

Statistical contributions to the analysis of environmental risks along the coastline

Contributi statistici all'analisi dei rischi ambientali lungo le linee costiere

Giovanna Jona Lasinio

DSPSA, Università di Roma “La Sapienza”, P.le A. Moro 5, 00185 Roma, Italy, email:
giovanna.jonalasinio@uniroma1.it

Arianna Orasi

APAT Coastal Defense Unit, Via Curtatone 3, 00185 Rome Italy email:
arianna.orasi@apat.it

Fabio Divino

Dipartimento di Scienze e Tecnologia per Ambiente e Territorio, Università del Molise,
Contrada Fonte Lappone, 86090 - Pesche (IS), Italy email: fabio.divino@unimol.it

Pier Luigi Conti

DSPSA, Università di Roma “La Sapienza”, P.le A. Moro 5, 00185 Roma, Italy, email:
pierluigi.conti@uniroma1.it

Riassunto: In questo lavoro è messa in evidenza l'importanza del contributo statistico nella definizione della climatologia costiera indispensabile al fine della comprensione dell'evoluzione dei litorali e dei processi idrodinamici che influenzano la fascia costiera. Il modello SWAN è utilizzato per la determinazione del trasporto a costa del clima ondoso; in questo lavoro ne viene presentata un'applicazione nel Mar Adriatico. Per l'implementazione di tale modello è stata costruita una procedura di post-processing che permettesse di ottenere delle condizioni al contorno corrette a partire dai dati di analisi del modello WAM del ECMWF calibrate tramite le osservazioni della rete ondometrica italiana. I risultati dell'applicazione della procedura hanno evidenziato una consistente riduzione della sottostima delle altezze significative dell'onda.

Keywords: Wave climate, Significant wave height, Calibration, Conditional simulation, Local regression

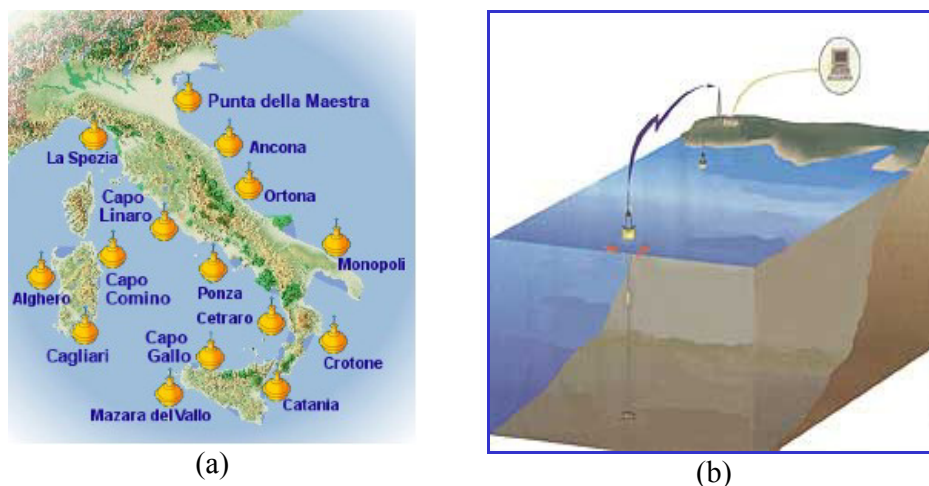
1. Introduction

Better understanding and prediction of sea state are an essential need for many application domains, such as navigation and security of ships, survey of the coastal zone, etc. These activities are among the APAT's¹ institutional tasks. These activities need a large amount of climatic and maritime transport information which is not directly available along the entire coast line. However direct observations of phenomena are seldom available. The majority of studies carried out on these subjects, are based on

¹ Agency for Environmental Protection and technical services

deterministic models output, usually climatic forecasts produced at several spatial and temporal resolution. Since 80s, deterministic models based on the dynamics and thermodynamics of the atmosphere have been used to forecast the weather with increasing reliability. Medium range forecasts are now useful up to 5 days on the average. More recently, say in the last decade, sea surface winds produced by meteorological models have been found accurate enough to be taken as a basis for operational marine forecasts, in terms of waves height generated by the wind during sea storms in the ocean. The principal provider for global numerical wave forecasts in Europe is the European Centre for Medium-Range Weather Forecasts (ECMWF), which runs as global medium range (3-5 up to 10 days forecast, 55 km spatial resolution) as high resolution short term (3 days, 28 km resolution.) models in the Mediterranean Area. Despite all efforts and improvements in the model implementation and the increase in the amount of assimilated data, still it is the statistical post-processing of the fields the key to improve, assess and verify the quality of the forecasting system beyond the deterministic limit in a particular area. Indeed, the wave height is often underestimated by numerical models. There is common agreement on the fact that wind at sea level is incorrectly predicted due to topographic effects. Since energy transfer from wind to waves is modelled by non linear parameterizations the effects are amplified in the wave forecasts (Cavaleri and Bertotti, 2004). Despite all the efforts and the improvements in the implementation of the wave model and the increase in the amount of assimilated data in the last decades, the differences between observed waves and numerical fields of analysis are still relevant in the Mediterranean basin. However a valuable source of direct measurements of significant waves height is the Italian National Sea Wave Measurement Network (RON). This is a buoys network distributed around the Italian coast (see Figure 1 (a)), measuring several quantities of interest for waves analysis: significant wave height, peak period, wave direction and others..)

Figure 1: (a) Italian buoy network (RON); (b) Buoy-receiver system



The network is made of real time sensors located in deep water (~100m depth) and local receivers as illustrated in Figure 1 (b). WAM data do not describe events near shore, but, once corrected/calibrated, they can be used to define boundary conditions for the implementation of an another class of deterministic models: the SWAN model (Simulating Waves Nearshore, Booij et al., 1999, Ris et al., 1999).

In this paper the statistical post-processing of WAM 6 hours Analysis data on the basis of RON data is presented. This study is part of an active collaboration between the DSPSA, University of Rome “La Sapienza”, and the APAT Coastal Defense Unit, involving in the research group people from diverse background (physics, engineering, mathematics and statistics). The research main target is to build appropriate boundary conditions for the SWAN model.

2. Methodology and results

The statistical methods applied to post-process the WAM Analysis data by means of the comparison with RON observations are described. The post-processing procedure has been organized in sequential steps after an accurate exploratory data analysis. Each step in the procedure deals with one single source of uncertainty, through a stochastic approach. The correction of WAM output with buoy data has three main uncertainty sources to be taken into account:

1. overall (static) variability
2. time variability
3. spatial variability

Each step in our procedure deals with one of the above mentioned points.

2.1 Exploratory Data Analysis

It is well known that data quality influences every step of any statistical procedure. When missing data are present in a time series, it is important to verify when they occur, in order to assess whether the relationship between data absence and phenomenon under study would bias the statistical results. If missing data do not appear to follow any specific design, data are said *missing at random*. In this case the only consequence on the statistical procedure is a reduction in the estimate precision. No WAM missing data were found, thus only RON buoy data have been tested for quality. The first set of RON buoys were installed in 1989. They used to transmit data every 3 hours in calm conditions, the data transmission switching to 30 minutes when the significant wave height (H_{m0}), would go over a specific threshold. The system was improved in 2002, all acquisition, transmission and storage are now at 30 minutes. In order to make the series homogeneous with respect to the analysis of high waves episodes, the two data sets have been aligned on a 30 minute basis, filling the periods of under-threshold values between the 3 hourly data with artificial missing. However episodes of lack of transmission or other causes generate real missing periods in the buoy dataset. Missing data have been classified according to *generating causes*. As the focus is mainly in sea storm and extreme events, RON buoy data have been examined for the length of the gaps and the time of occurrence. When a gap happen to appear during a sea storm it is classified as *not at random*. This procedure has been applied to all RON buoy series. The shorter gaps (less than 6 hours) have been found almost always significantly related to high waves just before or after the gap, on the other hand, gaps longer than 6 hours did not show a significant relation with high waves before or after the gap. This is probably because long gaps are often due to un-moorings caused by accidental collisions with ships, power or internal buoy failures or similar events. In the period from 1992 to 2002, most of the gaps were found greater than 6 hours and the ratio between the 6 hour gaps and the total number of gaps is around 4-5% in all the

stations considered. With a reasonable degree of confidence, this suggested that the missing data are generally not directly related to the presence of high waves, except in the few very short episodes due perhaps to transmission failures in rough sea. In the second period, from 2002 to 2003,, most of the gaps (around 80% in all the cases) were found in the short range (<6 hour). In this case a statistical test showed that the Hm0 before or after the gap was randomly distributed, in other words, there is no evidence that short gaps were related to the occurrence of rough sea. The long gaps (more than 6 hours) were not found to be related to the presence of high waves before the gap in the first as in the second period.

The second problem that has been addressed is the alignment of buoy and WAM data. WAM model outputs are on a $0.25^\circ \times 0.25^\circ$ grid, then for each buoy it is necessary to find which grid points around it can be considered. This is carried out by computing correlations between time-aligned WAM and buoy data over two different time windows (all available data and year by year) and two classes of Hm0 :

- C1: observations with $0.3\text{m} < \text{Hm0} < 2\text{m}$;
- C2: observations with $\text{Hm0} \geq 2\text{m}$.

The number of WAM's grid nodes to be considered has been chosen according to the variability of the wave direction, the more variable is this measure, the larger the number of points around a buoy considered. It has been found that for each class (C1 and C2) there were more than one grid point which was highly correlated with the buoy. As a consequence, their diversity was tested by testing the correlation coefficients differences (Cohen, Cohen 1983). As the correlation coefficients at all considered grid points didn't result to be significantly different, only the node that was found most correlated with the buoy data in both classes (C1 and C2) has been chosen. In order to find if directional intervals are characteristic to a Hm0 class, frequency distribution (histograms) have been built for each buoy and for the chosen WAM's grid points. This information has been used in the calibration procedure as described in section 2.2.

2.2 Calibration of WAM model data

In order to estimate the relationship between observed buoy data and WAM values, two calibration steps have been considered:

- step 1** application of a generalized additive model for a preliminary calibration of WAM data
- step 2** application of a dynamic regression model (estimated through Kalman Filter) in order to account for the temporal variability.

In step 1 the global variability is dealt with by implementing the following *generalized additive model* (GAM) (Hastie and Tibshirani, 1990) model:

$$Y_{buoy} = m(x_{mod}) + \sum_{j=1}^{k-1} \tau_j \delta_j + \varepsilon \quad (1)$$

where Y_{buoy} denotes observations from a single buoy, x_{mod} indicates aligned WAM model data, $m(x)$ is a smooth function, δ_j is the indicator of the observed wave direction ($j=1, \dots, k=24$) and it is such that $\sum_{j=1}^k \delta_j = 1$; τ_j are unknown coefficients quantifying the

effect of the direction, and ε is the random error term assumed to be Gaussian with a zero mean and constant variance. Equation (1) describes the relationship between the

buoy and the WAM data by means of a non parametric term $m(x)$ and a parametric term τ_j . $m(x)$ is obtained via *Local Linear Regression* (LLR) (Fan and Gijbels, 1996). The model is applied separately for the C1 and the C2 class of wave height. The result of this step is the predicted value of wave height obtained from WAM Hm0 and direction as described in equation (1). More precisely, if we denote by $\hat{m}, \hat{\tau}_j$ the estimated

elements of equation (1), the calibrated value is obtained as $\hat{Y} = \hat{m} + \sum_{j=1}^{k-1} \hat{\tau}_j \delta_j$

The procedure in step 1 tries to correct the WAM Hm0 data without considering the fact that the calibration procedure is applied on time series. To exploit the time variability a second calibration step has been introduced. Let $\tilde{X}_t$ denote the WAM corrected value at time t and Y_t denote, as above, the observed buoy value, then:

$$\begin{aligned} Y_t &= \alpha_t + \beta_t \tilde{X}_t + \varepsilon_t \\ \alpha_t &= \varphi_\alpha \alpha_{t-1} + \eta_t \\ \beta_t &= \varphi_\beta \beta_{t-1} + \xi_t \end{aligned} \quad (2)$$

In (2), the regression coefficients α_t and β_t are modelled as autoregressive processes of order 1, φ_α and φ_β are the autoregressive coefficients and ε, η, ξ are Gaussian error terms with zero mean and variances $\sigma_\varepsilon^2, \sigma_\eta^2, \sigma_\xi^2$ respectively. This type of models can be easily estimated through a Kalman Filter (KF). Final results of the calibration are the predicted values obtained from equations (2).

The calibration procedure results are shown here for a weak sea storm in the Adriatic Sea. The storm started at 6pm on December 2, 2002 and was over by 6am on December 10: 30 records have been analyzed, one every 6 hours. The two steps procedure has been applied to 4 WAM grid points near each Adriatic buoy. Even in the worst cases (highest MSE values), a considerable improvement was obtained; the calibration procedure was still able to correct WAM values in a satisfactory way when changes in Hm0 are not abrupt. In Table 1 the RMSE errors are shown.

2.3 Spatialization of calibration results

Calibration can only provide results near the buoy locations. In order to make these results significant on a larger scale, a conditional simulation model (Lantuéjoul, 2002) (MCMC based) has been implemented. Generally speaking, it is assumed that the correlation structure (spatial variation) of the wave field represented in the WAM output is correct, so it can be exploited to “spread” the calibration results to the entire field. Other approaches have been considered, such as universal kriging (Cressie, 1993) and kriging with external drift (Wackernagel, 1995). However, no one returned satisfactory results. In general terms, the wave random field Y is represented as a Gaussian field:

$$p(Y) \propto \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \gamma_{ij} (Y_i - Y_j)^2 \right\} \quad (3)$$

where Y is the vector of n values observed on a grid S , σ^2 is the marginal variance, equal for all field components, and $\{\gamma_{ij}\}$ is a set of weights controlling the intensity of the interaction between sites pairs. In our approach the field is Markovian than the

stochastic behaviour of each Y_i depends only on neighbouring values. The neighbourhood structure of the field is defined in terms of spatial proximity and it is denoted introducing ∂i for each location $i \in S$.

Table 1: Average error computed as Rooted MSE for December 2002 storm.

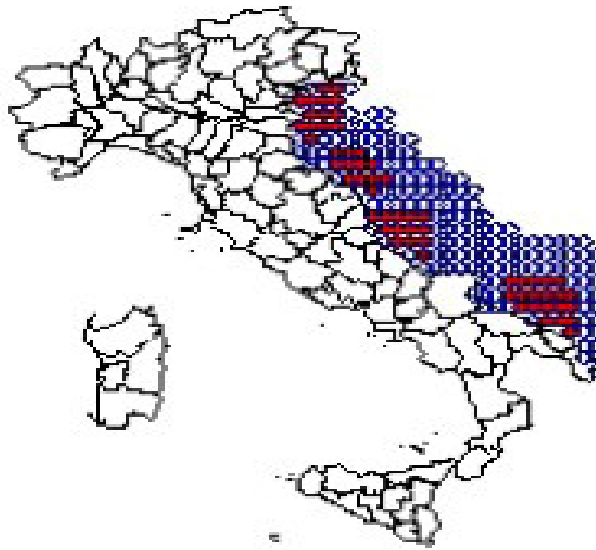
Buoy	RMSE WAM (m)	RMSE calibrated values (m)
Ancona	0.7489745	0.2264918
Chioggia	0.6708362	0.2305599
Monopoli	0.5225357	0.2070085
Ortona	0.4017648	0.1547558

With these assumptions is relatively easy to find the *local characteristic* of the field (conditional distributions $p(Y_i | Y_{\partial i}, \sigma^2) \sim N\left(\bar{Y}_i, \frac{\sigma^2}{n_i}\right)$) and then, under quite general assumptions, to find the joint distribution of the field itself. In the final implementation of our procedure we assume a weighted interactions between sites. The weights structure is based on the spatial variability of the WAM data, it depends on the distance between locations:

$$\gamma(d_{ij}) = \exp\left\{-\frac{d_{ij}^2}{r^2}\right\} \quad (4)$$

where d_{ij} is the distance between two location i, j .

Figure 2: WAM grid and the 59 calibrated values (red dots) in the Adriatic sea



The results obtained in the spatialization procedure are briefly illustrated. In the WAM grid and the 59 calibrated points available in the Adriatic are shown. It is clear that 59 locations are not enough to apply common geostatistical techniques as kriging to spread calibration results all over the area. In order to implement the conditional simulation it is necessary to specify a neighbourhood system to compute $\bar{Y}_i$ and the other quantities which depend on neighbouring sites. We chose a circular neighbourhood of radius 50km as it minimizes the MCMC standard errors estimates. The parameter r in (4) has been estimated for each simulation run (one for each time during the storm) from the WAM data. The final values of the field to be used in the SWAN implementation (for more details see paragraph 2.4) are obtained as ergodic mean of the simulated fields. We used a burn-in period of 10000 iterations and 5000 iterations to build the ergodic mean. In Figure 3 a contour map of Hm0 data at 6am of December 5, 2002. The off-shore stripe in the Adriatic Sea was used as boundary condition for the SWAN model. The dots represent the Hm0 values measured by the four RON buoys of Chioggia, Ancona, Ortona, Monopoli respectively from the top to the bottom of the pictures. The land and part of the Tyrrhenian sea are in black (we focus only on the stripe in the Adriatic sea on which the algorithm is applied) while the Adriatic Sea out of the area of interest is in white.

2.4 The implementation of the SWAN model

The proposed post-processing protocol, allows the use of corrected significant wave height fields as boundary conditions in the SWAN model. SWAN is a third generation numerical model for estimating wave parameters in coastal areas from given bottom, wind, currents and off-shore conditions. The wind field is obtained from the ECMWF meteorological analysis. The model has been at first applied on a large coarse resolution grid using corrected WAM data as boundary conditions and then on a nested higher resolution grid to obtain the details of the wave climate for the coastal area of interest. An example of SWAN results on the nested grid, where bottom induced refraction and wave breaking can be observed, is presented in Figure 4.

Figure 3: Contour map of Hm0 data at 6am of December 5, 2002. (a) Wam output, (b) Spatialization of Calibrated Wam data.

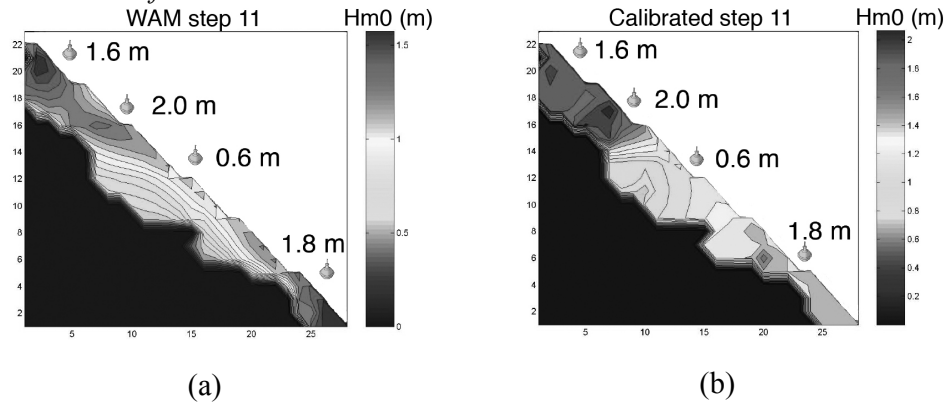
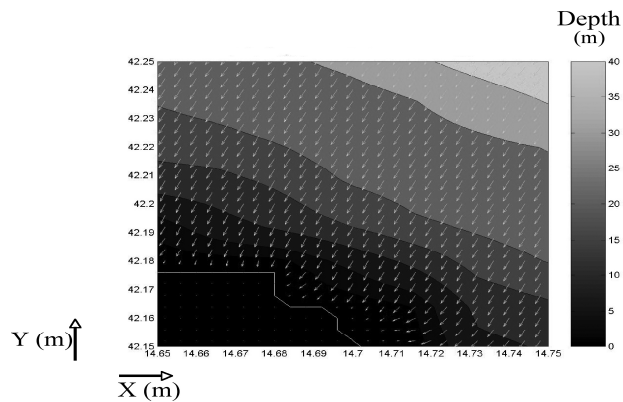


Figure 4: SWAN's result form the nested grid run. Arrows represent significant wave height and mean direction.



3. Conclusions

The use of Statistical techniques allowed to considerably improve the SWAN coastal prediction. This is a very relevant result in terms of coastal risk assessment and in the understanding of coastal dynamics. Here the contribution of statistical procedures to the integration of different data sources (such as observations and model forecasts) proved to be an essential tool in order to achieve consistent results. Further improvements of the proposed procedure are possible. For instance no clear account of the overall uncertainty is given. In order to solve this problem a Bayesian Hierarchical model is under study, in this setting being possible to account for several uncertainty sources simultaneously. However it has to be noticed that the present approach is computationally very efficient, while the estimation of a complex Bayesian model may be computationally intensive.

References

- Bruschi, A., Conti, P.L., Corsini, S., Divino, F., Ferrari, C., Inghilesi, R., Jona Lasinio, G., Leone, M., Lux, M., Morucci, S., Orasi, A. (2007) Assimilation of marine observations in a coastal modeling system. A case study: the Adriatic Sea to appear in *Bollettino di Geofisica Teorica ed Applicata*
- Cohen, J., Cohen, P., (1983) *Applied multiple regression/correlation analysis for the behavioral sciences*, 2nd ed. . Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.
- Cressie, N., (1993) *Statistical Analysis for spatial data*, Wiley
- Fan, J., Gijbels I., (1996) *Local Polinomial modelling and ist applications.i CRC Press.*
- Lantuéjoul, C., (2002) *Geostatistical Simulation: Models and Algorithms*, Springer, Berlin.
- Wackernagel, H., (1995) *Multivariate Geostatistics*. Springer-Verlag.