references therein). But, this method may not be appropriate for applications which involve persistence statistics. Indeed, if the blocks are small the probabilities of sojourns in a set as well as autocorrelation functions may not be well reproduced. If the blocks are long the resampled time series tends to be an exact replication of the observed time series and no innovation is brought by the simulated time series.

As far as we know this method has never been used to resample sea state time series. But (Hogben *et al.*, 1987) proposed a sampling method where observed time durations of sojourn over given levels are arranged at random. The main advantage of this method is that it can be applied when only persistence data

are available. A discussion on duration statistics for sea state time series can be found for example in (Soukissian *et al.*, 2001) and (Jenkins, 2002).

4.2 Resampling Markov chains

This method consists in assuming that the observed time series is Markovian and a non parametric method is used to estimate transition kernel. Finally, realizations of the chain are simulated with this empirical kernel. Generally, the transition kernel is estimated locally using nearest-neighbor estimators.

As concerns meteorological applications, nearest-neighbor resampling was first proposed by (Young, 1994) to simulate daily minimum and maximum temperatures and precipitations. Independently, (Lall *et al.*, 1996) used an analog method to generate hydrological time series, and (Buishand *et al.*, 2001) used basically the same method for multi-site generation of artificial meteorological time series.

Nearest-neighbor resampling for time series $\{y_t\}$ is based on a very simple idea. Let us assume that we have already generated the $t-1$ first values $y_1^{sim}, \cdots, y_{t-1}^{sim}$. We search in the data the k nearest neighbors of y_{t-1}^{sim}. One of these neighbors is randomly selected and the observed value for the date subsequent to the selected point is adopted as the simulated values at time t (Buishand *et al.*, 2001)..

Various discrete probability distributions or kernels may be used to select randomly 1 of the k nearest neighbors. (Lall *et al.*, 1996) suggested to use a kernel which assigns a higher probability to the closer points, as for instance $p_j = \frac{1/j}{\sum_{i=1}^{k} 1/i}, j = 1, \cdots, k$ where the point $j = 1$ is the closest point and $j = k$ the most distant. (Monbet *et al.*, 2005a) have suggested a more sophisticated method which permits to sample points which has not been observed, as in smoothed bootstrap methods (Efron *et al.*, 1993). This method, named Local Grid Bootstrap (LGB), has also been used to simulate realizations of the multivariate sea state processes (H_s, T), (H_s, T, U) (Monbet *et al.*, 2004) and (U, Φ) (Ailliot, 2004). It was shown that this method restores most of the characteristics of the data, such as the marginal distribution, the distribution of storm durations and inter-arrivals as well as the autocovariance functions. This method is illustrated on wind data in section 7.

4.3 General discussion

As mentioned above, resampling methods can be used for simulation. Some of them can also be adapted to perform forecast and reconstruction. For in-

stance, (Ailliot *et al.*, 2003) used a nearest neighbor method to simulate time series of (H_s, T, Θ_m) at several locations $(x_0, \cdots, x_L)$ along a line to study the profitability of a maritime line in Aegean sea (see also Section 7). (Caires *et al.*, 2005) proposed a non parametric regression method to correct outputs of meteorological models. And another method, based on a non-parametric Hidden Markov model is described in (Monbet *et al.*, 2005b) to reconstruct H_s time series given wind time series (see also (Marteau *et al.*, 2004a)). However, one of the main drawback of non parametric methods is that their descriptive power is about null.

All the methods introduced in this section can be used with noisy and missing data. Indeed, missing data do not affect the nearest neighbor search, and aberrant data may have a positive probability to be resampled but this probability can be estimated. One of their main advantage is that they can be easily adapted for various time series (wind, wave, circular time series, ...) because of the few assumptions required. For circular data, a particular attention has to be paid to the choice of the distance used for nearest neighbor search and kernel estimation (Athanassoulis *et al.*, 2002). If simulation is combined with cross-reconstruction, the dimension of the time series is not a difficulty. Indeed, when the dimension is high, simulation can be performed in two steps: the time series of a subset of components are generated first, then the other components are simulated given the first time series (see (Ailliot *et al.*, 2003) and Section 7).

Some asymptotic convergence properties of resampling and reconstruction methods have been studied (see (Monbet *et al.*, 2005a), (Monbet *et al.*, 2005b) and references therein). In practice, it is known that a large amount of data is needed for estimation in non parametric methods. An other well known drawback is the difficulty to choice the smoothing parameters such as the bandwidth parameters which appear in the probability density function estimators.

As concern the computational aspects, algorithms for resampling are generally very simple to implement but they can be time consuming. A solution consists in working with algorithms based on tree structures for the nearest neighbor search (Monbet *et al.*, 2004).

5 Parametric models

In this section, we discuss various parametric models which have been proposed for wind and sea state time series. Linear autoregressive models are discussed in section 3, and we focus now on finite state space Markov chain models and on non linear autoregressive models. A section is also devoted to

circular time series.

5.1 Finite State Space Markov chains

Let us briefly describe the general principle of the methods considered in this section. gt;

When the process has a continuous state space, it is discretized in a finite number of classes. This allows to use a representation as a finite state space process.

This finite state space process is then supposed to be a Markov chain whose transition matrix is estimated from the observed sequence.

New realizations of the process are finally simulated given the estimated transition matrix.

The success of this method is undoubtedly mainly due to its simplicity, and the main limitation is probably the number of parameters which have to be estimated in the case where no assumption is made on the shape of the transition matrix. Various models have been proposed to reduce the number of parameters. For instance, (Vik, 1981), assumed that H_s is a first-order Markov chain with tridiagonal transition matrix, which means that only the transitions to the closest states are allowed. The transition matrix is estimated in such a way that the stationary distribution of the Markov chain and the mean durations of persistence above certain levels match with those of the observed time series. A more general approach is proposed in (Raftery, 1985) who introduced the Mixture Transition Distribution model. If r is the order of the Markov chain, it is assumed that

$$P(Y_t = j | Y_{t-1} = i_1, ..., Y_{t-r} = i_r) = \lambda_1 q_{i_1,j} + ... + \lambda_r q_{i_r,j}$$

with $Q = (q_{i,j})$ is a stochastic matrix and $(\lambda_1, \ldots, \lambda_2)$ are positive parameters such that $\lambda_1 + ... + \lambda_r = 1$. This model has been fitted to time series of wind speed (Raftery, 1985) and wind direction (Raftery *et al.*, 1994), (MacDonald *et al.*, 1997). See also (Berchtold *et al.*, 2002) for a recent review on Mixture Transition Distribution models.

In (Ailliot *et al.*, 2001) another approach is proposed for (H_s, Θ_m) where the data are preprocessed before using a first-order Markov chain model. The preprocessing consists in detecting the slope change times by fitting a linear spline curve H_{lin} to the H_s time series. A marked point process (τ, H) is then associated to H_{lin}, where τ denotes the dates when the slope changes and H the significant wave height at these dates. Then a Markov model is proposed

for the trivariate process (H, τ, Θ_m). It was found that this model permits to reproduce the cdf of H_s and Θ_m, as well as the mean duration of the storms.

5.2 Nonlinear autoregressive models

In this part, we describe various autoregressive models which have been introduced recently to model some non linearities which can be observed on wind and sea state time series.

Artificial Neural Networks - During the last decades, artificial neural networks (ANN) was successfully used to solve problems in ocean, coastal and environmental engineering applications. ANN can be seen as particular regression models in which the link function simulate the circulation of the information in a biological neural network. The parameters of the regression model are generally estimated by maximum likelihood or least square methods. See for instance (Gurney, 1997) for an introduction to ANN. ANN was used in order to model the evolution of U (Stephos, 2000), (More *et al.*, 2003) and (H_s, T) (Deo *et al.*, 2001), (Makarynskyy *et al.*, 2004). The results obtained by these authors show that ANN models permit to obtain better short term forecasts than those obtained with linear autoregressive models. (Arena *et al.*, 2004) illustrate how multivariate ANN can be used to reconstruct H_s time series in buoy networks.

Time-varying autoregressive models - (Huang *et al.*, 1995) use a time-varying autoregressive model of order 2 in order to forecast the wind speed. The coefficients of the model are estimated in real time following the idea of (Young *et al.*, 1991). The autoregressive model of order r is given by:

$$U_t = \sum_{i=1}^{r} a_{t,i} U_{t-i} + \epsilon_t \tag{5}$$

Let $\Psi_t = (a_{t,1}, \cdots, a_{t,r})$, then

$$\begin{cases} \Psi_t = \psi_{t-1} + \Gamma_{t-1} \\ \Gamma_t = H\Gamma_{t-1} + \Omega_t \end{cases} \tag{6}$$

where I is the identity matrix, H is a diagonal matrix and Ω is a zero white noise vector. Γ is a vector of dummy parameters which is either a white noise process or a random walk process. The unknown parameter vector Ψ is either a

random walk process or a smoothed integrated random walk process depending on the matrix H. The parameters of model (5)-(6) are estimated by a Kalman filter with state vector Ψ given observation vector $\{u_1, \cdots, u_t\}$.

(Scotto *et al.*, 2000) have proposed a Self-Exciting Threshold AutoRegressive model for the process $Y = H_s$. It is assumed that:

$$Y_t = \sum_{i=1}^{r} a_i^{(S_t)} Y_{t-i} + b^{(S_t)} + \sigma^{(S_t)} \epsilon_t$$

with $S_t = i$ if and only if $Y_{t-d} \in [r_i, r_{i+1}]$ for a fixed integer d. Here, $r_1 < r_2 < \cdots < r_M$ are parameters of the model and ϵ_t denotes a Gaussian white noise. It is a switching autoregressive model, in which the regime at time t only depends on the last values of the process. In practice, the identified model has 2 regimes, the evolution in the different regimes being described by $AR(10)$ models and $d = 7$, which corresponds to a delay of 21 hours. The authors compare the results obtained with this model and those corresponding to an $AR(22)$ model. The sequences simulated with the Self-Exciting Threshold AutoRegressive model was shown to have features closer to the data than those simulated with the model $AR(22)$, in particular as concern the marginal distribution and the autocorrelation function.

(Ailliot *et al.*, 2003) proposed a Markov Switching Autoregressive model (MS-AR) for the wind intensity. In MS-AR models the observed proposed $\{Y_t\}$ is represented by an autoregressive model of order r with parameters depending on a non observable Markov chain $\{S_t\}$. This hidden variable represents the "weather type". More precisely, a bivariate process $\{S_t, Y_t\}$ follows a MS-AR model if gt;

$\{S_t\}$ is a Markov chain on a finite space $\{1, ..., M\}$ with $M > 0$ the number of regimes. This process is supposed to be hidden.

Conditionally to $\{S_t\}$, $\{Y_t\}$ is a non-homogeneous Markov chain of order $r \geq 0$ on $\mathbf{Y} \subset \mathbf{R}^d$. More precisely, we assume that the conditional distribution of Y_t given $\{Y_{t'}\}_{t'<t}$ and $\{S_{t'}\}_{t'\leq t}$ only depends on S_t and $\overline{Y}_{t-1} = (Y_{t-1}, \cdots, Y_{t-r})$. Then, it is assumed that for each $s_t \in \{1, ..., M\}$ and $\overline{y}_{t-1} \in \mathbf{Y}^r$, the conditional distribution $P(Y_t | \overline{Y}_{t-1} = \overline{y}_{t-1}, S_t = s_t)$ is a gamma distribution with mean $\sum_{i=1}^{p} a_i^{(s_t)} y_{t-i} + b^{(s_t)}$ and standard deviation $\sigma^{(s_t)}$.

MS-AR models are simple generalizations of Hidden Markov models (HMM), which correspond to the case $r = 0$. In (Ailliot *et al.*, 2003), a non homogeneous MS-AR models is also proposed for the process (U, Φ), where the transition probabilities of the hidden Markov chain $\{S_t\}$ depend on the wind direction. More precisely, it is assumed that $\{S_t\}$ is a non homogeneous Markov chain

on $\{1, \cdots, M\}$ with a transition matrix such that

$$P(S_t = j | S_{t-1} = i) \simeq \pi_{i,j} exp(\kappa^{(j)} cos(\Phi_t - \phi^{(j)}))$$

where $\Pi = (\pi_{i,j})_{i,j \in \{1,\cdots,M\}}$ is a stochastic matrix, $(\kappa^{(j)})_{j \in \{1,\cdots,M\}}$ are positive parameters and $(\phi^{(j)})_{j \in \{1,\cdots,M\}}$ denote parameters in $[0, 2\pi[$. This model is validated on wind data in North Atlantic in (Ailliot, 2004), and it is shown that the model restores the marginal distributions, the autocorrelation functions and the distribution of the durations of storms as well as their inter-arrivals. This model is also validated on wind data in Aegean Sea in section 7.

GARCH model - GARCH models have been proposed in (Toll, 1997) for the process $Y = U$. In this paper, it is supposed that U is Markovian of order r, the conditional distribution of Y_t given $(Y_{t-1}, \cdots, Y_{t-r})$ being described by a gamma distribution with mean

$$\mu_t = \sum_{i=1}^{r} a_i Y_{t-i} + b$$

and variance

$$\sigma_t^2 = \alpha + \sum_{i=1}^{p} \lambda_i (Y_{t-i} - \mu_{t-i})^2 + \sum_{i=1}^{q} \kappa_i \sigma_{t-i}^2$$

The identified model is of order r=2 and it is shown that it makes it possible to predict the heteroscedasticity present in the wind time series.

5.3 Models for circular data

In the section 5, we have introduced a first family of model for directional time series. Circular time series can also be described by autoregressive models. Let $\{\Phi_t\}$ be a stationary process with values in $\mathbb{R}/2\pi\mathbb{Z}$. Mainly three types of autoregressive models have been proposed in the literature for such time series, and the use of these models for time series of wind direction is discussed in (Breckling, 1989). gt;

Models obtained by "wrapping" a real valued process (Wrapped Autoregressive model). One supposes that $\Phi_t = Y_t$ modulo 2π with $\{Y_t\}$ a real valued process which follows an autoregressive model.

Models obtained by using a "link function", $g : \mathbb{R} \to (-\pi, \pi)$ strictly monotonous and checking $g(0) = 0$. This function is used to transform a real valued autoregressive process $\{Y_t\}$ into a process $\Phi_t = g(Y_t)$ with values on the torus $\mathbb{R}/2\pi\mathbb{Z}$.

Models specifying directly the density of the conditional distribution $P(\Phi_t|\Phi_{t-1},\cdots,\Phi_{t-k})$. It is for example the case of the autoregressive model of Von-Mises proposed initially in (Breckling, 1989). The Von-Mises distribution with parameters (θ_0,k) is defined by its density:

$$f(\theta)=\frac{1}{2\pi I_0(k)}e^{\kappa\cos(\theta-\theta_0)}$$

for $\phi\in R/2\pi\mathbb{Z}$ with $k>0$ the concentration and $\theta_0\in R/2\pi\mathbb{Z}$ the mean direction. The Von-Mises autoregressive model is then defined in the following way: it is supposed that $P(\Phi_t|\Phi_{t-1},...,\Phi_{t-k})$ follows a Von Mises distribution with parameters (θ_t,k_t) given by

$$\kappa_t e^{i\theta_t}=\kappa_0 e^{i\phi_0}+\kappa_1 e^{i\phi_{t-1}}+\cdots+\kappa_p e^{i\phi_{t-p}}$$

with $\kappa_0,\kappa_1,\cdots,\kappa_p\in\mathbb{R}^+$ and $\Phi_0\in\mathbb{R}/2\pi\mathbb{Z}$. In this model, the concentration k_t changes in time (heteroscedastic model) but it can also be assumed to be constant.

HMM have also been proposed for time series of wind direction. In (MacDonald *et al.*, 1997), discrete (multinomial) distributions are used whereas in (Holzmann *et al.* , 2005) continuous (wrapped-normal) distributions are used.

5.4 General discussion

A parametric model is generally build for a particular task and describes particular features of the data. For instance, ANN are currently used for short term forecasts (for time lag varying from 3 hours to 48 hours) and reconstruction. (Pittalis *et al.*, 2003) and (Tsai *et al.*, 2002) used ANN to reconstruct H_s and (H_s,T) missing data in a buoy network. ANN have also been extensively applied to correct wind outputs of meteorological model (see (Giebel *et al.*, 2003) and references therein). One drawback of ANN is the large number of parameters that they involve and their lack of interpretability. MS-AR and GARCH models are more parsimonious and physically interpretable. They can be used to generate artificial sequences which restore specific features of the data (heteroscedasticity, existence of several meteorological regimes, ...) as it is illustrated in section 7.

The form of the model has to be specified and this is generally time expensive because it requires a precise study of the data before the construction of the model. It also limits the possibilities to export the model to other time series. However, once the model is chosen, the estimation of the parameters can be performed automatically and should require less data than for non-parametric models. In general, the parameters are estimated using the

methods of maximum likelihood or least square. When maximum likelihood is used the existence of missing values may complicate the statistical inference. Statistical criteria such as Akaike (AIC) or Schwartz Bayesian (BIC) criteria can be used to select the best model.

Most of the parametric models discussed in this section are usual for time series, and softwares are available to fit these models.

6 Validation and comparison method

In the previous sections, various stochastic models are proposed for wind and sea state time series. And we need criteria to chose among these models. In this section, we present a general validation method, based on Monte Carlo tests, which can be used to measure the ability of a model to simulate realistic synthetic time series.

The most widespread method for model validation consists in comparing certain statistics calculated from the observations with those corresponding to the considered model. In general, several criteria are used, such as the matching of the mean and the variance of the marginal distributions, or more generally its cdf. When the temporal dependence is important for the applications, other features are also considered, like the autocorrelation functions or the distribution of the time duration of sojourns below or above given levels.

Meanwhile, the authors often perform only visual comparisons. Such approach remains not entirely satisfactory because it does not make it possible to decide whether the observed differences are significant or not. A more formal method, based on Monte Carlo tests, is proposed below. For the sake of simplicity, it is presented in the simple case of comparing means, but its generalization to other statistics is straightforward.

Let $\{y_t\}_{t=1,\cdots,T}$ be an observed sequence of a real valued process $\{Y_t\}$ with mean m and $\{Z_t\}$ a process corresponding to the model which has to be validated. The mean of the marginal distribution of $\{Z_t\}$ is denoted m_0. We want to test

$$H_0 : m = m_0 \quad \text{versus} \quad H_1 : m \neq m_0 \tag{7}$$

The considered test statistic is the empirical mean $\overline{Y} = \frac{1}{T}\sum_{t=1}^{T} Y_t$, and the associated decision rule is given by

$$H_0 \text{ is rejected if } \overline{y} = \frac{1}{T}\sum_{t=1}^{T} y_t \in R(\alpha)$$

where α is the level of the test. The critical region $R(\alpha)$ is such that $P_{H_0}(\overline{Y} \in R(\alpha)) = \alpha$.

In order to compute $R(\alpha)$, we need to know the distribution of the test statistic $\overline{Y}$ when H_0 is true. When the model is complex, it is not always possible to derive the exact distribution of the test statistic. In this case, we can use the Monte Carlo method described hereafter to approximate this distribution: gt;

Simulate B time series of length T with the model:

$$\begin{array}{c} \{z_1^{(1)}, \cdots, z_T^{(1)}\} \\ \vdots \\ \{z_1^{(B)}, \cdots, z_T^{(B)}\} \end{array}$$

Compute the empirical mean $\overline{z}^{(i)} = \frac{1}{T}\sum_{t=1}^{T} z_t^{(i)}$ for each simulated sample $i = 1, \cdots, B$

Approximate the distribution of $\overline{Y}$ under H_0 by the empirical distribution of $\{\overline{y}^{(1)}, \cdots, \overline{y}^{(B)}\}$. This allows to compute an approximation of $P_{H_0}(\overline{Y} \in R)$ for any region $R \subset \mathbb{R}$ or equivalently deduce an approximative critical region $\tilde{R}(\alpha)$ such that $\frac{1}{B} card(\{i \in \{1, \cdots, B\} | \overline{z}^{(i)} \in \tilde{R}(\alpha)\}) = \alpha$.

Finally, H_0 will be accepted if and only if $\overline{y} \in \tilde{R}(\alpha)$.

This framework can be applied to compare other features like cdf or autocorrelation functions. For this, a test statistic has to be chosen. As concern cdf, the most popular test statistic is probably the Kolmogorov-Smirnov distance. However, it is well known that it is more sensitive near the center of the distribution than at the tails. Due to this limitation, many analysts prefer to use the Anderson-Darling statistic which gives more weight to the tails. But, as this is an integrated deviation, like the Cramer-Von-Mises or chi-square distance, it can mask local differences. In (Ailliot *et al*, 2005), a more sensitive test statistic was proposed, which permits to measure locally the goodness-of-fit. This method is used in the next subsection to validate and compare three models.

7 Example: simulation of multivariate sea state time series along a maritime line

In this section, an example is proposed in order to illustrate some of the models discussed in the previous sections. The data sets considered in this part have also been used in (Ailliot *et al.*, 2003). The initial objective of this study was to estimate the profitability of the maritime line Piraeus-Heraklion (Aegean Sea) for a given passenger ship. For that, we had at our disposal two data sets:

gt;

3 years of data (from October 1999 to September 2002) describing the sea state conditions (i.e. (H_s, T, Θ_m)) at $N = 17$ points along the maritime line. These points will be denoted $(x_1, \cdots, x_N)$.

11 years (from January 1992 to December 2002) of wind data at a point denoted x_0, located near the middle of the maritime line (see 1)

The sea state data have been produced using the WAM model by the National Center for Marine Research (NCMR), and the wind data are reanalysis data produced by ECMWF. The amount of sea state data is clearly to small in order to get reliable estimates of the variability of the sea state conditions, and thus we have proposed to use a stochastic generator. In (Ailliot *et al.*, 2003), this sea state generator is combined with a crossing simulator. This crossing simulator is based on polar diagrams which describe the response of the passenger ship under consideration in different conditions. These diagrams permit to compute, according to the sea state conditions on the maritime line, if the crossing can be done in normal conditions, or has to be delayed or canceled. And as a final result, we can deduce estimates of the distribution of canceled and delayed crossings. In this paper, we will only focus on the sea state generator.

The Piraeus-Heraklion line is located in Aegean Sea, which is a relatively closed sea. Thus we can consider that the waves are essentially generated by local winds and it seems natural to perform the simulation in two steps. At first, a stochastic model is used to simulate artificial wind conditions at x_0, and then the sea state conditions corresponding to these artificial wind conditions are cross-reconstructed. These two steps are described more precisely hereafter.

Simulation of artificial wind conditions

In order to simulate artificial wind conditions at x_0, we have tried three different methods: gt;

The TGP method discussed in section 3. The transformation (4) is used to

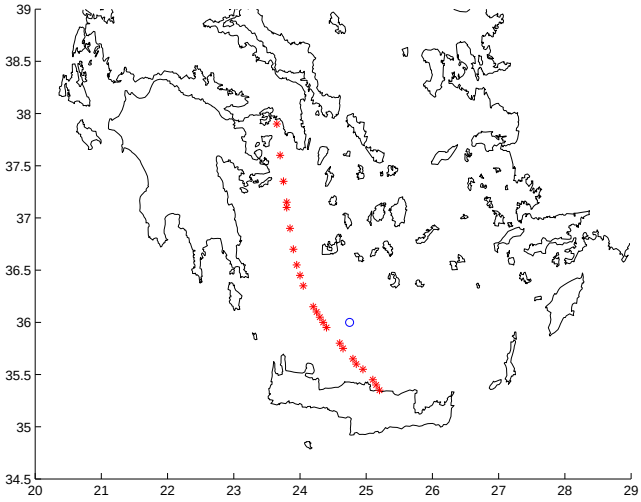

Figure 1. Map of the Aegean sea. Maritime line $(x_1, \cdots, x_N)$ (stars) and point x_0 (circle)

transform the original time series and an exact simulation method is used to simulate the underlying Gaussian process.

The LGB method discussed in section 4.

The non-homogeneous MS-AR model introduced in section 5. In order to simulate the wind direction, we have used a simple finite state Markov chain of order 1 (see section 5.1).

These different methods have been fitted to the wind data for the months of August. Then, they have been validated using the method described in section 6. In order to check the realism of the simulated sequences, the list of criteria just below has been used. gt;

F_U: cdf of the marginal distribution of U

F_Φ: cdf of the marginal distribution of Φ

$F_{(U,\Phi)}$: cdf of the bivariate distribution of $\{U, \Phi\}$

F_{extr}: cdf of the monthly maxima of U

C_U: autocorrelation function of U

$F_{[U>1/2]}$: cdf of sojourn durations above level $1/2 \max(u)$, with $\max(u)$ the largest wind speed in the observed time series

$F_{[U<1/3]}$: cdf of sojourn durations below level $1/3 \max(u)$

The cdf F_U, F_Φ and $F_{(U,\Phi)}$ are important criteria since the distribution of the wind is strongly related to the distribution of the sea state parameters. F_{extr} describes the interannual variability of the process at a monthly scale. The autocorrelation function C_U is a usual measure of linear dependence in time. Finally, the cdf of sojourn durations $F_{[Y>1/2]}$ and $F_{[Y<1/3]}$ describe the time duration of the stormy and calm conditions. These durations are also strongly related to the severity of the sea state conditions.

For each criteria, Monte Carlo tests have been run on the basis of $N = 1000$ synthetic time series, each of them having the same length as the initial wind time series. The results are given in Table 1. According to this table, the TGP method successfully reproduces the bivariate marginal distribution of the process but fails to reproduce its dynamics. This is also illustrated on Figure 2. According to this figure, the model underestimates the first coefficients of the autocorrelation function and also the durations of the calm and stormy periods. It indicates that the transformed time series is not Gaussian, although its bivariate marginal distribution is Gaussian. On the opposite, the LGB method successfully describes the dynamics of the process but can not restore the marginal distribution of U and (U, Φ). According to Figure 3 this method simulates too many data close to the mode of the distributions. It is a well known problem of this kind of algorithm, and better results could perhaps be obtained with an other choice of the bandwidth parameters. However, the search of appropriate parameters is fastidious since there is no automatic criterion. As concerns the non-homogeneous MS-AR model, the model with $M = 2$ regimes and autoregressive models of order $p = 1$ has been selected with BIC, and the results obtained with this model are better than the ones corresponding to LGB and TGP. Indeed, this model reproduces all the criteria, except the bivariate marginal distribution. According to Figure 4, the model simulates too many wind from the north-west and not enough from the north, but the distribution of the wind intensity in each sector seems to be well reproduced. A better description of this bivariate distribution could probably been obtained by using a more sophisticated model for the wind direction and a higher number of regimes.

An other advantage of this model over the other models considered in this section is its physical interpretability. The maximum likelihood estimates are given in Table 2. According to this table, one important difference between the two regimes is the value of the conditional standard deviation σ, which is higher in the second regime. It indicates that the second regime is associated to weather conditions in which the wind speed evolves quickly, whereas the first regime corresponds to steady wind speed conditions (low volatility). This is illustrated on Figure 5. And the two regimes are associated to different wind directions (see Figure 6), the first one corresponding mainly to Northerlies and the second one to Westerlies. This relation is described through the parameters ϕ and κ.

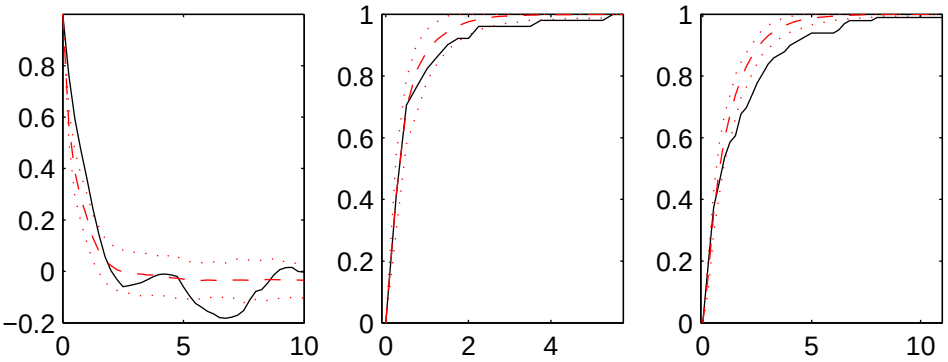

Figure 2. Autocorrelation function of the wind intensity U (left), cumulative distribution functions of sojourn durations over $1/2\max(U) = 8.05ms^{-1}$ (middle) and below $1/3\max(U) = 5.35ms^{-1}$(right). The time in abscissa is expressed in days. Solid: observation, dashed: TGP model, dotted 95% interquantile interval for TGP model.

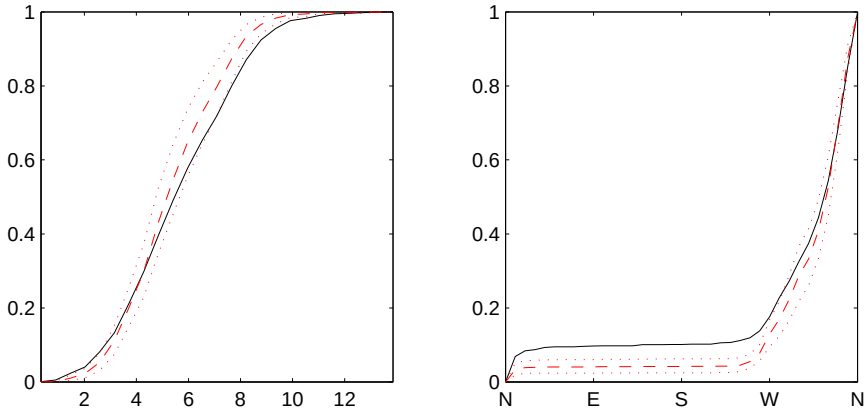

Figure 3. cdf of the marginal distribution of the wind intensity U (left) and the wind direction Φ (right). Solid: observation, dashed: LBG model, dotted 95% interquantile interval for LGB model.

Reconstruction of the sea state conditions The sea state conditions $\{H_s^{(sim)}, T^{(sim)}, \Theta_m^{(sim)}\}$ corresponding to the simulated wind $\{U^{(sim)}, \Phi^{(sim)}\}$ are reconstructed by a nearest neighbor resampling method. In practice, at each time t and each location x_i, $(H_s^{(sim)}(x_i,t), T^{(sim)}(x_i,t), \Theta_m^{(sim)}(x_i,t)) = (H_s(x_i,s^*), T(x_i,s^*), \Theta_m(x_i,s^*))$ where s^* is solution of

$$\begin{aligned}\min_s \Big\{\omega_1||U^{(sim)}(t)e^{\sqrt{-1}\Phi^{(sim)}(t)} - U(s)e^{\sqrt{-1}\Phi(s)}|| \text{ amp; amp; amp;}\\ +\omega_2||H_s^{(sim)}(x_{i_0},t-1)e^{\sqrt{-1}\Theta_m^{(sim)}(x_{i_0},t-1)} - H_s(x_{i_0},s-1)e^{\sqrt{-1}\Theta_m(x_{i_0},s-1)}||\text{amp; amp; amp;}\\ +\omega_3||T^{(sim)}(x_{i_0},t-1) - T(x_{i_0},s-1)||\Big\}\text{amp; amp; amp;}\end{aligned}$$

where $\omega_1, \omega_2, \omega_3$ denote fixed weights, $||.||$ denotes the euclidean norm and x_{i_0} is the point of the line $\{x_1, \cdots, x_N\}$ which is the closest to x_0 ($i_0 = 16$). Here,

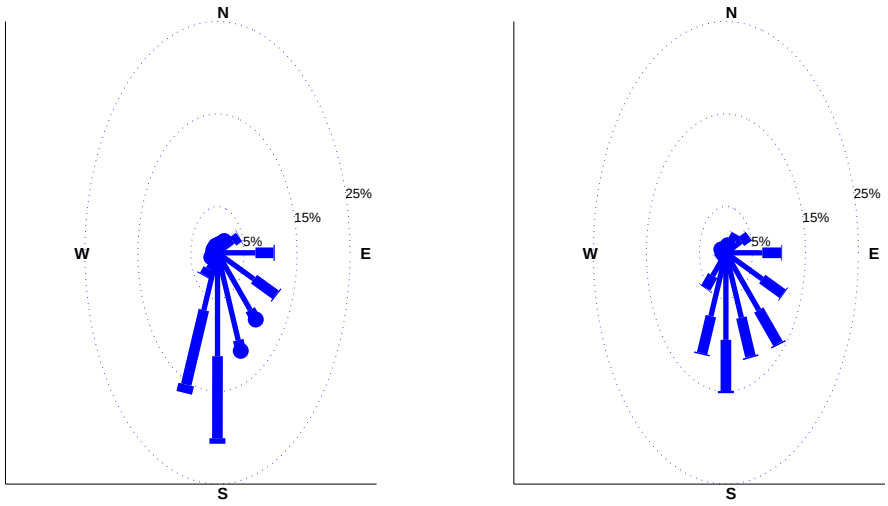

Figure 4. Wind roses for the data (left) and the data simulated with the non homogeneous MS-AR model (right)

	TGP	LGB	MS-AR
F_U	.282 *[.003]*	.000 *[.001]*	0.078 *[.005]*
F_Φ	.003 *[.000]*	.001 *[.000]*	.004 *[.000]*
$F_{(U,\Phi)}$	0.210 *[.001]*	.000 *[.001]*	.000 *[.001]*
F_{extr}	.012 *[.012]*	.025 *[.009]*	.137 *[.008]*
C_U	.000 *[.001]*	.006 *[.003]*	.001 *[.000]*
$F_{[U>1/2]}$	.000 *[.007]*	.046 *[.006]*	.015 *[.002]*
$F_{[U<1/3]}$	.000 *[.007]*	.068 *[.002]*	.042 *[.003]*

Table 1
Results of the Monte Carlo tests for the wind time series. The first value is the observed statistic w^{obs} and the value in bracket the cut-off value w_α with $\alpha = 0.05$. The null hypothesis is rejected at the level α if $w^{obs} < w_\alpha$

	$a_1^{(i)}$	$b^{(i)}$	$\sigma^{(i)}$	$\pi_{i,i}$	$\phi^{(i)}$	$\kappa^{(i)}$
Regime 1 (i=1)	0.85	0.82	1.27	0.96	1.64	1.47
Regime 1 (i=2)	0.53	2.11	2.24	0.95	1.59	2.71

Table 2
Maximum likelihood estimates for the non-homogeneous MS-AR model

$\omega_1, \omega_2, \omega_3$ are chosen empirically.

We have used this simple method to compute the sea-state conditions corresponding to the artificial wind conditions simulated with the non homogeneous MS-AR model. Then, in order to check the realism of these artificial sea state conditions, we have performed Monte Carlo tests, and the list of criteria just

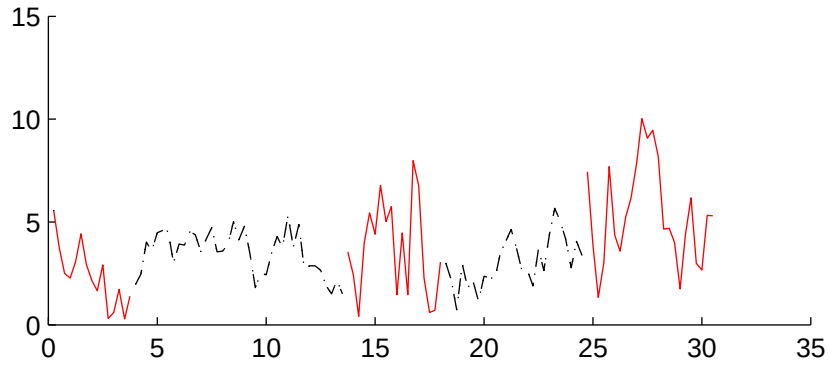

Figure 5. Evolution of the wind speed at x_0 in August 2002. The dash-dotted [resp. solid] line represents the date when the first [resp. second] regime is the most likely. The regimes have been identified using the Viterbi algorithm

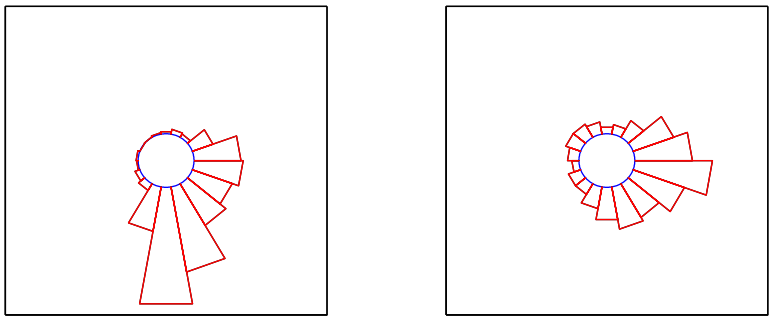

Figure 6. Distribution of the wind direction in the two regimes (identified using the Viterbi algorithm). Regime 1 on the left and regime 2 on the right.

below has been used.

gt;

F_{H_s} : cdf of the marginal distribution of H_s

F_{Θ_m} : cdf of the marginal distribution of Θ_m

F_T : cdf of the marginal distribution of T

$F_{(H_s,\Theta_m)}$: cdf of the bivariate marginal distribution of (H_s, Θ_m)

$F_{(H_s,T)}$: cdf of the bivariate marginal distribution of (H_s, T)

$F_{(U,H_s)}$: cdf of the bivariate marginal distribution of (U, H_s)

C_{H_s}: autocorrelation function of H_s

$F_{[H_s>1/2]}$: cdf of sojourn durations above level $1/2 \max(H_s)$

$F_{[H_s<1/3]}$: cdf of sojourn durations below level $1/3\max(u)$

Table 3 shows that all the considered statistics are well reproduced at location 16 (located near the middle of the maritime line), except the marginal distributions of Θ_m and (H_s, Θ_m). This bivariate distribution is shown on Figure 7 and the lack of fit is in accordance to the one identified on the simulated wind time series: the proportion of wave coming from the north is underestimated. The joint distribution of U and H_s is shown on Figure 8.

F_{H_s}	F_{Θ_m}	F_T	$F_{(H_s,\Theta_m)}$	$F_{(H_s,T)}$
.007 *[.004]*	.000 *[.002]*	.017 *[.011]*	.000 *[.001]*	.017*[.011]*
$F_{(U,H_s)}$	C_{H_s}	$F_{[H_s>1/2]}$	$F_{[H_s<1/3]}$	
.027 *[.005]*	.024 *[.001]*	.011 *[.001]*	.031 *[.005]*	

Table 3
Results of the Monte Carlo tests for the sea state time series at location 16. The first value is the observed statistic w^{obs} and the value in bracket the cut-off value w_α with $\alpha = 0.05$. The null hypothesis is rejected at the level α if $w^{obs} < w_\alpha$

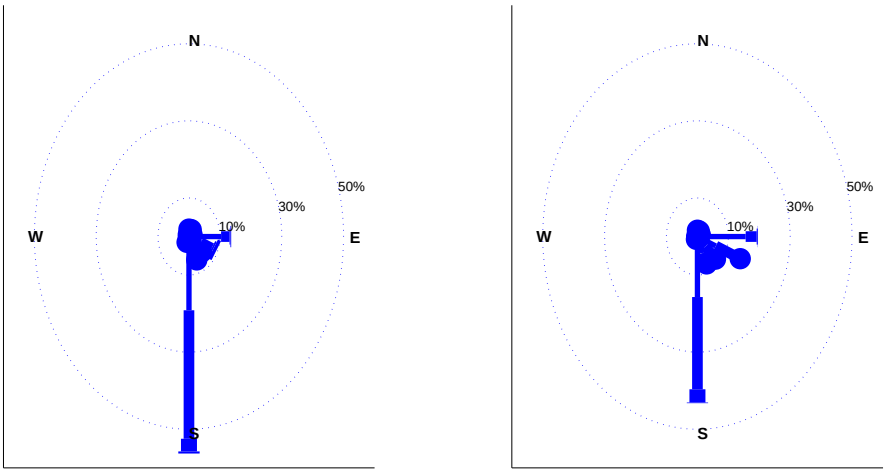

Figure 7. Wave roses for the data at location 16 (left) and for the simulated sequences (right)

8 Concluding remarks

In this paper, we make a review of stochastic models for metocean time series. The models are classified in three groups: gt;

non parametric models

models based on Gaussian approximations

other parametric models

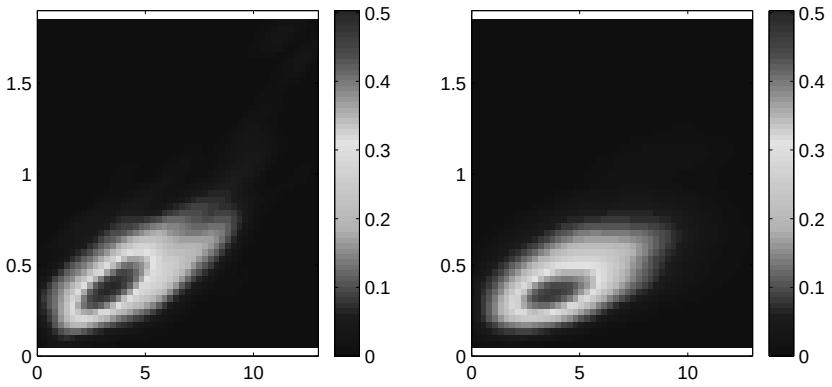

Figure 8. Joint distribution of U (x-axis) and H_s (y-axis). Data on the left and model on the right

For each group of models, we discuss the possible uses of the models, their advantages and drawbacks, etc. Then a quantitative method is proposed to measure the ability of a model to restore chosen statistical feature like marginal distribution, covariance functions, durations, etc., and this method allows to validate or compare models for given data. And finally an example is discussed where a stochastic model is used to generate a multivariate time series $\{U, \Phi, H_s, T, \Theta_m\}$ at several location along a ferry line in Aegean Sea (Greece).

Finally, a lot of tools and methods are available for modeling metocean time series and the choice of a model depends on the nature of the studied process (univariate or bivariate, intensity and/or direction, ...), of the considered location and also on the objectives of the users.

The review focus on models for time series at the scale of the sea state. And, as a consequence, we have neglected many usual and interesting aspects of metocean studies, such as, for instance, linear and non linear models for waves, extremes, spatial process.

References

[1] Ailliot, P., Prevosto, M., (2001). Two methods for simulating the bivariate process of wave height and direction. *Proc. of ISOPE Conf.*, **III**, 15-18.

[2] Ailliot, P., Prevosto, M., Soukissian, T., Diamanti, C., Theodoulides, A., Politis C., (2003). Simulation of sea state parameters process to study the profitability of a maritime line. *Proc. of ISOPE Conf.*, **III**, 51-57.

[3] Ailliot, P., (2004). *Modèles autorégressifs à changement de régimes markovien - Application à la simulation du vent.* PhD Thesis. Université de Rennes 1.

[4] Ailliot, P., Monbet, V., Prevosto, M., (2006a). An autoregressive model with time-varying coefficients for wind fields. *Environmetrics*, **17**(2), 107-117.

[5] Ailliot, P., Frenod, E., Monbet, V. (2006b). Long term object drift forecast in the ocean with tide and wind. *To appear in Multiscale Modeling and Simulation.*

[6] Arena, F., Puca, S., (2004). The reconstruction of significant wave height time series by using a neural network approach. *J. Offshore Mechanics and Artic Eng.*, **126**(3), 213-219.

[7] Athanassoulis, G.A., Skarsoulis, E.K., Belibassakis, K.A., (1994). Bivariate distributions with given marginals with an application to wave climate description. *Applied Ocean Research*, **16**(1), 1-26.

[8] Athanassoulis, G.A., Belibassakis, K.A., (2002). Probabilistic description of metocean parameters by means of kernel density models. 1. theoritical background and first results. *Appl. Oc. Research* **24**, 1-20.

[9] Athanassoulis, G.A., Stephanakos, C.N., (1995). A nonstationary stochastic model for long-term time series of significant wave height. *Journal of Geophysical Research, Section Oceans* **100**(C8), 14149-14162.

[10] Baxevani and I. Rychlik (2004). Fatigue Life Prediction for a Vessel Sailing the North Atlantic Route. *Proc. ISOPE PACOMS* .

[11] Berchtold, A. and Raftery, A.E. (2002). The Mixture Transition Distribution (MTD) model for high-order Markov chains and non-Gaussian time series. *Statistical Science*,**17**, 328-356.

[12] Borgman L.E., Sheffner N.W. (1991). *Simulation of time sequences of wave height, period and direction.* Technical report US Army Corps of Engineers.

[13] Boukhanovksy, A., Lavrenov, I., Lopatoukhin, L., Rozhnov, V., Divinsky, B., Kosyan, R., Abdalla, S., Ozhan, E., (1999). Persistence statistics for Balck and Baltic seas. *Proc. of Int. MEDCOAST conf. "Wind and Wave climate, 99"*, 199-210.

[14] Box, G.E.P., Jenkins, G.M., (1976). *Time series analysis, forecasting and control.* (revised edn.) Holden-Day, San Francisco.

[15] Breckling, J. (1989). *The Analysis of Directional Time Series.* Lecture Notes in Statistics Series.

[16] Brockwell, P., Davis, R., (1991). *Time series: theory and methods, 2nd edition.* Springer Verlag, New York.

[17] Brown, B.G., Katz, R.W, Murphy, A.H. (1984). Time series models to simulate and forecast wind speed and wind power. *J. of clim. and appl. meteor.* **23**, 1184-1195.

[18] Buishand, T.A., Brandsma, T., (2001). Multi-site simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling. *Water Resources Research*, **37**, 2761-2776.

[19] Caires, S., Sterl, A., (2005). A new non-parametric method to correct model data: Application to significant wave height from the ERA-40 reanalysis. *J. Atmospheric and Oceanic Tech., In press.*

[20] Castino, F., Festa, R., Ratto, C.F. (1998). Stochastic modelling of wind velocities time series, *J. of Wind Engineering & Industrial Aerodynamics*, **74-76**, 141-151.

[21] Cunha, C., Guedes Soares, C. (1999). On the choice of data transformation for modelling time series of significant wave height. *Ocean Engng*, **26**, 489-506.

[22] Daniel, A.R., Chen, A.A. (1991). Stochastic simulation and forecasting of hourly average wind speed sequences in jamaica. *Solar Energy*, **46**(1), 1-11.

[23] DelBalzo D.R., Schultz, J.R., Marshall, D.E., (2003). Stochastic time-series simulation of wave parameters using ship observations. *Ocean Engng.* **30**(11), 1417-1432.

[24] Deo, M.C., Jha, A., Chapekar, A.S., Ravikant, K., (2001). Neural network for wave forcasting.*Ocean Engng*, 28, 889-898.

[25] Efron, B., Tibshirani, R., (1993). *An introduction to the bootstrap.* Chapman Hall, New York.

[26] Ferreira, J.A., Guedes Soares, C., (2002). Modelling bivariate distributions of significant wave height and mean wave period,*Applied Ocean Research*, **24**(1), 31-45.

[27] Fouques, S., Myrhaug, D, Nielsen, F., (2004). Seasonal Modeling of Multivariate Distributions of Metocean Parameters With Application to Marine Operations. *Journal of Offshore Mechanics and Arctic Engineering* **126**(3), 202-212.

[28] Giebel, G., Landberg, L., Kariniotakis, G., Brownsword, R., (2003). State-of-the-Art on Methods and Software Tools for Short-Term Prediction of Wind Energy Production, *Proc. of the European Wind Energy Conference & Exhibition EWEC 2003, Madrid, Spain, June 16-19.*

[29] Gioffre M., Gusella V., Griogriu M. (2000). Simulation on non-Gaussian field applied to wind pressure fluctuations. *Probabilistic Engineering Mechanics*, **15**, 339-345.

[30] Grigoriu, M., (1995). *Applied Non gaussian Processes*, Prentice-Hall.

[31] Guedes Soares, C., Ferreira, A.M., (1996). Representation of non-stationary time series of significant wave height with autoregressive models, *Prob. Engng. Mech.*, **11**, 139-148.

[32] Guedes Soares, C., Cunha, C., (2000). Bivariate autoregressive models for time series of significant wave height and mean period, *Coastal. Engng.*, **40**, 297-311.

[33] Gurney K., (1997). *An Introduction to Neural Networks*, UCL Press.

[34] Härdle, W., Horowitz, J., Kreiss, J.P., (2003). Bootstrap methods for time series. *Int. Stat. Review* **71**(2). 435-459

[35] Ho, P.C., Yim, J.Y., (2005). A study of the data transferability between two wave-measuring stations. *Coastal Engng, In press.*

[36] Hogben, N., Standing, R.G., (1987). A Method for Synthetising Time History Data from Persistence Statistics and its use in Operational Modelling. *Underwater Technology, Society for Underwater Technology*, **13**(4), 11-18.

[37] Holzmann, H., Munk, A., Suster, M., Zucchini, W. (2005). Hidden Markov models for circular and linear-circular time series. *Environmental and Ecological Statistics.* To appear

[38] Huang, Z., Chalabi, Z.S., (1995). Use of time series analysis to model and forecast wind speed. *J. Wind Eng. Ind. Aerodyn.*, **56**, 311-322.

[39] IAHR, (1989). List of sea state parameters, *J. Waterway Port Coast. Ocean. Eng.*, **115** (6), 793-808.

[40] Izquierdo, P., Guedes Soares, C., (2005). Analysis of sea waves and wind from X-band radar. *Ocean Engineering*, **32**(11-12), 1404-1419.

[41] Jenkins, A.D., (2002). Wave duration/persistence statistics, recording interval, and fractal dimension. *International Journal of Offshore and Polar Engineering*, **12**(2), 109-113. 2002

[42] Lahiri, S.N. (2003). *Resampling Methods for Dependent data*, Springer, New York.

[43] Lall, U., Sharma, A., (1996). A nearest neighbor bootstrap for resampling hydrologic time series. *Water Ressources Res.*, **32**, 679-693.

[44] MacDonald, I. L., Zucchini W., (1997). *Hidden Markov and Other Models for Discrete-Valued Time Series*. New York: Chapman and Hall.

[45] Makarynskyy, O., Pires-Silva, D., Makarynska, D., Ventura-Soares, C., (2004). Artificial neural networks in wave prediction at the west coast of Portugal, *Computer & Geosciences, In press*

[46] Marteau, P.F., Monbet, V., Ailliot, P., (2004a). Non parametric modeling of cyclo-stationary markovian processes - part 2. *Proc. ISOPE Conf.*, **III**, 145-151.

[47] Marteau, P.F., Monbet, V., (2004b). Conditional prediction of Markov processes using non parametric Viterbi algorithm - Comparison with MLP and GRNNmodels, *WSEAS Trans. on systems*, **3**(2), 346-351.

[48] Medina, J.R., Gimenez, M.H., Hudspeth, R.T., (1991). A wave climate simulator. *Proc. 24th Int. Ass. Hydrolic Res. Congress*, 521-528.

[49] Monbet, V., Prevosto, M., (2001a). Bivariate Simulation of Non Stationary and Non Gaussian Observed Processes . Application to Sea State Parameters. *Applied Ocean Research*, **23**, 139-145.

[50] Monbet V., Marteau P.F., (2001b). Continuous Space Discrete Time Markov Models for Multivariate Sea State Parameter Processes. *Proc. ISOPE Conf.*, **III**, 10-14.

[51] Monbet V., Marteau P.F., (2004). Non parametric modeling of cyclo-stationary markovian processes. *Proc. ISOPE Conf.*, **III**, 138-144.

[52] Monbet V., Marteau P.F., (2005a). The Local Grid Bootstrap for Stationary Multivariate Markov Processes, *J. Statistical Planning and Inference, in press.*

[53] Monbet V., Ailliot P., (2005b). L^1-convergence of smoothing densities in non parametric state space models *submitted to Stat. Inf. for Stoch. Proc.*

[54] More, A., Deo, M.C., (2003). Forecasting wind with neural networks. *Marine structures*, **16** , 35-49.

[55] Nfaoui, H., Buret, J., Sayigh, A.A., (1996). Stochastic simulation of hourly average wind speed sequences in Tangiers (Morocco). *Solar Energy*, **56**(3), 301-314.

[56] O'Carroll, (1984). Weather Modelling for Offshore Operations. *The Statistician*, **33**, pp 161-169.

[57] Pittalis, S., Bruschi, A., Puca, S., Tirozzi, B., (2003). Reconstruction of sea events and extreme value analysis. *Proc. ISOPE Conf.*, **III**.

[58] Poggi, P., Muselli, M., Notton, G., Cristofari, C., Louche, A., (2003). Forcasting and simulating wind speed in Corsica by using an autoregressive model. *Energy conversion and management*, **44**, 3177-3196.

[59] Popescu, R., Deodatis, G., Prevost, J.H. (1998). Simulation of homogeneous nonGaussian stochastic vector fields. *Prob. Engng. Mech*, **13**(1), 1-13.

[60] Raftery, A.E., (1985). A model for higher-order Markov chains. *J. Roy. Statist. Soc. Ser. B*, **47**, 528-539.

[61] Raftery, A.E., Tavare, S. (1994). Estimation and modelling repeated patterns in high-order Markov chains with the mixture transition distribution (MTD) model. *J. Roy. Statist. Soc. Ser. C - Applied Statistics*, **43**, 179-200.

[62] Rychlik, I., Johannesson, P., Leadbetter, M.R., (1997). Modelling an dstatistical analysis of ocean-wave data using transformed Gaussian processes. *Marine Structures*. **10**, 13-47.

[63] Scotto, M.G., Guedes Soares, C., (2000). Modelling the long-term time series of significant wave height with non linear threshold models. *Coastal Engng*, **40**, 313-327.

[64] Shi S.G., (1991). Local Bootstrap. *Ann. Institut. Statist. Math.*, **43**, 667-676.

[65] Soukissian, T., Theochari, Z., (2001). Joint occurrence of sea states and associated durations, *Proc. ISOPE Conf.*, **III**, 33-39.

[66] Stefanakos, C.N., (1999). *Nonstationary stochastic modelling of time series with applications to environmental data*. PhD Thesis. NTAU.

[67] Stefanakos, C.N., Athanassoulis, G.A., (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.

[68] Stefanakos, C.N., Athanassoulis, G.A., Barstow, S.F., (2002), Multiscale Time Series Modelling of Significant Wave Height. *Proc. ISOPE Conf.*, vol. **III**, pp 66-73.

[69] Stefanakos, C.N., Belibassakis, K.A., (2005). Nonstationary Stochastic Modelling Of Multivariate Long-term Wind And Wave Data, *24th Int. Conf. on Offshore Mechanics and Arctic Engineering, OMAE'2005*.

[70] Stephos, A., (2000). A comparison of various forecasting techniques applied to mean hourly wind time series. *Renewable Energy*, **21**, 23-35.

[71] Toll, R.S.J., (1997). Autoregressive conditional heteroscedasticity in dayly wind speed measurements, *Theor. Appl. Climatol.*, **56**, 113-122.

[72] Tsai, C.P., Lin, C., Shen, J.N., (2002). Neural network for wave forcasting amoung multi-stations, *Ocean Engineering*, 29, 1683-1695.

[73] Turlach, B.A., (1993). Bandwidth selection in kernel density estimation: a review. *Techn. report*.

[74] Vanhoff, B., Elgar, S., (1997). Numericallu simulation non-Gaussian sea surfaces, *J. of Waterway, Port, Coastal and Ocean Engng*, **124**, 68-72.

[75] Vik, I. (1981). *The tile series generating module-Modelling aspects* Ocean Research- operational criteria, CNRD 3-2 task 2.

[76] Waeles, B., Le Hir P., Silva Jacinto R. (2004). Modélisation morphodynamique cross-shore d'un estran vaseux. *C. R. Geoscience*, **336**, 1025-1033.

[77] Walton, T.L., Borgman, L.E. (1990). Simulation of non-stationary, non-gaussian water levels on the great lakes, *J. of Waterways, Ports, Coastal and Ocean Division, ASCE*, **116**(6).

[78] Wang, D.W., Hwang P. (2001). An operational method for separating wind sea and swell from ocean wave spectra. *J. Atmospheric and Oceanic Technology.* 18(12), 2052-2062.

[79] Woolf, D.K., Challenor, P.G., (2002). Statistical comparison of satellite and model waves climatologies, *Proc. 4th symp. WAVES*

[80] Yim, J.Z., Chou, C., Ho, P., (2002). A study on simulating the time series of significant wave height near the keelung harbor. *Proc. ISOPE Conf.*, vol III.

[81] Young, P.C., NG, C.N., Lane, K., Parker, D., (1991). Recursive forcasting, smoothing and seasonal adjustement of non-stationary environmental data, *J. Forecast*, **10**, 58-89.

[82] Young, K.C., (1994). A multivariate chain model for simulating climatic parameters from daily data, *J. Appl. Meteorol.*, 33, 661-671.