EPiC Series in Engineering
Volume 3, 2018, Pages 1656–1663
HIC 2018. 13th International Conference on Hydroinformatics

A wave measurements HF radar data set in the Malta-Sicily channel: data quality, validation and gap filling

Marco Picone[1*], Arianna Orasi[1], Aldo Drago[2], Fulvio Capodici[3], Giuseppe Ciraolo[3], Gabriele Nardone[1], Joel Azzopardi[2], and Adam Gauci[2] and Anthony Galea[2].

[1] ISPRA, CN-COS, via Vitaliano Brancati 60, Rome 00144, Italy
[2] University of Malta, Physical Oceanography Research Group, Dept. of Geosciences, Msida MSD 2080, Malta
[3] University of Palermo, DICAM, Viale delle Scienze Ed. 8, Palermo 90128, Italy

Abstract

The CALYPSO HF radar network is a permanent and fully operational observing system currently composed of four CODAR HF stations. The system is providing real-time hourly maps of sea surface currents and wave data in the Malta-Sicily Channel since 2012. Significant wave height derived from the HF radar wave measurements are confirmed to be a reliable source of wave information even in case of extreme events. However, it is noticed that the HF radar wave data are subject to differing interfering noise in the signal from unknown sources that may be competing with transmissions in the same frequency band. These interferences lead to frequent gaps and/or outliers that affect the continuity and reliability of the data set. The aim of this work is to estimate missing values and to detect possible outliers building and fitting a Markov chain mixture model on the significant wave height data collected at the four stations. It is verified that the proposed procedure is sufficiently robust since the model estimates succeed to classify radar observations with a high percentage of missing data and to equally highlight spikes and outliers.

* Corresponding author: marco.picone@isprambiente.it

G. La Loggia, G. Freni, V. Puleo and M. De Marchis (eds.), HIC 2018 (EPiC Series in Engineering, vol. 3), pp. 1656–1663

1 Introduction

With the partial financing by the EU under the Operational Programme Italia-Malta 2007-2013, the CALYPSO project and its Follow-On activity have delivered a permanent and fully operational HF radar observing network capable of recording (in quasi real-time with hourly updates) surface currents and wave data in the Malta-Sicily Channel (A. Drago, 2014). The availability of CALYPSO data is allowing new insights on the hydrodynamical signals in the area especially on the mesoscale and sub-mesoscale variability (S. Cosoli, 2015). The combination of HF radar data to numerical models supports applications to optimize intervention in case of oil spill response as well as for search and rescue, security, safer navigation and improved meteo-marine forecasting. CALYPSO data can also support the operational monitoring of sea conditions in critical areas such as in proximity to ports. However, scattered waves from long range cells can be corrupted and cannot be correctly interpreted by the radar. This interference can drastically affect the spatial coverage of the computed combined sea current fields and of the main wave parameters. Such data gaps in both space and time are highly restrictive on the quality of the service provision to users. HF radar data streams need therefore to be processed to fill in the gaps by reliable guesses. This kind of work has already been performed on sea surface current data using Machine Learning techniques by (A. Gauci, 2016) to fill in missing data in a high resolution grid. Data quality control, is a crucial aspect in the validation of a marine dataset, to identify anomalous values, missing information, and transmission errors. It consists in a sequence of operations that aim to remove all incorrect measurements. Standard validation procedures, such as threshold tests and spike (outliers) detection as described in the main literature, can remove only measurements that appear to be clearly out of a reasonable range, and is often unable to process different time series that are strictly correlated, such as data from neighboring stations. Incomplete datasets are a serious obstacle to the analysis of these data, from the computation of simple descriptive summaries to the estimation of sophisticated models. Moreover, validation can be problematic in case of a considerable number of missing values in the dataset. In order to account for all these issues, a flexible statistical model is here proposed in order to cluster measurements, validate data and assign missing information. Although in the literature only few works have been proposed in oceanographic topics, these models are very useful to cluster and validate any kind of data by choosing appropriate distributions, such as positive data (almost all oceanographic and meteorological measurements), circular data (typically directional data), univariate or multivariate data, longitudinal data or spatial data. Latent class models allow the clustering of multivariate correlated data, such as wave data from different stations or even other marine parameters (periods, directions), by defining the joint distribution of the data as a mixture of different distributions (F. Lagona, 2012).

2 Material and Methods

2.1 The Calypso SWH data

The CALYPSO HF radar system has been running operationally since 2012 and consists of antenna installations on the northern Malta and southern Sicilian shores at four selected sites. In particular, radial sites are installed at Ta' Barkat, limits of Xghajra in Malta, Ta' Sopu limits of Nadur in Gozo, and Pozzallo and Ragusa harbors in Sicily. Real-time information can be accessed from the project website http://www.capemalta.net/CALYPSO. Unfortunately after the installation and calibration of the first radars, strong interference was noted on the frequency band used by the radars. Spectral analysis revealed external transmissions at 13.5MHz which are active every day, especially in the afternoon. According to Resolution 612 of the International Telecommunications Unit (ITU), this band can be used by oceanographic radars on a secondary basis and hence other transmissions are

allowed. This interference has caused data gaps and spikes in both space and time affecting the quality and entirety of the whole dataset. The data used in the current study spans over three months, from 1st December 2016 to 10th January 2017, and comprises several climatologically relevant storms. This selected period is representative of the standard HF radar operation as it includes homogenous time series as well as missing values and outliers due to noise signals and interferences. The significant wave height (SWH), at half-hour intervals, is obtained by the radars at a number of close-in radar range concentric sectors, each sector being approximately 1.6 km wide (Figure 1).

Time series plots of SWH from radar, model and altimeter data are produced and compared to explore the agreement of the different data sources. In particular, satellite altimeter wave measurements are obtained by passes over the Malta-Sicily Channel from the Jason 2, Jason 3 and Saral Altika missions. While the numerical WAM model data is derived from the ISPRA regional model which produces the main wave parameters over a grid having a spatial resolution of 1/60 degrees and a temporal resolution of one hour.

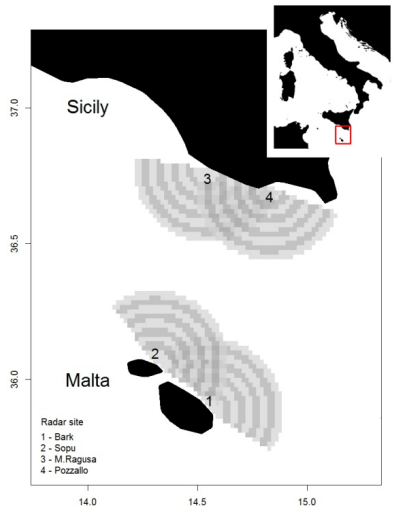

Figure 1: Calypso HF Radar sites

2.2 The HMM model

A sequence model is a model whose job is to assign a label or class to each unit in a sequence, thus mapping a regular series of observations to a sequence of labels. A Hidden Markov Model (HMM) is a probabilistic sequence model; given a sequence of units, it computes a probability distribution over possible sequences of labels and establishes the best label sequence.

A HMM follows the latent class approach and allows for temporal correlation, because the parameters of the mixture model depend on the states of the latent Markov chain.

In this context, the temporal evolution can be conveniently described as a multinomial process in discrete time (J. Bulla, 2012), represented by a sequence $\xi_{0:T} = (\xi_t, t = 0,...,T)$, with $\xi_t = (\xi_{t1} ...\xi_{tK})$ multinomial variables with one trial and K classes. For each time step t, the binary components in ξ_t represent the class membership.

The specification of a HMM is completed by assuming that the observations $z_{0:T} = (z_t , t=0,...,T)$ are conditionally independent, given as a realization of the Markov chain. As a result, the conditional distribution of the observed process, following a given latent process, takes the form of a product density, namely:

$$f(z_{0:T}|\xi_{0:T}) = \prod_{t=0}^{T}\prod_{k=1}^{K}\left(f_k(z_t)\right)^{\xi_{tk}}$$

where $f_k(z), k = 1, \dots, K$ are K univariate or multivariate density distributions.

For classification purposes, these densities are usually assumed to be known up to a number of parameters that indicate the locations and the shapes of K clusters. In environmental studies, normal distributions have been widely adopted for continuous unbounded data. In case of strictly positive values, such as significant wave height observations, skew normal distributions can be adopted, or in a more simple way a log-transformed time series can be used in order to remove any condition on the data domain. Given $z^*=(z^*_1, \dots, z^*_S)$, the log transformation of the time series $z=(z_1, \dots, z_S)$ from S different measurement stations, we assume that observations z^* are conditionally independent given a realization of the Markov chain, and adopt a family of normal densities $f(z^*_s;\beta_s)$, $s=1, \dots, S$, indexed by a parameter β_s that indicates the location and the shape. As a result, we obtain:

$$f(z_{0:T}|\xi_{0:T}) = \prod_{t=0}^{T}\prod_{k=1}^{K}\left(\prod_{s=1}^{S} f(z_s^*;\beta_s)\right)^{\xi_{tk}}$$

The classification procedure is implemented by computing the marginal distribution of the observed data, known up to the set of parameters $\beta = (\beta_1, \dots, \beta_S)$ and the realization of the Markov chain that depends on K initial probabilities $\pi_k=P(\xi_{0k}=1)$, $k=1,\dots,K$, and a transition probabilities KxK matrix $\pi_{hk}=P(\xi_{tk}=1|\xi_{t-1,h}=1)$, $h,k=1, \dots, K$.

Parameters can be estimated through numerical optimization or the likelihood approaches (J. Bulla, 2012). Each observation can be validated according to the class membership and the confidence interval associated to the mixture model. The number K of latent classes has been chosen according to the use of the common goodness of fit indexes based on the likelihood function (BIC, ICL).

3 Results and discussion

3.1 The analysis of Calypso SWH data

Comparisons of HF radar, simulated and satellite-derived significant wave heights (Ta' Barkat site in Figure 2) reveal that there is a good agreement between all these data series; there is also an optimal distance from the radar station for the relevance of the HF radar wave data which is found to be most accurate at the intermediate distances from the coast and deteriorating both closer to coast as well as further offshore.

Figure 3 displays, in the diagonal subplots, the density distributions of the SWH at the respective four radar sites. The density distributions reveal that SWH in the selected period were milder along the Sicilian coasts compared those near the Maltese Islands. In the same plot, in the upper right corner, the values of the linear correlation of the four sites are reported: they are high between the respective pairs of the Maltese and the Sicilian radars. The scatter plots, in the lower left corner, confirm the agreement of the neighborhood sites SWH distributions. This result suggests that as an alternative to the use of the standard univariate analysis, it is possible to aggregate data into bivariate time series for the respective Maltese and Sicilian radar observations.

Even though the selected period of the radar observations are quite complete, some gaps and outliers are still present. The analysis of the missing value pattern is reported in Table 1.

The table reports in the first column the combination of the half-hour SWH that are observed ('O') or are missing ('M') at the four HF radar sites. For instance, the first row indicates the number of half-hour SWH that are contemporarily observed at the four locations, the second row the SWH that are

observed everywhere, but are missing at Pozzallo and so on. In the selected period, Pozzallo reports the most critical situation with 43.7% of missing values and a longer gap of about 6 days.

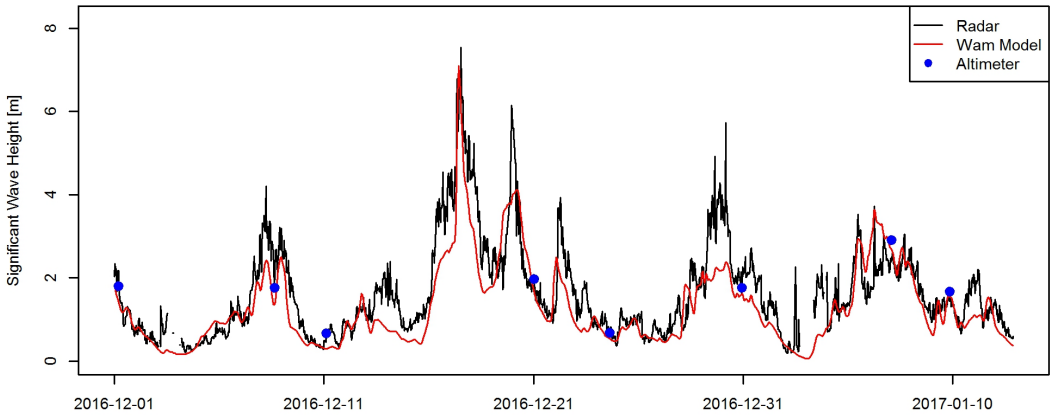

Figure 2: SWH time series at Ta' Barkat from HF Radar (intermediate ring), WAM and satellite.

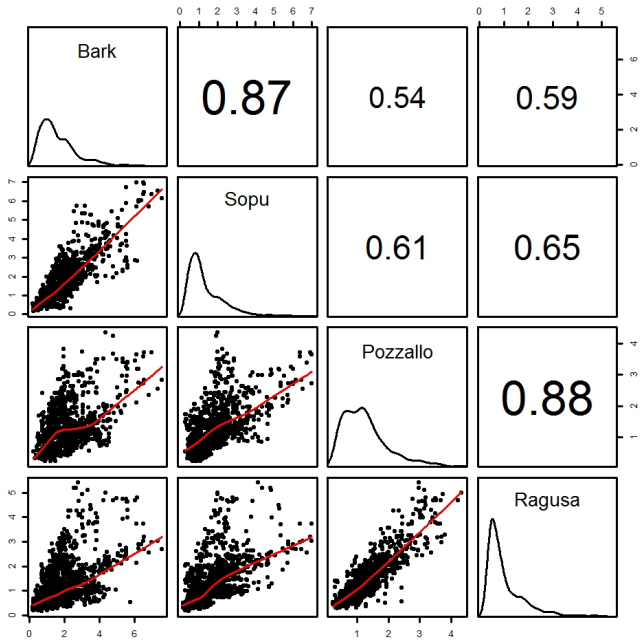

Figure 3: Scatterplots and correlations between the four radar sites. The subplots along the diagonal show the density distributions.

Count	Bark	Sopu	Pozzallo	Ragusa
1059	O	O	O	O
691	O	O	M	O
25	O	O	M	M
81	O	M	O	O
124	O	M	M	O
17	O	M	M	M
9	M	O	O	O
4	M	O	M	O
1	M	O	M	M
10	M	M	O	O
30	M	M	M	O
9	M	M	M	M

Table 1: Missing Pattern.

3.2 The HMM model application

Since the SWH observations are strictly positive data, $z^*_{0:T}$ corresponds to the log-transformation of the original series. The HMM has been estimated in terms of density parameters, transition and posterior probabilities, through maximum likelihood methods, by introducing $K=6$ latent classes.

Figure 4 displays the overall estimation (in black) over the log-transformed data, which is the weighted sum of 6 Gaussian components (coloured densities). Missing data are estimated as expected values of the posterior distribution and subsequently a confidence band for each observation has been computed by bootstrapping 1000 values starting from the HMM. The estimated missing values are represented as red dots and the lower and upper band limits (green lines) correspond to the 2.5^{th} and 97.5^{th} percentile as shown in Figure 5.

Another important feature of this method is that the use of upper and lower confidence bands provides a very effective method to highlight values that can be suspicious if they fall outside these bands. Blue circles in Figure 5 are values outside the confidence limits that have to be subjected to expert judgement since they could be possible extreme SWH values.

4 Conclusions

The implemented procedure proves to be a powerful statistical tool for the quality control and diagnostic analysis of SWH observations by HF radar.

The overall HMM estimated distribution shows a good fit with respect to the observed data and the implemented procedures provide a useful methodology for data validation and quality control purposes. The reconstructed data are able to reproduce the time series pattern although the confidence intervals are still wide in the case of a high percentage of missing values (as in the case of Pozzallo). At the same time, the 2.5^{th} and 97.5^{th} bands suggest values to be investigated as possible extreme values or spikes.

An improvement of the proposed model could be reached following two paths.

Primarily, although the proposed normal distribution mixture fits reasonably the log-transformed observed data, additional univariate or multivariate distribution functions could be tested to improve the performance of the model. Several distribution functions have been already tested in literature for wind and wave data such as Weibull, Rayleigh and Gamma without any data manipulation (I. Pobočíková, 2017).

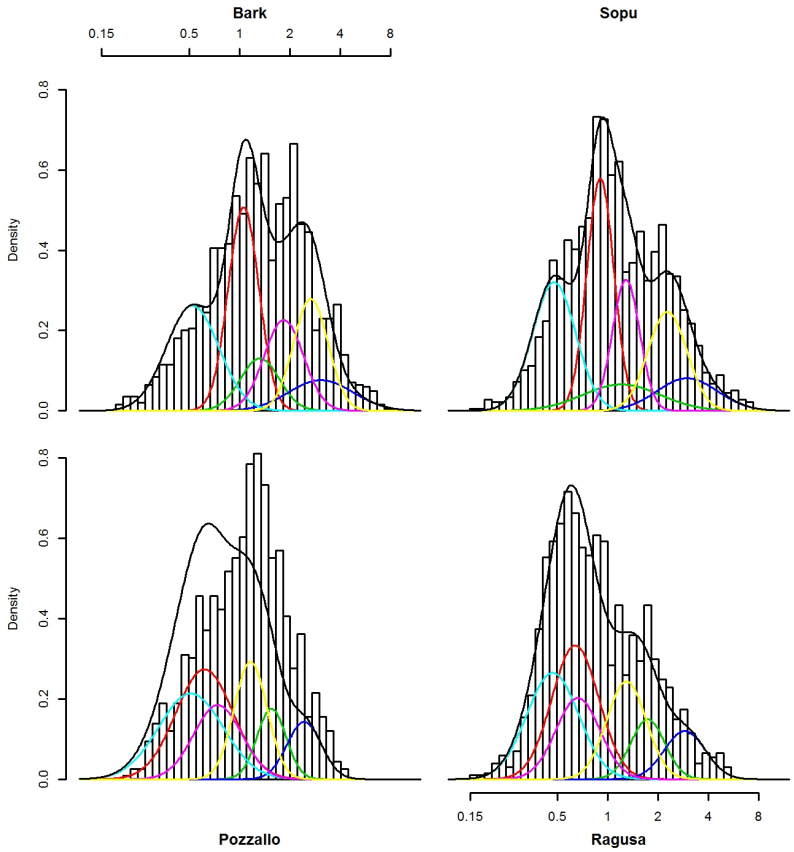

Figure 4: *K*=6 Gaussian densities (coloured lines) on wave height data in log-scale, and global distribution (black) at the four HF radar sites.

On the other hand, it is therefore interesting the introduction of explicative variables strictly connected to the wave motion such as wave periods and wave directions: this further information allowing for a joint classification of physical sea state.

5 Acknowledgement

The CALYPSO Project was partially funded by the Operational Program Italia-Malta 2007-2013.

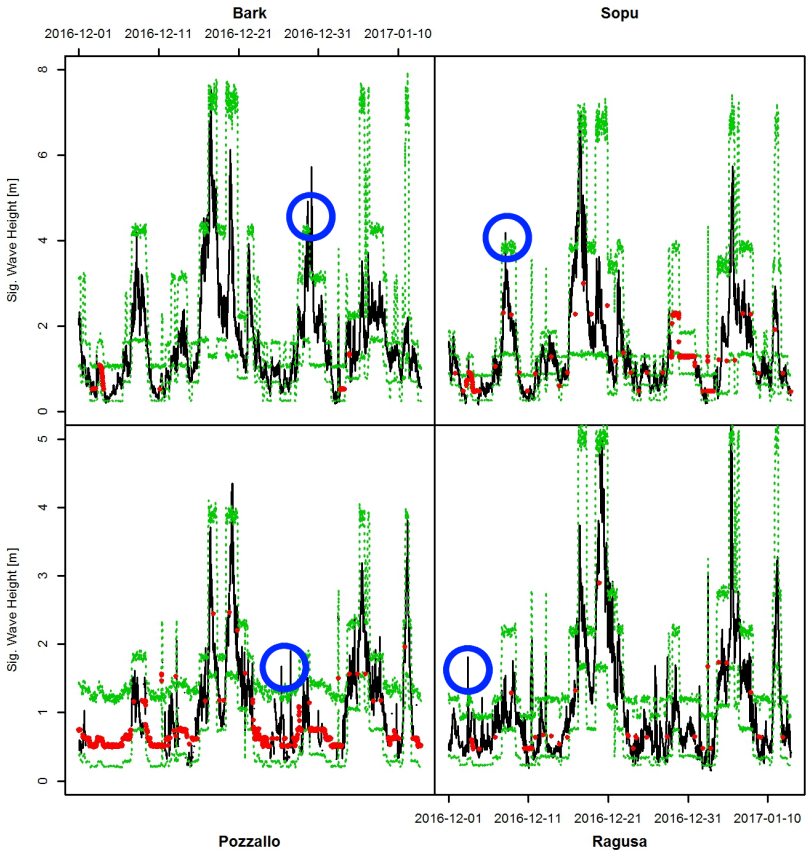

Figure 5: 2.5th and 97.5th percentiles (green) for observed data (black) and computed data (red) according to the estimated model at the four HF radar sites. Possible outliers to be investigated are shown in blue circles.

References

A. Drago, G. C. (2014). CALYPSO – An operational network of HF radars for the Malta-Sicily Channel. *Proceedings of the Seventh International Conference on EuroGOOS.* Lisbon, Portugal: H. Dahlin, N.C. Fleming and S. E. Petersson.

A. Gauci, A. D. (2016). Gap filling of the CALYPSO HF Radar Sea Surface Current Data through Past Measurements and Satellite Wind Observations. *International Journal of Navigation and observation, 3*, 1-9.

F. Lagona, M. P. (2012). Model-based clustering of multivariate skew data with circular components and missing values. *Journal of Applied Statistics, 39*, 927-945.

I. Pobočíková, Z. S. (2017). Application of Four Probability Distributions for Wind Speed Modeling. *Procedia Engineering, 192*, 713-718.

J. Bulla, F. L. (2012). A multivariate hidden Markov model for the identification of sea regimes from incomplete skewed and circular time series. *Journal of Agricultural, Biological and Environmental Statistics, 39*, 544-567.

S. Cosoli, A. D. (2015). Tidal Currents in the Malta-Sicily Channel from High-Frequency radar observations. *Continental Shelf Research, 109*, 10-23.