International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 09, September-2015

An Expert System for Modelling Wave - Height Time - Series

Rodolfo Piscopia
Freelance
Rome, Italy

***Abstract*— This paper describes an expert system designed for the analysis of an incomplete, non-stationary and non-Gaussian, long-term, time series of wave significant heights by means of specific linear parametric model. Using this system makes it possible to complete missing-value gaps, forecast wave-height short-term evolution or simulate arbitrarily long sequences of wave data preserving the key statistical properties of the observed series, including autocorrelation, persistence over threshold, non-Gaussian distribution and seasonality.**

The implemented improvements bear on specific key tasks of ARMA setup procedure, i.e. preliminary analysis, parameter estimation and optimal model-configuration identification. Specifically, a Seasonal Trend decomposition based on Loess robust method is applied to compute more stable and detailed seasonal trend, allowing assuming more confidently its deterministic nature. Moreover, aiming at accurately estimating the model parameters, a proficient method is taken in, which is based on the robust Whittle's approximation of the maximum log-likelihood function as well as on the direct-search, non-linear, multi-parameter, constrained, optimization technique called complex modified. Finally, an automatic expert system is developed, able to identify, almost correctly, ARMA orders by selecting the model with the smallest residuals variance and parameter numbers.

Confident applicability of the suggested procedure is tested by means of both Monte Carlo simulations and comparisons of generated series with observed one, this latter measured offshore Alghero – Italy. Analysis of results clearly indicate that the accuracy in identifying the correct ARMA model is improved; furthermore, it is shown that the simulated time series exhibit all the primary statistical properties of the observed data, including winter and summer seasonal patterns as well as sea states sequencing, persistence and severeness.

Keywords — Wave climate; ARMA model; Wave forecast; Storm duration; Sea state persistence; Sea severeness

I. INTRODUCTION

For marine human activities and engineering applications, the understanding of sea-state sequences is important as well as the knowledge of extreme wave parameters, e.g. to evaluate a maritime traffic line efficiency, to guess a port/terminal operativeness or to assess risks of engineering processes. Actually, marine intervention and installation works involve long-lasting and complex operations. In these cases, the analysis of effects related to meteorological changes during specific operations is utmost relevant to disclose any possible critical situations and their related costs-growth. To these aims, linear models can be very helpful, being able to provide large database of information statistically equivalent to the observed one.

Additionally, recorded time series are usually incomplete due to several reasons, e.g. to instrument failures, accidental data loss or spikes rejection. Considering that the data incompleteness can seriously bias statistical inferences, makes obvious the relevance of a procedure able to recover missing values by ensuring same statistical sample properties.

Autoregressive, moving-average models (ARMA) are a specific class of the linear parametric family that, in few words, replicate time processes by combining some their outcomes with a white noise.

In ocean engineering applications [1] and [2] have used ARMA to simulate individual waves in short-term elevation record, supposed to be stationary in time. Reference [3] have used ARMA to model the non-stationary, long-term, time-series of significant wave-height, whereas [4] proposed a new methodology for the analysis, missing-value completion and simulation of an incomplete, non-stationary, time-series of wave data. Further researches were pointed at verifying data transferability between two wave-measuring stations [5].

Generally, two main problems have to be solved in order to apply ARMA models to long-term series of wave parameters. One is the presence in the series of missing-value gaps, which can sometimes be relatively long, and the other is the series non-stationarity and non-normality. Accordingly, gap filling as well as data transformation procedures are required. Furthermore, the common and challenging task of the model identification, i.e. selecting the most suitable ARMA order, has to be tackled.

Here, the work of [4] is extended by improving the following tasks: the seasonal component assessment, the model parameter estimation and the choice of the optimal ARMA configuration. Namely, a technique called STL robust (Seasonal Trend decomposition based on Loess) able to compute more accurate seasonal components is adopted [6]. Furthermore, the more robust Whittle's approximation of the maximum log-likelihood function is used to estimate ARMA coefficients, the set of which is found out by a proficient, nonlinear, constrained, multi-parameter, optimization technique. Finally, an expert system has been developed allowing automating the struggle step of model identification that, for mixed ARMA process, is quite tricky and someway affected by subjective interpretation.

Different Monte Carlo simulations have been carried out with a double purpose: the validation of the parameter estimation procedure and the verification of the automatic expert system proficiency. The obtained results have been gratifying, making possible to confidently say that the proposed enhancements are very efficient.

In the following, each step of the adopted ARMA modelling procedure is accurately described and the results of Monte Carlo simulations are illustrated. Afterwards, the application to real wave data is fully described and the comparison between two different techniques to normalize-denormalize the series is plainly outlined. Finally, the conclusions are drawn out.

II. LINEAR PARAMETRIC MODELS

Starting from the Box and Jenkins definition [7], the family of linear models has been developed with the conception of several subtypes that roughly follow a common setup procedure. Here only the ARMA model is considered.

Regarding a second-order stationary and Gaussian time series $\{z_t\}$, the autoregressive and moving average parts of an ARMA(p,q) model define z_t respectively as the combination of p previous terms of the series plus the combination of $q+1$ terms of a white noise (i.e. a stationary random process with zero mean and variance equal to σ_r^2). Introducing the back-shift operator defined as $B: B^n z_t = z_{t-n}$, an ARMA(p, q) can be written as $\phi(B) z_t = \theta(B) a_t$, being $\phi(B) = 1 - \phi_1 B - \ldots . - \phi_p B^p$ and $\theta(B) = 1 - \theta_1 B - \ldots . - \theta_q B^q$.

The standard ARMA setup procedure can be resumed as follows [7]: preliminary analysis, model identification, parameter estimation, model verification and optimal model-configuration selection. In what follows each task is accurately delineated.

III. PRELIMINARY ANALYSIS

With reference to the stationarity of significant wave-height series, it is typically not satisfied and, according to the common knowledge of the environmental process, a seasonal component is expected to exist. According with [8], a non-stationary time series z_t can be decomposed as $z_t = \tilde{z}_t + \mu_t + \sigma_t X_t$, where $\tilde{z}_t$, μ_t, σ_t are the deterministic functions, respectively, of long-term trend, seasonal mean and standard deviation. In what follows, the long-term trend is not considered.

The seasonal mean and variance can be defined as follows by introducing the Buys-Ballot double index, i.e. by re-indexing the time series z_t as [9]:

$$\left\{ z_t, 1 \le t \le N \right\} = \bigcup_{j=1}^{Y} \left\{ z_\tau^j, 1 \le \tau \le M \right\} \tag{1}$$

$$\mu_\tau = \frac{1}{K} \sum_{j=1}^{Y} z_\tau^j \ , \ 1 \le \tau \le M \tag{2}$$

$$\sigma_\tau^2 = \frac{1}{K-1} \sum_{j=1}^{Y} \left(z_\tau^j - \mu_\tau \right)^2 \ , \ 1 \le \tau \le M \tag{3}$$

where Y and M are integers respectively equal to the series length in year and the annual number of observations, $N=YM$ is the series numerosity, τ is the index within the annual cycle, K is the number of existing values per each observation index τ (if no missing values affects the wave series then clearly $K=Y$). The deseasonalized series is computed according to the following expression:

$$y_\tau^j = \left(z_\tau^j - \mu_\tau \right) / \sigma_\tau \ , \text{ with } 1 \le j \le Y \text{ and } 1 \le \tau \le M \tag{4}$$

This approach, however, produces seasonal components possibly affected by large sample variability, especially when the time series length is not enough extended or when many missing-value gaps exist. This large variability contrasts with the assumed deterministic nature of the seasonal component and cannot therefore be accepted by both physical and stochastic points of view.

Following [6], here is preferred a more robust method, derived from the STL one (Seasonal-Trend decomposition based on Loess). This technique, used for both the mean and variance seasonal components, is split into five tasks (here, only the evaluation of the mean component is illustrated as the variance computation is straightforwardly derivable).

1. Identify the seasonal mean series μ_τ^* by (2) and reduce it to zero average. Compute the new time series $X_\tau^j = \left| z_\tau^j - \mu_\tau^* \right|$.
2. Define the scaling factor $u_\tau^j = X_\tau^j / (c\xi)$, where ξ is the median value of X_τ^j and c is a constant (equal to 6 and 36 respectively for the mean and variance seasonal components).
3. Estimate the weighting factor η_τ^j, for each X_τ^j, according to:

$$\eta_\tau^j = \begin{cases} \left[1 - \left(u_\tau^j \right)^2 \right]^2 & \text{if } u_\tau^j < 1 \\ 0 & \text{if } u_\tau^j \ge 1 \end{cases} \tag{5}$$

4. Evaluate the weighted seasonal mean component, for each index τ in the annual cycle, as:

$$m_\tau^* = \sum_{j=1}^{Y} X_\tau^j \eta_\tau^j \Big/ \sum_{j=1}^{Y} \eta_\tau^j \tag{6}$$

5. Smooth m_τ^* by means of the interpolator called Loess. Specifically, considering a time window centered at τ, with amplitude W, m_τ^* is smoothed according to:

$$\hat{m}_\tau^* = \sum_{i=\tau-W/2}^{\tau+W/2} m_i^* \nu_i(\tau), \text{ with } \nu_i(\tau) = \left[1 - \left(\frac{2|\tau - i|}{W} \right)^3 \right]^2 \tag{7}$$

For completing the missing-value gaps, the following procedure has been adopted. When the gap length is very small (dealing with one or two observations), the missing values are interpolated from neighbors. Otherwise, the smoothed mean seasonal component (7) is transformed in a Fourier series. The missing values are therefore replaced by [3]:

$$\tilde{z}_\tau^j = \bar{\mu}_\tau + a \cos(\omega \tau) + b \sin(\omega \tau) \tag{8}$$

being $\bar{\mu}_\tau$ the average value of the smoothed mean seasonal component, $\omega = 2\pi/M$, a and b the Fourier coefficients given by:

$$a = \frac{2}{M} \sum_{\tau=1}^{M} \mu_\tau \cos(\omega \tau), \ b = \frac{2}{M} \sum_{\tau=1}^{M} \mu_\tau \sin(\omega \tau) \tag{9}$$

With reference to normality of the significant wave-height series, it is verified by performing a *t*-student test; if the tested hypothesis is rejected, a series transformation is applied. Two different approaches were implemented and compared: the first involves the classical Box-Cox transformation [10]; the second entails the Probability Level-Equivalence Transformation (PLET) used by [11].

Using the Box-Cox formula, the time series is transformed as:

$$\hat{z}_t(\lambda_{BC}) = \begin{cases} \left(z_t^{\lambda_{BC}} - 1\right) \big/ \lambda_{BC} & when \quad \lambda_{BC} \neq 0 \\ \log z_t & when \quad \lambda_{BC} = 0 \end{cases} \tag{10}$$

Differently, the PLET method is based on the percentile equivalence among the standardized normal distribution (Φ) and the wave-height best-fitting distribution (P_z). The standardized normal series is therefore obtained by:

$$\hat{z}_t = \Phi^{-1}\left(P_z\left(z_t\right)\right) \tag{11}$$

The inverse transformation used in the simulation task is:

$$z_t = P_z^{-1}\left(\Phi\left(\hat{z}_t\right)\right) \tag{12}$$

IV. MODEL PARAMETERS ESTIMATION

This task is here completed in two steps: the preliminary estimation and the accuracy refining. The method of moments [7] is adopted for the former, whereas the maximum log-likelihood method along with the Whittle's function approximation is implemented for the latter.

Actually, for a Gaussian stationary process, the approximated expression of log-likelihood function is [12]:

$$\tilde{L}_W(z \mid \boldsymbol{\psi}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log\left[S(\lambda, \boldsymbol{\psi})\right] d\lambda + \frac{1}{N} z^T A(\boldsymbol{\psi}) z \tag{13}$$

being $\boldsymbol{\psi}=(\phi_1,\ldots, \phi_p, \theta_1,\ldots, \theta_q, \sigma_r)$ the parameters vector to estimate, $A(\boldsymbol{\psi})$ the inverse of the process covariance-matrix and $S(\lambda, \boldsymbol{\psi})$ the parametric power spectrum. The latter is the Fourier transform of the autocorrelation function and can be seen as expressing the energy level of each periodicity composing the time series. Its expression, computed by parametric method, can be written as [13]:

$$S(\lambda;\boldsymbol{\psi}) = 2\sigma_r^2 \frac{\sum_{i=1}^{q}\left[1-\theta_i\cos(2\pi i\lambda)\right]^2 + \left[\theta_i \sin(2\pi i\lambda)\right]^2}{\sum_{j=1}^{p}\left[1-\phi_j\cos(2\pi j\lambda)\right]^2 + \left[\phi_j \sin(2\pi j\lambda)\right]^2} \tag{14}$$

Equation (13) can be rewritten in terms of the series periodogram (P), i.e. the series power-spectrum computed by Fourier method, as follows [14]:

$$\tilde{L}_W(z \mid \boldsymbol{\psi}) = \frac{1}{2\pi}\left[\int_{-\pi}^{\pi} \log\left[S(\lambda, \boldsymbol{\psi})\right] d\lambda + \int_{-\pi}^{\pi} \frac{P(\lambda)}{S(\lambda, \boldsymbol{\psi})} d\lambda\right] \tag{15}$$

The maximization of (15) is here carried out by a direct-search, non-linear, multi-parameter, constrained, optimization technique called complex modified [15] [16], which has been proved to perform very efficiently [17].

Aiming to enlighten the proficiency of the maximum log-likelihood method, two set of tests were carried out. In the first one, three different spectra were firstly defined by assigning p, q, σ_r, ϕ_j and θ_t in (14) and then randomly perturbed by adding a white noise drawn out from $U[-0.05\sigma_r; 0.05\sigma_r]$. The resulting frequency distributions were assumed as periodograms to be fitted by the complex modified method with (15) as target function. Fig. 1 shows the results along with those achieved using the spectral least-square target-function, given by [4]:

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\frac{S(\lambda;\boldsymbol{\psi}) - P(\lambda)}{P(\lambda)}\right]^2 d\lambda \tag{16}$$

The fig. 1 and the herein reported table clearly state that, even if both (15) and (16) fulfil the optimization by producing a nearly perfect fitting of ARMA spectra, (15) is more efficient to accurately estimate each parameter value, reducing the maximum relative approximation-error form 200% to 2% (for AR(4) – AR(10)) or from 20% to 3% (ARMA(10,10)).

In the second test case, a Monte Carlo simulation was carried out by modelling 500 synthetic series, generated from an ARMA(1,1) with σ_r = 1.0 and both (ϕ, θ) ranging from 0.0 to 1.0, step 0.2. For each generated series, (15) was used to estimate the ARMA parameter values. The difference between the assigned values and the finally estimated one have been represented, in fig. 2, as relative errors in a box-plot form. The relative errors obtained by using the widespread method of moments [18] along with those achieved by minimizing (16) are also reported. The illustrated results, revealing a great error variance reduction (at least halved) as well as an unbiased zero averages, confidently confirm the proficient improvement achieved by the herein implemented method.

V. VERIFICATION OF ESTIMATED MODELS AND SELECTION OF THE OPTIMAL ONE

To verify ARMA stationarity and invertibility, respectively, all roots of following (17) should lie externally to the unit circle (here IMSL® ZPORC routine is used):

$$1 - \sum_{i=1}^{p}\phi_i x^i = 0\,; \quad 1 - \sum_{i=1}^{q}\theta_i x^i = 0 \tag{17}$$

If one of the tested hypothesis is rejected the model is not considered further, otherwise a Portmanteau test is completed. Namely, if an ARMA is stationary and invertible as well as properly identified with accurately estimated parameters, the model residuals, given by

$$r_\tau = z_\tau - \left(\sum_{i=1}^{p}\phi_i z_{\tau-i} + \sum_{j=1}^{q}\theta_j a_{\tau-j}\right) \tag{18}$$

have nearly null random values.

Model = AR(4)			**Model = AR(10)**			**Model = ARMA(10.10)**					
fixed	*MLM*	*LSM*	*fixed*	*MLM*	*LSM*	*fixed*		*MLM*		*LSM*	
σ_r = 1.00	σ_r = 1.01	σ_r = 2.98	σ_r = 1.00	σ_r = 0.98	σ_r = 3.04	σ_r = 1.00		σ_r = 1.01		σ_r = 1.03	
ϕ_j	ϕ_j	ϕ_j	ϕ_j	ϕ_j	ϕ_j	ϕ_j	θ_i	ϕ_j	θ_i	ϕ_j	θ_i
0.94	0.95	2.81	-0.82	-0.78	-2.38	-0.41	0.87	-0.42	0.87	-0.47	0.84
-0.62	-0.63	-1.87	0.12	0.12	0.41	-0.15	-0.17	-0.14	-0.18	-0.13	-0.14
0.03	0.02	0.10	0.16	0.21	0.71	0.80	-0.38	0.84	-0.37	0.86	-0.33
-0.20	-0.20	-0.59	0.62	0.56	1.54	0.31	0.03	0.36	0.07	0.38	0.09
-	-	-	0.18	0.20	0.83	0.80	-0.21	0.84	-0.19	0.89	-0.19
-	-	-	0.02	0.01	0.28	0.92	0.58	0.92	0.56	1.13	0.68
-	-	-	0.75	0.72	2.06	-0.67	0.38	-0.68	0.39	-0.62	0.46
-	-	-	0.99	1.00	3.03	0.72	0.09	0.71	0.07	0.84	0.16
-	-	-	0.45	0.42	1.19	0.81	0.18	0.86	0.20	0.92	0.22
-	-	-	0.93	0.89	2.72	-0.41	-0.81	-0.44	-0.83	-0.45	-0.79

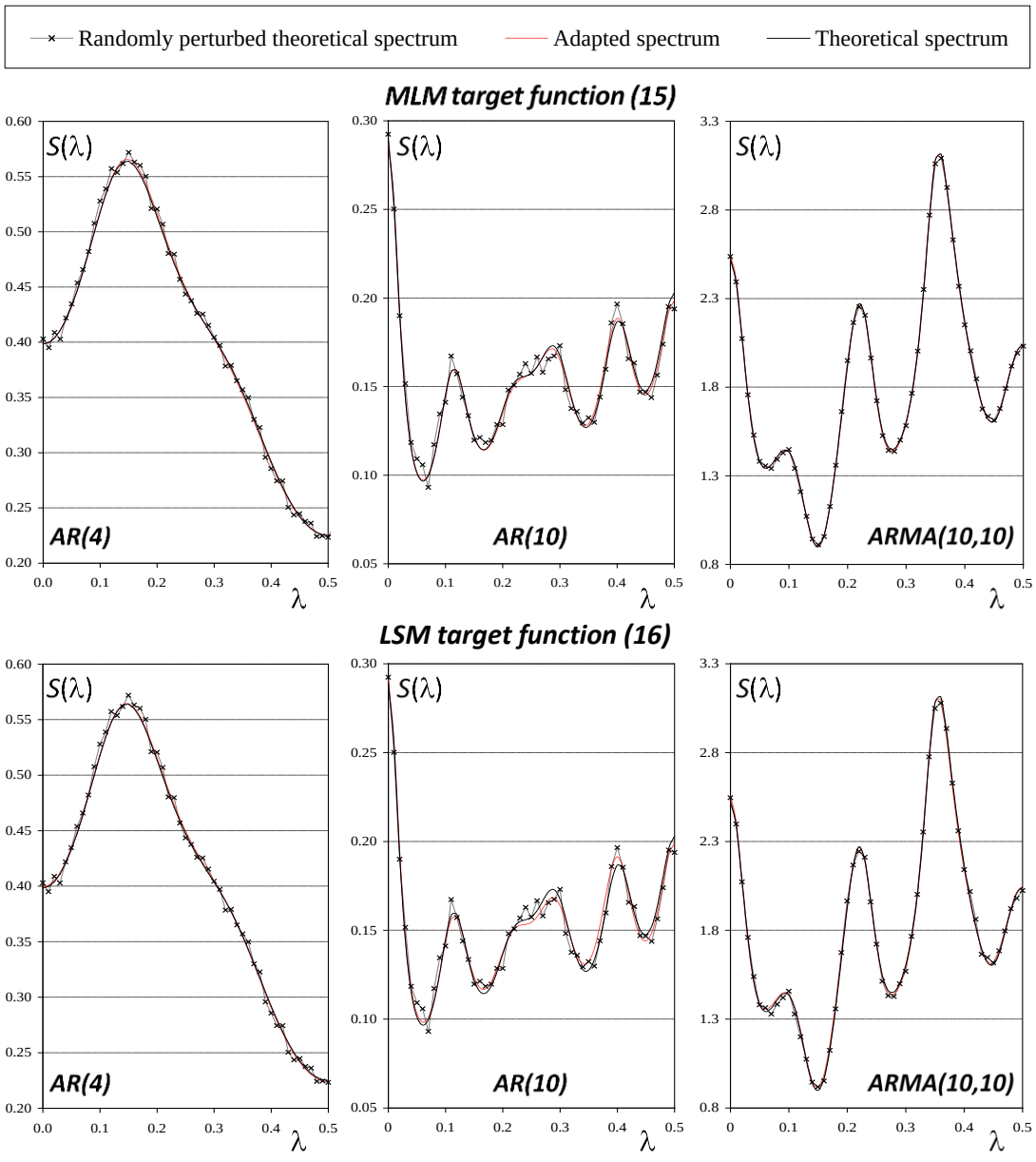

Fig. 1. Comparison between synthetics periodograms and spectral distributions obtained by the complex modified optimization technique.

Fixed parameter values

	a	b	c	d	e	f	g	h	i	j	k	l	n	o	p	q	r	s	t	u
p	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
q	1	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1
ϕ_1	#	#	#	#	0.2	0.2	0.2	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.6	0.8	.8	.8	.8
θ_1	0.2	0.4	0.6	0.8	#	0.4	0.6	0.8	#	0.2	0.6	0.8	#	0.2	0.4	0.8	#	0.2	0.4	0.6

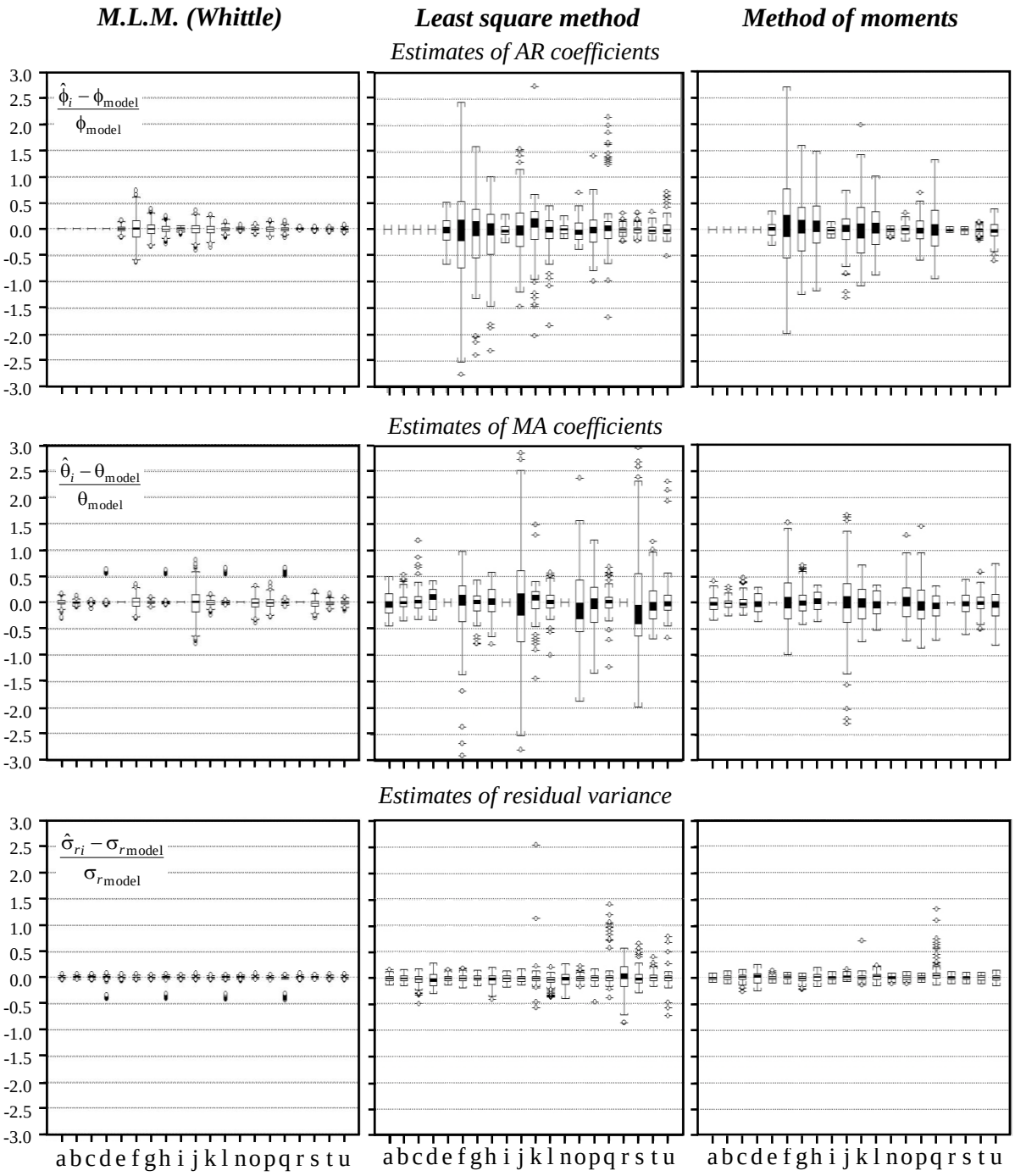

Fig. 2. Results of the Monte Carlo simulation for the ARMA(1,1) parameter estimation obtained by means of three different techniques: the methods of moments (right panels), the spectral Least Square Method (central panels) and the Whittle's approximation of the Maximum log-Likelihood Function (left panels).

To verify this hypothesis, the Ljung-Box Portmanteau-test is used [19]. A weighted sum (Q) of residual autocorrelation coefficients is computed according to the following expression:

$$Q = N(N+2)\sum_{k=1}^{s}\frac{\rho_k^2}{N-k} \qquad (19)$$

where N is the residuals number, k is the time lag, ρ_k is the autocorrelation coefficient (computed by IMSL® ACF routine) and s is the maximum lag (here s=75 is adopted). Q is then compared to the quantile of a χ^2 distribution, with s degrees of freedom, at the level of probability P. If Q is greater than χ^2 (s), the test is rejected and at least one of the examined autocorrelation coefficients is statistically different from zero, to the fixed significance level P.

If the tested ARMA is stationary, invertible, with random residuals, it will be considered for the final task of optimal model selection, i.e. the choice of model order (p, q). Aiming at automating this specific task, many methods based on some patterns of different ACF functions have been proposed [20]. Here a different approach is used.

Starting from both the definition and meaning of a linear model, the more efficient configuration can be defined as the one that outlines the process correlation structure by using the lowest parameter number and, at the same time, produces random residuals with the lowest variance. Considering that the latter generally decreases as the former increases, makes it necessary to choose the "optimal configuration" on the basis of statistical indices. In the present work, the following is taken into consideration [21]:

$$BIC = (N\text{-}p\text{-}q)\ln\left(\frac{N\hat{\sigma}_r^2}{N\text{-}p\text{-}q}\right) + (p+q)\ln\left[\frac{N(\sigma_r^2-\hat{\sigma}_r^2)}{p+q}\right] \qquad (20)$$

where $\hat{\sigma}_r^2$ is the variance of series-residuals.

The combination of p and q that minimize (20) is regarded as the ARMA "optimal configuration".

To show the consistency of the proposed procedure, two series of tests were carried out. The first one was performed analyzing the identified ranking of ten different AR, MA and mixed ARMA models of arbitrary orders. The obtained results are reported in Table I and support, although not on a statistical basis, the developed expert-system robustness. Namely, the true model order is ranked for seven times in the first two positions and it is always high-ranked. Furthermore, only one attempt gave the sum p+q of the fitted model underestimated by more than one order (settled ARMA(2,3) – selected MA(3)).

The second Monte Carlo simulation was carried out with the goal of comparing the proposed expert-system proficiency with that of the three best-performing ACF pattern-selection methods; namely, the Corner, the EACF and the SCAN methods. The efficiency of these latter methods was found out from [20].

The ARMA(2,1) $(1-0.8B)(1-0.5B)z_t = (1+0.5B)a_t$ with $\sigma_r^2 = 1.0$, was used for simulating 1000 series of 1000 terms, which were successively analyzed.

TABLE I. RESULTS OF THE "OPTIMAL MODEL IDENTIFICATION" TEST FOR TEN DIFFERENT AR, MA AND ARMA MODELS OF ARBITRARY ORDER.

FIXED		*Identified*		***Fixed ranking***
p	*q*	*p*	*q*	BIC
1	1	1	1	**1st**
1	2	1	2	**1st**
1	3	1	3	**1st**
2	1	1	1	3rd
2	3	0	3	4th
3	1	3	0	*2nd*
3	2	3	2	**1st**
5	7	4	8	*2nd*
6	4	6	5	6th
20	2	18	20	*2nd*

The occurrence of the identified combinations (p, q) obtained by the different methods are summarized in Table II. All the methods fairly spread out the identified configurations but, in this specific case, any of them significantly underestimate the model total order (p+q).

The expert system performs better than the corner method. Namely, the former achieves nearly equal results in selecting the true model configuration (scoring just a 1% less than the latter), but it shows greater sensitivity in both recognizing the minimum model orders and identifying the correct influence of the AR and MA model parts. Actually, when one of the identified model order is equal or greater than the fixed one, the expert system 35% of times overestimates the other one whereas the corner method underestimates it 49% of times.

Moreover, the expert system 21% of times bias the autoregressive character of the process with the MA one whereas the corner method makes the same error for the 41% of times. The ESACF and SCAN methods have instead a worse hit percentage for the correct model identification and have the same biasing character of the Corner one.

On these bases, it could be stated that the implemented expert system works nicely well, slightly better than the best performing pattern selection method here considered. Moreover, it has to be highlighted that the expert system automatically provides in output a list of ranked models, opening to chances of trying different configurations having similar statistical index values.

TABLE II. NUMBER OF IDENTIFIED COMBINATIONS (p, q) FOR THE 1000 SERIES GENERATED FROM AN ARMA (2,1).

	Expert system		***Corner method***		***ESACF method***		***SCAN method***	
(1,0)	-	-	-	-	-	-	-	-
(0,1)	-	-	-	-	-	-	-	-
(2,0)	3	0%	-	-	-	-	88	9%
(0,2)	-	-	-	-	-	-	-	-
(3,0)	59	6%	86	9%	-	-	309	31%
(0,3)	-	-	-	-	-	-	-	-
(1,1)	-	-	-	-	-	-	-	-
(2,1)	**472**	**47%**	**483**	**48%**	**339**	**34%**	**206**	**21%**
(3,1)	80	8%	7	1%	30	3%	6	1%
(1,2)	12	1%	155	15%	110	11%	195	20%
(2,2)	108	11%	7	1%	92	9%	10	1%
(3,2)	22	2%	2	0%	59	6%	4	0%
(1,3)	45	4%	253	25%	303	30%	179	18%
(2,3)	156	16%	7	1%	66	7%	2	0%
(3,3)	42	4%	-	-	-	-	-	-

TABLE III. RATES OF UNDER-SPECIFICATION OF THE TOTAL ARMA ORDER.

ω	0	3	5	7	9	∞
Expert system	0%	32%	40%	59%	62%	100%
Corner method	0%	0%	2%	47%	49%	100%
ESACF method	0%	0%	0%	7%	3%	87%
SCAN method	9%	9%	9%	46%	49%	100%

Finally, the influence of outlier occurrences on the methods performance has been analyzed. Accordingly, a new Monte Carlo simulation similar to the previous one was carried out; the same number of series, with equal numerosity, were generated and some spiked values were introduced at fixed time lags (i_1=0.25N, i_2=0.5N, i_3=0.75N). The Monte Carlo simulation was replicated five times, varying the outlier magnitude $\omega\sigma_r^2$. The obtained results were analyzed to compute the total order underestimation percentage. The results reported in Table III show that the expert system is the more affected one, as could be expected, inasmuch as it is based on indices being functions of the series and residual variances.

VI. SERIES SIMULATION

An infinite moving average representation [22] is here adopted with 100 terms. The implemented procedure has been compared against the widely used S*plus*® software and results have shown the statistical equivalence of generated series.

With reference to the series inverse transformation, the following is noteworthy; when the hydrologist preferred Box-Cox transformation is considered, depending on λ_{BC}, the generated series can contain data with no physical correspondence. Actually, if λ_{BC}=0, an exponential inverse transformation is required, making the simulated peak values significantly increase and so possibly involving maximum wave-height of 50m, value that actually have never been observed in the Central Mediterranean Sea [24] or by any ocean buoy all over the world [23]. Moreover, if $\lambda_{BC} \neq 0$, the transformed series could have negative values, again with no physical sense. To overcome in some extends these problems, a method to remove negative value is applied retaining the original occurrence of calms (generally defined as sea states having $H_{mo} \leq 0.20m$); namely, a constant quantity is added to the generated series so that the original number of calms N_c is equal to the number of series elements $x_t \leq 0.2$. Afterwards, the loess-smoothing procedure is used to trim off the peak values. These shortcomings are eliminated by using PLET.

VII. APPLICATION TO OBSERVED WAVE DATA

The analysed time series of significant wave-height was recorded by the RON directional wave-buoy located one mile offshore the Alghero coast – Sardinia (Italy), at a depth of about 100m. The analysed wave record was observed from 1 July 1989 to 31 December 2000, with a time interval of three hours, resulting in a series numerosity of 33615 observations.

For a deeper description of both the Italian Data Buoy Network (RON), managed by ISPRA – Oceanographical Service, and the measured data see respectively [25] and [26].

The missing values are 1224, equal to 3.6% of the data. The frequency distribution of the gap-length showed that almost all gaps cover less than one day and that many of them (equal to the 75% of gaps) can be simply recovered by neighbouring interpolation. The remaining 80% of missing data have been recovered by (8) with $\bar{\mu}_\tau$ = -0.169, a = -0.190, b = -0.043

The time series shows no significant daily non-stationarity but there is a clear seasonal component (fig. 3) having periodicity of about three months (2200 hours), which were identified and removed according to the herein illustrated procedure. Fig. 4 shows that the seasonal component is markedly affecting the mean value of the observed time series, whereas its standard deviation is nearly invariant within the averaged year. It is noteworthy that the application of the STL robust method give seasonal components much more stable than those computed by the classical averaging method as well as more detailed than those estimated by the Fourier representation (fig. 4).

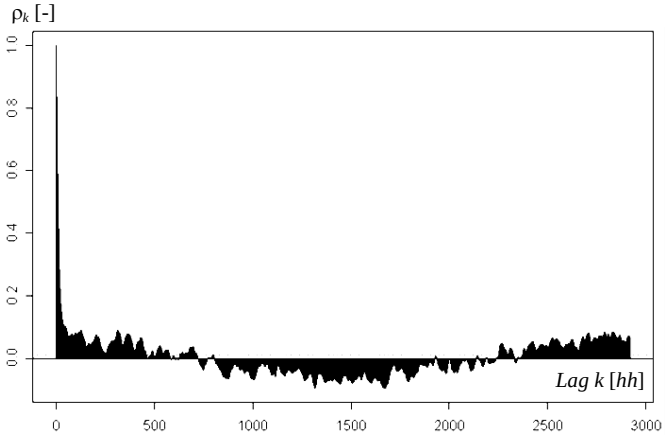

Fig. 3. Autocorrelation function of the observed time series at Alghero.

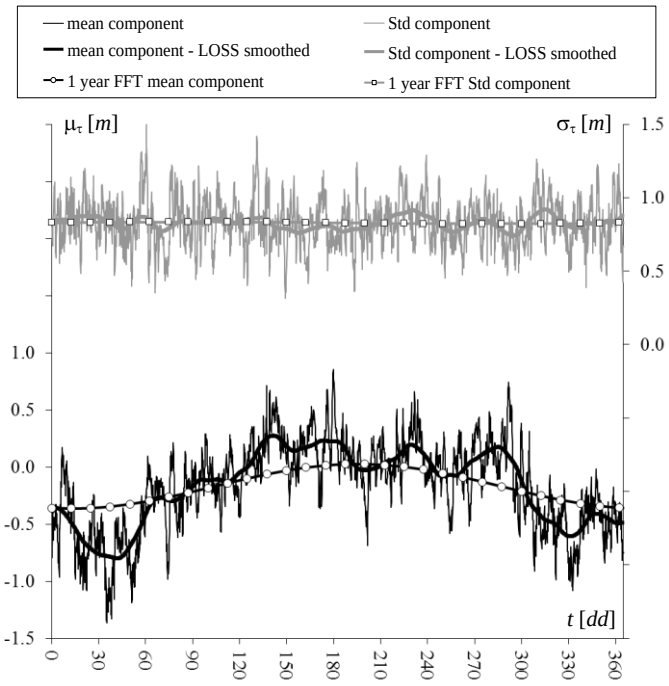

Fig. 4. Seasonal components (mean and standard deviation) of the observed time series at Alghero estimated by the classical hydrological method, by low order Fourier transform method and by the STL robust method (here, the window amplitude used by the loess smoothing is equal to 240 steps, which is nearly equal to one month). Time scale ranges from July to June.

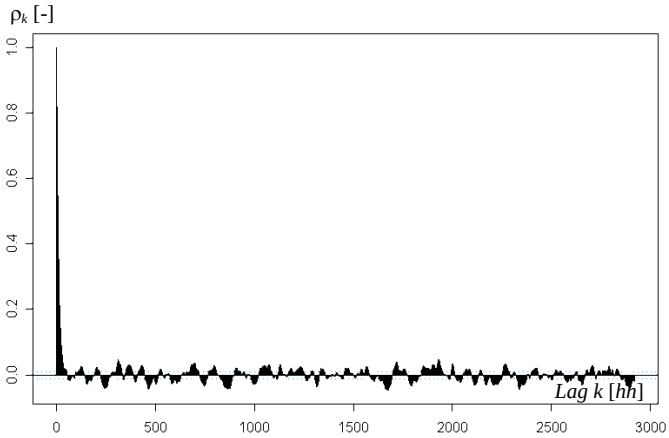

Fig. 5. Autocorrelation function of the detrended, deseasonalized series at Alghero.

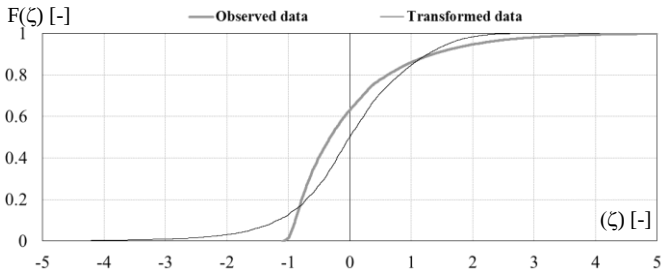

Fig. 6. Cumulate frequency distributions of the observed time series as well as those (overlapped) of the standardized series by both Box-Cox and Probability Level-Equivalence transformations, as a function of the standard variable (ζ).

Fig. 5 shows the autocorrelation of the detrended time series giving clear evidence of the performance of the applied method. Fig. 5 also shows a small residual fluctuation of the autocorrelation with a periodicity of about half a week. No physical causality could be disclosed in this cycling; accordingly, it has been considered tolerable and no more effort to remove it from the detrended series has been done.

The time series turned to be non-Gaussian as well, and both (10) and (11) have been applied to recover the time series normality; the obtained results are shown in fig. 6. Both transformations have performed accurately and efficiently by turning the non-Gaussian time series into a Gaussian distributed one; nonetheless, the inverse Box-Cox transformation has showed to be failing by recovering all the storm parametric characteristics when the reverse task of generation is considered. Namely, several attempts were carried out by varying the Box-Cox exponent value λ_{BC} in the range between 0 and 1, with a step of 0.01, as well as the width of the loess-smoothing window (see fig. 7 where some of the achieved results are reported). Unfortunately, none of the tested combinations gave fully satisfying results by obtaining a simultaneous reasonable agreement between non-exceedance cumulate frequency distributions of both storm duration and its peak-value wave-height. Actually, fixing λ_{BC} = 0.0 gives a good agreement between the duration cumulate frequencies but produces too many sea state with unreal giant waves. Conversely, setting λ_{BC} = 0.5 gives a fairly good agreement between the storm peak-value wave-height distributions but produces overestimated durations, with sea storminess greatly increased; actually, sea states with H_{mo} over the 5*m* threshold persist 40% more than the observed one.

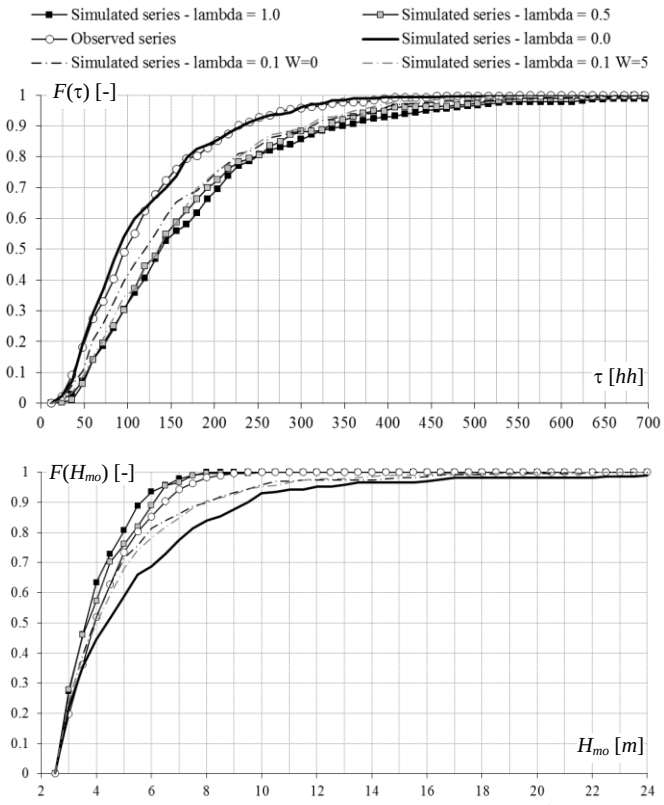

Fig. 7. Comparison between the non-exceedance cumulate frequency distributions of the storm duration (up) and its peak-value wave-height (down), computed from the observed and simulated time series using the Box-Cox transformation, with different λ_{BC} and different width of loess window.

With the aim to improve the generation task, the probability equivalence transformation was adopted. Accordingly, a probability law should be chosen to properly model the significant wave-height distribution. To this aim, several different probability laws were considered in literature, mainly focused on the distribution upper-tails (see, among others [18], [27], [28]). Thought the efforts made, the achieved results do not give any clear evidence of a true distribution [29]. Generally, it is considered that GEV type III describes better the upper tail, at the cost of larger deviations for small H_{mo} values, while the log-normal distribution fits better the distribution mode. Thus, GEV seems more appropriate for extreme-value analysis (see [26]), while the lognormal distribution seems more suitable for moderate-value analysis (e.g. fatigue-life analysis, estimation of the wave-energy resource, operativeness analysis, etc.).

Here, several probability functions were tested and fitted to the observed data by using the complex modified method to minimize the overall least square error. The goodness of fit was verified by using the Kolmogorov-Smirnov test. The distribution functions having higher K-S confidence level were: GEV type III, three parameters lognormal, four parameters power-lognormal, Beta, gamma, χ^2 and *f* (see [30] for distribution details). Each of the above functions was then used in the series simulation, but only the GEV type III and the Beta distributions gave fully satisfying results. Namely, the lognormal and power-lognormal distributions involve exponential transformation and therefore present drawbacks

similar to the Box-Cox transformation with too many sea state with unreal giant waves.

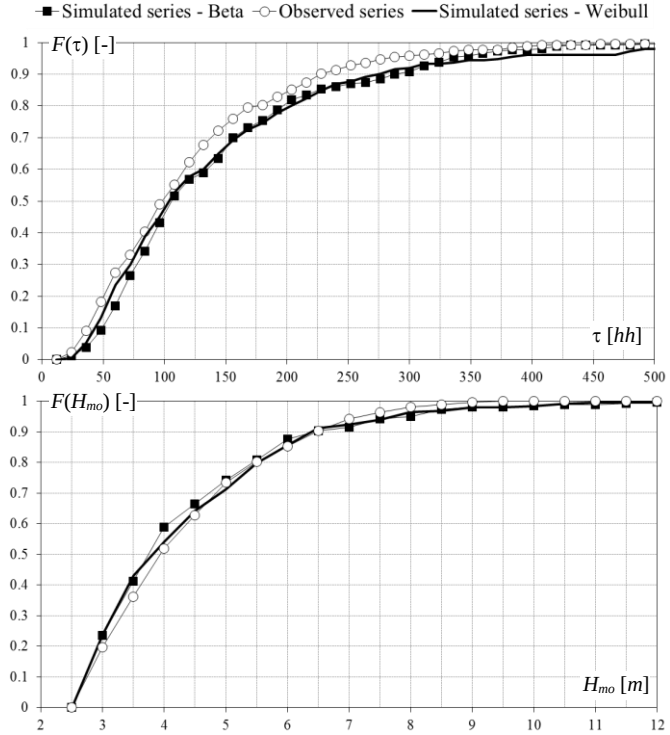

Fig. 8. Comparison between the non-exceedance cumulate frequency distributions of storm duration (up) and its peak-value wave-height (down), computed from the observed and simulated time series using PLET with Beta and GEV III wave-height theoretical probability distributions.

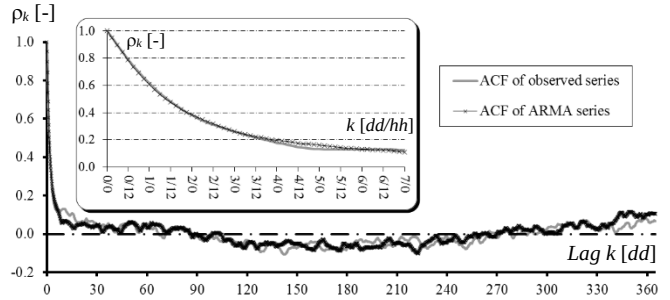

Fig. 9. Comparison between the non-exceedance cumulate frequency distributions of storm duration (up) and its peak-value wave-height (down), computed from the observed and simulated time series using PLET with Beta and GEV III wave-height theoretical probability distributions.

Conversely, the gamma, χ^2 and f distributions gave slightly biased distribution lower tails, giving rise to storm persistence too long. Accordingly, only the results obtained implementing the GEV and Beta distributions are reported in fig. 8; the attained improvements are evident inasmuch as generated series comply both frequency distributions of storms duration and wave-height peak-value quite well. Finally, the Beta distribution was chosen for the generation task at Alghero owing to its slightly better performance in replicating the duration of the storm with extreme peak-value wave-height (upper tail of the frequency distribution). The parameters of the best fitting Beta distribution resulted A=0.0, B=96.6, k_1=1.275, k_2 =100.0.

The optimal configuration of the linear parametric model at Alghero resulted in a second order Auto Regressive one, with parameters equal to ϕ_1 =1.017, ϕ_2 =-0.068 and residual variance σ_r^2 =0.087.

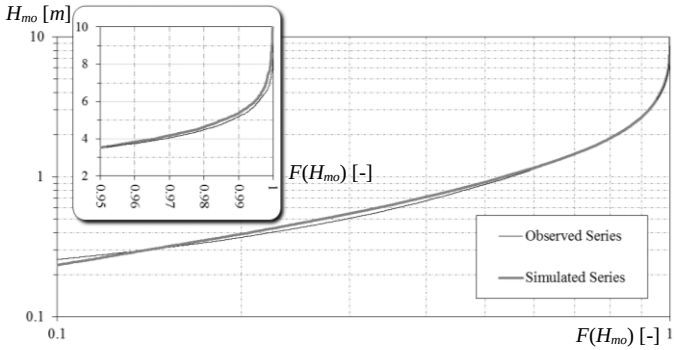

Fig. 10. Comparison between the non-exceedance cumulate frequency distributions of H_{mo} observed and simulated by the AR(2) model.

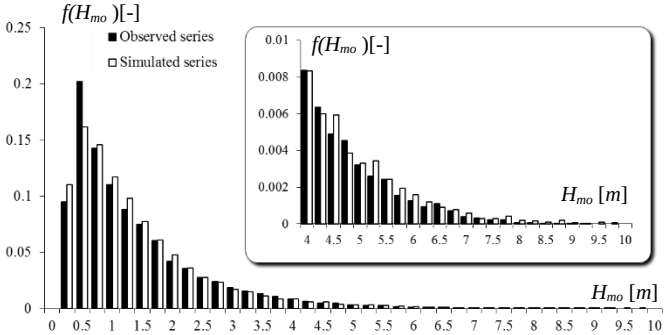

Fig. 11. Comparison between the non-exceedance cumulate frequency distributions of storm duration (up) and its peak-value wave-height (down), computed from the observed and simulated time series using PLET with Beta and GEV III wave-height theoretical probability distributions.

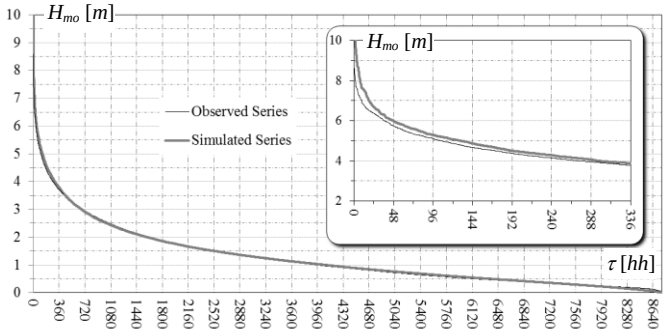

Fig. 12. Comparison between the non-exceedance cumulate frequency distributions of storm duration (up) and its peak-value wave-height (down), computed from the observed and simulated time series using PLET with Beta and GEV III wave-height theoretical probability distributions.

Fig. 9 shows the autocorrelation functions of the generated and observed series, revealing a nice matching for lags less or equal to a week as well as a nearly perfect seasonal trend.

Figs. 10 and 11 respectively show the not-exceedance and occurrence frequencies of the significant wave-height. The achieved results are quite gratifying, in that distributions are in very nice agreement starting from values greater than 0.5m, value generally assumed as the lower threshold below which data are discarded both in extreme and climatic analysis. In addition, the over-threshold persistence curves are pretty overlapped (fig. 12); it is relevant to stress that the persistence curves almost exactly overlap for $H_{mo} > 3m$, given that many

engineering activities are limited or broken down by the occurrence of such sea states.

TABLE IV. STATISTICAL SUMMARY OF BOTH THE SERIES OBSERVED AT ALGHERO AND THE AR(2) SIMULATED ONE.

	Min	1st Qu.	Mean	Mean lower conf.	Mean upper conf.	Std Err Mean	Median	3rd Qu.	Max	Variance	Std Dev.	Skewness	Kurtosis
Observed	0.06	0.40	1.23	1.22	1.24	0.01	0.90	1.60	9.10	1.21	1.10	1.81	4.09
Simulated	0.00	0.47	1.25	1.23	1.26	0.01	0.92	1.65	13.16	1.28	1.13	2.06	6.43

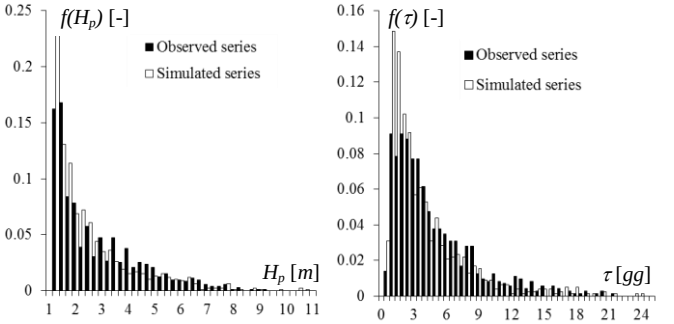

Fig. 13. Comparison between the occurrence frequencies of storm-peak significant wave-height (H_p - on the left) and storm duration (τ - on the right) for the observed time series and the simulated one.

All these fine agreements reflect themselves into the descriptive statistics of the observed and simulated series (see Table IV). The only dissonant note is the mismatch between observed and simulated data kurtosis in Table IV, which points out a greater peakedness of observed frequency mode (fig. 11).

Taking all these features in mind makes possible to say that the correlation structure of the observed data is very well reproduced into the synthetic time series.

By considering the derived dataset of the storm features, characterized by the peak-value of the significant wave-height and by its duration over the threshold of 1*m*, the agreement is slightly less gratifying but still suitable.

Namely, fig. 13 shows the occurrence frequencies of the storm duration and peak-value. The achieved results indicate that the simulated mild storms, characterized by a peak-value of the significant wave-height in the range of 1÷3*m*, are more frequent than the observed ones. On the contrary, the simulated violent storms, characterized by a peak-value of the significant wave-height in the range of 3÷5*m*, are fairly less frequent than the observed one. For the extreme storms, characterized by a peak-value of the significant wave-height greater than 5*m*, both simulated and observed ones exhibit nearly the same frequencies.

Furthermore, the simulated storms show a little bit shorter duration. Actually, results indicate that the simulated short storms, characterized by a duration less or equal to 3 days, are more frequent than the observed ones. On the contrary, the simulated persistent storms, characterized by a duration in the range of 3÷10*gg*, are fairly less frequent than the observed one. Finally, the very long storms, characterized by a duration greater than 10*gg*, show nearly the same frequencies for both the simulated and observed series.

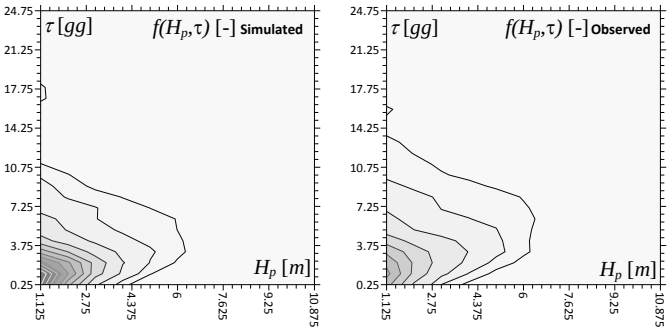

Fig. 14. Comparison between bivariate occurrence frequencies of storm-peak significant wave-height (H_p) and storm duration (τ) for the observed series (on the right) and the simulated one (on the left).

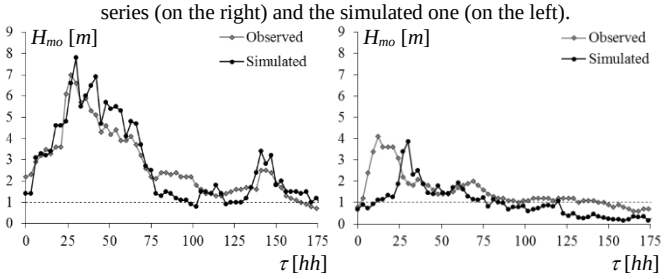

Fig. 15. Evolutions of real and simulated sample storms for severe (left) and mild (right) sea-state conditions.

These trends reflect themselves into the bivariate density of occurrence frequency for the simulated series (fig. 14), which appears more peaked near the origin. Fig. 14 states also that the simulated rare storm events (small occurrence frequency) last shorter than the observed one, independently from the peak wave-height.

The described deviations can be explained considering the evolution of sample storms for both the observed and generates series, in mild and severe conditions, reported in fig. 15; as shown, the storm decreasing-tails decay slightly faster for the generated events than for the observed ones. Moreover, the decay stops when generated wave-heights became very small while the observed ones show a more persistent behaviour; accordingly, a possible successive event reaches faster the sea-storm censoring-threshold of 1*m*. These elements concur to give a slightly greater duration of the observed storms.

VIII. CONCLUSION

This paper describes an improved methodology for the analysis, missing-value completion and simulation of an incomplete, non-stationary and non-Gaussian time series of wave significant height. The method analyses a finite-length time series to identify an ARMA model, which can be used to recover missing values, forecast short-term wave evolution or to generate arbitrarily long sequences of wave data, preserving the primary statistical properties of original dataset, including persistence over threshold, autocorrelation, non-Gaussian distribution and seasonality.

Three main improvements to the general ARMA fitting procedure are introduced: a robust estimation of the seasonal components, and accurate method to compute model parameters along with an automatic expert system for the model optimal-configuration selection. These implementations

are effectively verified to be fine improvements. Namely, STL method computes very regular as well as detailed seasonal components; in addition, the Whittle's approximation coupled with the complex-modified optimization-procedure give parameter-estimates with lower variance and unbiased mean when different series drawn out from the same process are analysed. Finally, if the observed series is unaffected by significant outliers, the expert system is able to automatically and properly identify the true model with rates very close to 50% and, generally, slightly overestimating the model total order and correctly identifying the right prevalence of the AR and MA model parts.

The proficiency of the herein proposed methodology is demonstrated in this paper through comparisons of simulated series with data observed by the directional buoy of RON, located offshore Alghero coast – Sardinia, Italy. The achieved results point out that statistical properties of the observed and simulated time series are almost nearly equivalent; this nice agreement embraces winter and summer seasonal patterns, sea state sequencing, over-threshold persistence, occurrence and cumulate frequency distribution of significant wave-height as well as both the cumulate frequency distribution of the storm duration and its wave-height peak-value.

Accordingly, the described ARMA-modelling procedure is an efficient tool in representing the wave-height climate. Its straightforward application is accordingly associated to the comprehension of sea state conditions, which is of central importance for many offshore and nearshore activities. Actually, estimates of risk for critical scenarios are often defined as some over-threshold responses of complex and interlaced systems; Monte Carlo can therefore be the only way to derive the probabilities of interest. Accordingly, even if there is a huge amount of data collected on ocean waves, which is jet geographically sparse and time limited, ARMA models can be very helpful, being able to provide large database of observed statistically-equivalent information.

Moreover, taking into consideration the modern engineering-area of wave-energy conversion, a fresh and promising application of linear model is delineated. Actually, according with [31], a real-time control of converters is required to approach the optimal efficiency of wave-energy extraction; to this aim the knowledge of future incident wave elevation is mandatory. Treating wave surface fluctuations as a time series and applying an ARMA model, makes possible to predict incoming wave elevation only from its past history. Results achieved on real data from Galway Bay and Pico Island showed the proficiency of a linear model to render a very accurate prediction of the incoming swell waves for a lag up to two wave periods. The herein presented methodology can be promptly adapted to wave elevation time series excluding the seasonal components estimation.

REFERENCES

[1] P.TD. Spanos, "ARMA algorithms for ocean modelling". Trans. ASME, J. Energy Res. Tech., vol. 105, pp. 300-309, 1983.
[2] R.J. Sobey, "Correlation between individual waves in a real sea state", Coastal Eng., vol. 27, pp. 223-242, 1996.
[3] C. Guedes Soares, A.M. Ferreira, C. Cunha, "Linear models of the time series of significant wave height in the Portuguese coast", Coastal Eng., vol. 29, pp. 149-167, 1996.
[4] C.N. Stefanakos, G.A. Athanassoulis, "A unified methodology for the analysis, completion and simulation of nonstationary time series with missing values, with application to wave data", App. Ocean Res., vol. 23, pp. 207-220, 2001.
[5] P.C. Ho, J.Z. Yim, "A study of the data transferability between two wave-measuring stations", Coastal Eng., vol. 52, pp. 313-329, 2005.
[6] S. Grimaldi, "Linear parametric models applied to daily hydrological series". Journal Of Hydrologic Engineering, vol. 9, pp. 383-391, 2004.
[7] G.E.P. Box, G.M. Jenkins, "Time Series Analysis: Forecasting and Control", Holden-Day, San Francisco, 1976.
[8] J.D. Salas, "Analysis and Modelling of Hydrologic Time Series", in Handbook of hydrology, Maidment Editor, Eds. McGraw-Hill, 1993.
[9] G.A. Athanassoulis, C.N. Stefanakos, "A nonstationary stochastic model for long-term time series of significant wave height", J. Geoph. Res., vol. 100, pp. 16149-16162, 1995.
[10] G.E.P. Box, D.R. Cox, "An analysis of transformation", J. Roy. Stat. Soc., Ser B., vol. 26, pp. 211-252, 1964.
[11] C. Cunha, C. Guedes Soares, "On the choice of data transformation for modelling time series of significant wave height", Ocean Eng., vol. 26, pp. 489-506, 1999.
[12] J. Beran, "Statistics for Long-Memory Processes", Chapman and Hall, New York, 1994.
[13] S.M. Kay, S.L. Marple, "Spectral analysis – a modern perspective", Proc. of the IEEE, vol. 69, n° 11, pp. 1380-1419, 1981.
[14] M.S. Taqqu, V. Teverovsky, "Robustness of Whittle-type estimators for time series with long-range dependence", Stochastic Models, vol. 13, n° 4, pp. 723-757, 1997.
[15] M.J. Box, "A new method of constrained optimization and a comparison with other methods". Computer Journal, vol. 8 (1), pp. 42-52, 1965.
[16] P.E. Gill, W. Murray, "Numerical Methods for Constrained Optimization". Academic Press. 1975.
[17] R. Piscopia, "On the optimal fitting of a ten-parameter model to observed wave spectra". ISOPE 13th conf., vol. III, pp. 234-240, 2003.
[18] J.A. Ferreira, C. Guedes Soares, "Modelling the long-term distribution of significant wave height with the beta and gamma models", Ocean Eng., vol. 26, nº 8, pp. 713-725, 1999.
[19] G.M. Ljung, G.E.P. Box, "On a measure of lack of fit in time series models", Biometrika, vol. 65, pp. 297-303, 1978.
[20] W.S. Chan, "A comparison of some pattern identification methods for order determination of mixed ARMA models", Statistics & Probability Letters, vol. 42, pp. 69-79, 1999.
[21] P.J. Brockwell, R.A. Davis, "Introduction to Time Series and Forecasting". Springer Verlag, New York, 1996.
[22] T.W. Anderson, "The Statistical Analysis of Time Series". John Wiley & Sons, New York, 1971.
[23] R. Piscopia, R. Inghilesi, S. Corsini, L. Franco, 2004. "Italian Wave Atlas", ed. Univ. Roma III, US Lib. Congr. G1989.21.C7, pp. 134, 2004.
[24] P. Petrova, Z. Cherneva, C. Guedes Soares, "Distribution of crest heights in sea states with abnormal waves". Appl. Ocean Res., vol. 28, pp. 235-245, 2006.
[25] M. Bencivenga, G. Nardone, F. Ruggiero, D. Calore, "The Italian data buoy network (RON)", Proc. Advances in Fluid Mechanics IX, ed. WIT, pp. 321-332, 2012.
[26] R. Piscopia, R. Inghilesi, A. Panizzo, S. Corsini, L. Franco, "Analysis of 12 - year wave measurements by the Italian wave network", Proc. 28th ICCE (Cardiff, UK), pp. 121-133, 2002.
[27] M. Isaacson, N.G. Mackenzie, "Long term distribution of ocean waves: a review", J. Waterways, Port, Coastal and Ocean Eng., vol. 10, pp. 93-109, 1981.

[28] Y. Goda, K. Konube, "Distribution function fitting for storm wave data", Proc. 22nd ICCE Conf., vol. 1, pp. 1-14, 1990.
[29] C. Guedes Soares, M. Scotto, "Modelling uncertainty in long-term predictions of significant wave height", Ocean Eng., vol. 28, pp. 329-342, 2001.
[30] M. Evans, N. Hastings, B. Peacock, "Statistical Distributions", 3rd ed., John Wiley and Sons, 2000.
[31] F. Fusco, J.V. Ringwood, "Short-term wave forecasting for real-time control of wave energy converters", Sustainable Energy, IEEE Transactions, vol. 1, n° 2, pp. 99 – 106, 2010.