A NEW METHODOLOGY FOR EXTREME WAVES ANALYSIS BASED ON WEATHER-PATTERNS CLASSIFICATION METHODS

Sebastián Solari[1] and Rodrigo Alonso[1]

Extreme Value Analysis is usually based on the assumption that the data is independent and homogeneous. Historically the hypothesis of independence has received more attention than the hypothesis of homogeneity. The two most common ways of ensuring independence is to use annual maxima or peaks over threshold approaches. In wave and wind extreme analysis, the usual approaches to achieve homogeneous series have been to work to differentiate according to type of process generating the extreme value (e.g. differentiate between hurricanes and cyclones) and conduct directional analyzes. In this work an alternative approach is proposed, based on the use of cluster analysis methodologies to identify weather circulation patterns that results in extreme wave conditions. The proposed methodology is successfully applied to a case study in the Uruguayan South Atlantic coast. From the obtained results it seems that the proposed methodology is able to differentiate the data in homogenous subsets, not only in terms of the target variable (significant wave height) but also in terms of relevant covariables, like wave direction or sea level, and that the extreme value distribution of the whole data, obtained from the distributions fitted to each subset, is fairly insensitive to the number of weather patterns used in the analysis.

Keywords: extremes; waves; weather-patterns

INTRODUCTION

When performing extreme value analysis it is assumed that data are homogeneous, i.e. that all extremes came from the same parent distribution. Then, under the hypothesis that waves generated under different physical process would have different parent distributions, the engineer should recognize the different physical processes that lead to extreme condition before fitting extreme value distributions, e.g. in tropical areas, waves generated during hurricanes or typhoons (tropical cyclones) are analyzed separately from the rest of the data.

At mid-latitudes extreme waves are typically generated by synoptic scale processes (i.e. mid-latitude cyclones); however, at any region it is possible to recognize several distinctive storm types, usually associated to the area where cyclones were generated and to the path they followed until arriving to the study site (e.g. Sartini et al. 2015).

Directional extreme waves analysis is a relatively widespread practice for differentiate data in homogeneous populations, with the additional advantage that a design could be optimized by taking into account that different extreme distributions are associated to different wave directions. (e.g. Jonathan et al. 2008). However, there are two unresolved issues associated with the directional extreme analysis: (a) it is not straightforward to define the number and width of directional bins that should be use, and (b) directional classification of extremes is unable to differentiate among storms of different origin resulting in the same wave direction (particularly relevant in intermediate and shallow waters). In addition, directional analysis can only be applied to directional variables (like winds or waves), but not to scalar variables of interest in coastal engineering (like storm surges of precipitation).

An alternative to directional extreme analysis, aimed to identify homogeneous populations of extreme events, is the use of weather patterns. Weather patterns (WP) are typical synoptic conditions for a given area that are usually given as average fields of some atmospheric variable, like pressure or wind. These WP are commonly used on atmospheric sciences, and to some extent also in hydrology, in order to characterize commonly observed and distinctive conditions that give place to an expected behavior of some variable (like precipitation, wind, etc.).

The use of WP on coastal engineering application is quite limited. Pringle at al. (2014) proposed a classification of the circulation patterns that drive wave climate in KwaZulu-Natal coasts, South Africa. Their approach results in a set of WP that explain the wave climate but there is no indication on how to apply them in order to characterize average or extreme wave conditions. Latter, Pringle et al. (2015) analyzed the link between WP and extreme wave events for the same region. Camus et al. (2014) proposed a statistical downscaling framework based on the use of weather-type classification methods, with a focus on average wave conditions (no extreme conditions). More recently, Rueda et al. (2016a, 2016b) proposed a methodology for wave and storm surge extreme analysis based on weather-type classification methods. The methodology proposed by Rueda et al. (2016a, 2016b) is based on the use of a large number of weather patterns (approx. 100) to properly classify all daily maxima values. Even

[1] IMFIA-Universidad de la República, Julio Herrera y Reissig 565, 11300, Montevideo, Uruguay

though results obtained with this methodology are promising, in our opinion there are two issues that may prevent its widespread application in coastal engineering practice: first, the fact that it may result difficult to give physical interpretation to a large number of weather patterns; secondly, that the methodology focuses on modelling "daily extreme" conditions, something that departs from the common practice of modeling only "annual extreme" events (either through annual or monthly maxima or through peaks over threshold).

In this paper, we propose a methodology for performing extreme waves analysis, aiming to ensure that extreme data is differentiated in homogenous populations and to facilitated the definition of the number of populations to be used in the analysis. The methodology is intended to be used in mid-latitudes, where extreme wave conditions are generated by mid-latitude cyclones of synoptic scale.

OBJECTIVES

To develop and implement a methodology to the analysis of extreme wave conditions based on atmospheric circulation or weather pattern classification, that relies in the use of a reduced number of weather patterns and that focuses only on extreme events.

To apply this methodology to a case study on the Uruguayan South Atlantic coast.

METHODOLOGY

The proposed methodology is based on an automatic classification of the extreme wave events by means of weather-patterns, taking into account the spatial and time evolution of the pressure and wind fields that were registered (or hindcasted) previously and simultaneously to the occurrence of the extreme event. In line with Caravaglia et al. (2010), whom proposed a similar methodology for analyzing extreme precipitation events in Southern France, the methodology comprises the following four steps (outlined in Fig. 1):

1. Obtain a series of independent peak values from the whole series of wave height (e.g. from the hourly series).
2. Define homogeneous subsets using WP
3. Fit an Extreme Value (EV) distribution to each subset.
4. Reconstruct the EV distribution of the original set.

Details of the four steps are given below.

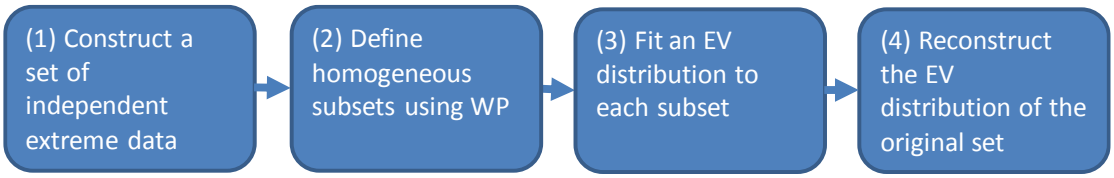

Figure 1. Outline of the proposed methodology.

Step 1: Construction of a set of independent extreme value data

The first step of the proposed methodology is divided in three sub-steps (see Fig. 2 for an outline of the three steps):

a. First, use a moving window approach to find a set of peaks of the target variable (significant wave height in this case).
b. Second, choose a threshold in order to reduce the length of the peaks data set and to retain only peak values with reasonable high value (this step follow the Bernardara et al. 2014 concept of the physical threshold).
c. Third, construct the set covariates for each retained extreme event. The set of covariates is composed from both: state variables that complete the characterization of the sea state (e.g. Tp, Dir, S.L., etc.), and the simultaneous and lagged fields of the atmospheric variables used for the definition of the weather-patters (Mean Sea Level Pressure SLP, Surface Winds, etc.).

The use of the moving windows results in a number of peaks per year just under what is obtained by dividing the length of the year by the width of the window, and warrants that the time between two consecutive peaks is equal or greater than the width of the window. The width of the window must be chosen by the analyst and is a free parameter of the methodology, as it is the "physical threshold". Along this work we used a width of seven days and a physical threshold equal to the 90% quantile of the whole data set.

Step 2: Definition of homogenous subsets

Our work hypothesis here is that homogenous subsets may be obtained by classifying the extreme value data set by means of weather patterns. Under this hypothesis, the problem of defining homogeneous subsets reduces to the classification of the data in accordance to a set of WP.

There may be many ways for defining a set of WP. In this work, we chose to construct the set of WP as the centroids of the clusters obtained applying the Nucleated Agglomerative Clustering technic (see e.g. Wilks 2011), based on k-means and Euclidian distance. Once the methodology for constructing the clusters is chosen, it is necessary to define:

a. The variable(s) to use in the cluster analysis.
b. If the variable(s) should be normalized in some way or not (i.e. if WP will be constructed based on the original value of the atmospheric variables or only taking into account the "shape" of the atmospheric field).
c. The spatial and time domain to be used in the cluster analysis.
d. The number of clusters (WP) to use.

In our case study, we have explored the use of the following atmospheric variables for the construction of the WP: Mean Sea Level Pressure, Mean Sea Level Pressure Anomaly and Surface Wind Speed. In addition, original and normalized values where used (normalization performed for each field, so only the shape of the fields is used in the clusters analysis).

Spatial and time domain were chosen after analyzing linear correlation maps between peak values of the target variable (significant wave height at our study site) and wind fields, for several time lags. For the estimation of the linear correlation between a scalar (significant wave height) and a directional (wind speed) variable, the following equation was used:

$$\rho_{i,j,t} = \max_{\theta} \left\{ \rho\left(H_{m0}, W_{i,j,t,\theta}\right) \right\} \qquad (1)$$

where $\rho_{i,j,t}$ is linear correlation estimated for node *(i,j)* and time lag *t*, $\rho(H_{m0},W_{i,j,t,\theta})$ is linear correlation between significant wave height peaks series (H_{m0}) and wind speed series at node *(i,j)*, with time lag *t*, projected in direction θ ($W_{i,j,t,\theta}$). For estimation of the maximum linear correlation, projection are taken every 1º.

Regarding the choose of the number of cluster to use in the analysis, we use the total variance (sum of intracluster variances) as a guide: while the increase in the number of cluster produce a significant reduction in the total variance, it is reasonable to enlarge the number of clusters in the analysis, otherwise it is not. This do not gives an objective criteria for choosing the number of cluster to use, but allows for the construction of variance reduction charts that orientate the decision; final decision corresponds to the analyst and must be made taking into account the resulting WP and its physical interpretation. The total variance is estimated as:

$$W = \sum_{g=1}^{G} \sum_{i=1}^{N_g} \left\| X_i - \overline{X}_g \right\|^2 \qquad (2)$$

where $g=1,\ldots,G$ are the clusters (WP), $i=1,\ldots,N_g$ are the elements in cluster g, X_i are the individual atmospheric fields and X_g is the average atmospheric field of cluster g.

Step 3: Fitting an EV distribution to each subset

The set of peak values is divided in subsets, following the result of the cluster analysis. Each subset is assumed homogeneous; i.e. given that all data in a subset is generated by similar atmospheric circulation processes, it is assumed that all data comes from the same parent distribution. Under this hypothesis it is possible to applied extreme value theory and to fit an EV distribution to the peaks in each subset.

In this work the peaks in each subset are fitted by a Generalized Pareto Distribution (GPD), following Solari et al. (2017) for the estimation of the threshold and L-Moments for the estimation of the parameters of the GPD (see Solari et al. 2017 for details).

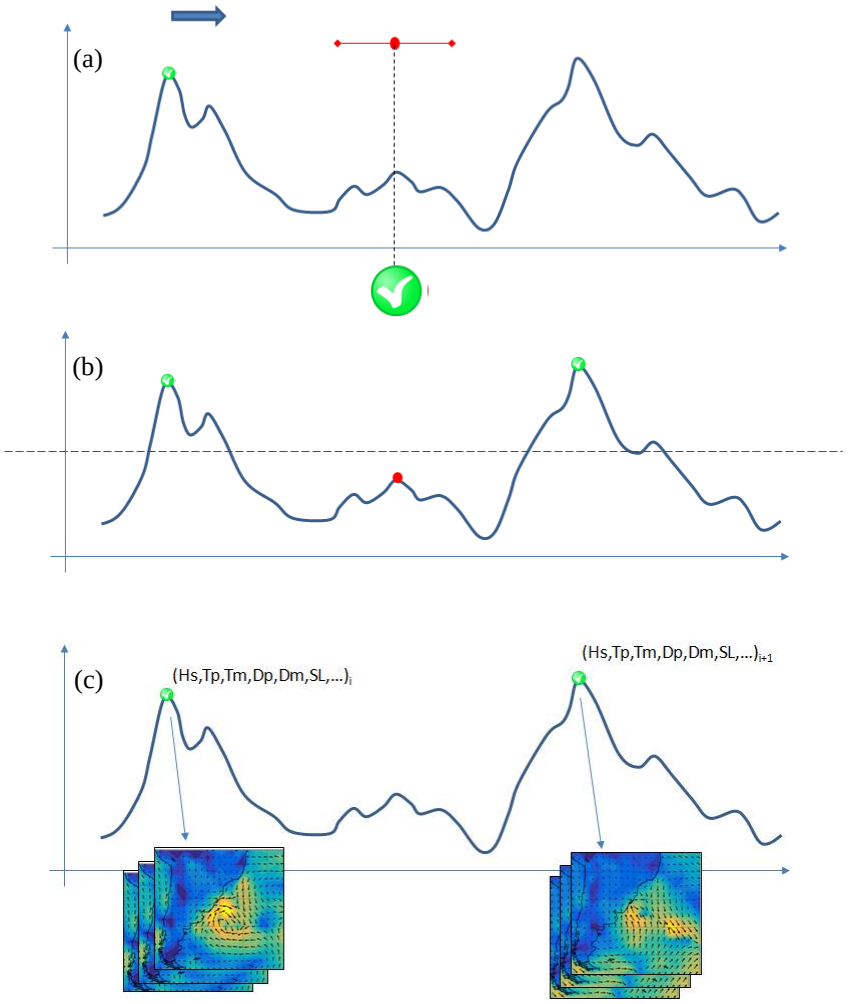

Figure 2. Outline of step (1) of the proposed methodology: Construction of a set of independent extreme value data.

Step 4: Reconstruct the EV distribution of the original data set from the EV distributions fitted to the subsets

In order to obtain the EV distribution of the whole data set (omni-WP distribution, in analogy with the omnidirectional distribution term used in directional analysis), as well as its confidence intervals, we propose to use the Monte Carlo simulation method outlined in Fig. 3.

The method is straightforward, with the only particularity of using a multivariate Poisson model for the simulation of the number of events per year that falls in each WP. Multivariate Poisson model is constructed using a Gaussian copula and marginal Poisson distributions.

The parameters of the distributions that make up the model are resampled before each repetition N_r, in order to fully reproduce the uncertainty of the whole model. For this a bootstrapping approach is used.

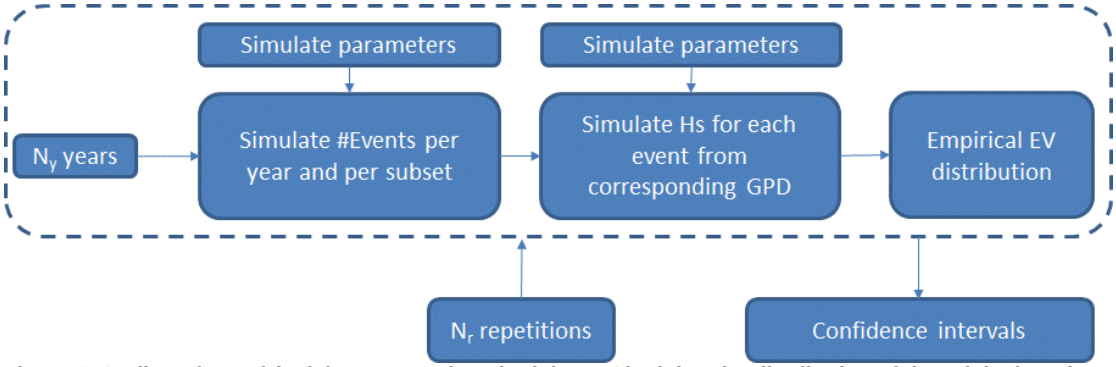

Figure 3. Outline of step (4) of the proposed methodology: Obtaining the distribution of the original set from the EV distributions fitted to the subsets.

APPLICATION

Case Study

Our study case is a series of hindcasted wave parameters (Alonso et al. 2015) and sea levels in front of the Uruguayan South Atlantic coast, at a depth of approx. 20 m (see Fig. 4). The series comprised the period 1980-2010 for wave parameters and 1993-2010 for sea levels, with 3-hourly time step. Fields of SLP and wind at 10 m height are obtained from NCEP Climate Forecast System Reanalysis (Saha et al. 2010). The zone is microtidal, so no distinction is made between astronomical tides and storm surges, and sea level is treated as a random variable.

Following Methodology section, all significant wave height peaks are identify using a seven days moving window and only those over the 90% quantile of the whole data set are retained for the analysis. Using this filtered data set, the correlation maps shown in Fig. 5 are constructed. From these, a spetaial domain (45ºW to 65ºW, 25ºS to 45ºS) and time lags (0, 6 and 12 hours) are chosen.

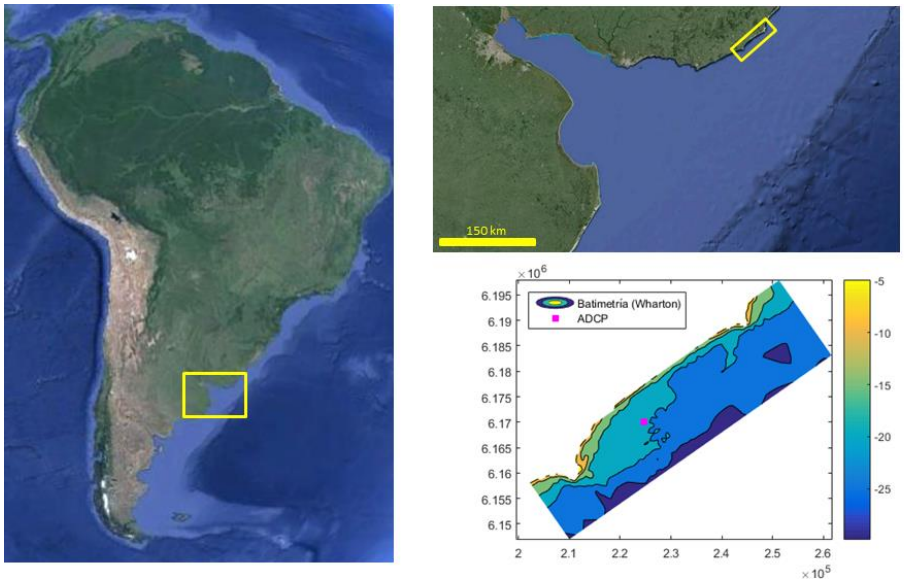

Figure 4. Location of the study case: hindcasted wave parameters and sea level in front of Uruguayan South Atlantic coast (violet square dot in the lower right panel).

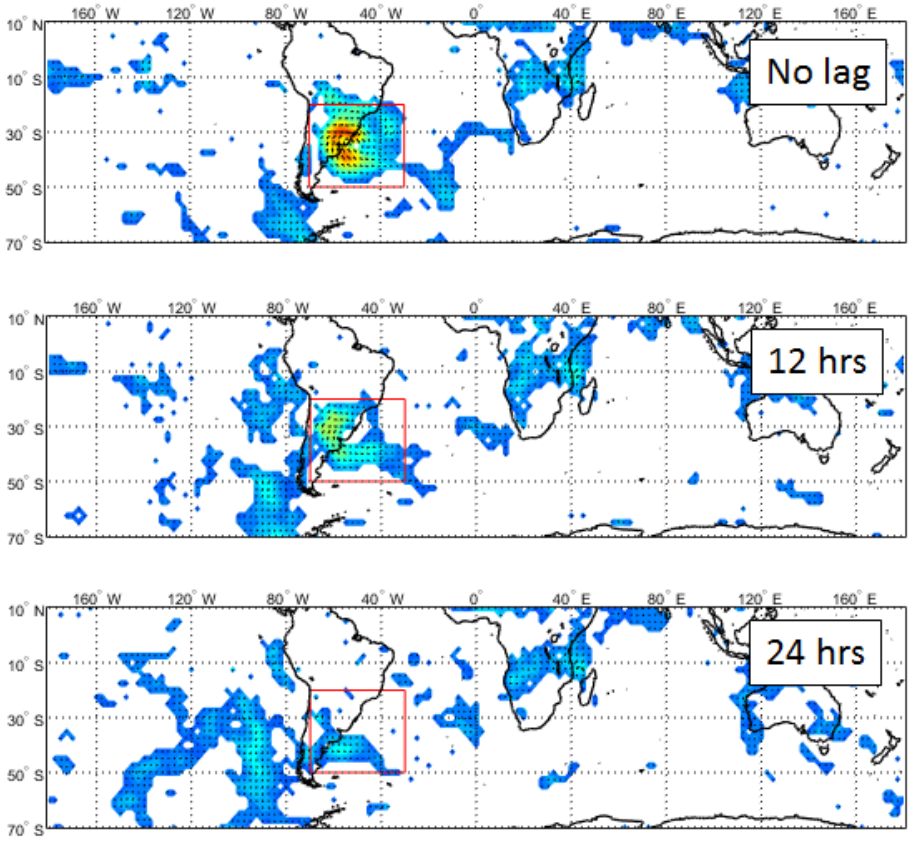

Figure 5. Correlation between the series of significant wave heights peaks and surface wind speed, for different time lags.

Results and Discussion

Fig. 6 shows variance reduction charts (Eq.2), as percentage of total variance obtained with only one cluster, constructed using different atmospheric variables. Firstly, it is observed that all graphs are very similar, whether they are constructed using SLP, SLP anomaly or surface wind fields, and whether the original values of these variables are used or the normalized fields (field shapes). Accordingly, in the following, we work with the clusters obtained using surface wind fields without normalization (bottom left frame in Fig. 6), since this variable is the one that has a more direct relation with the target variable and, for our particular case study, with the sea level variations (see Santoro et al., 2013).

From the variance reduction graph obtained with the wind fields (bottom left frame in Fig. 6) it is clear that separating the initial population into two clusters leads to a reduction of the total variance of approx. 25%, and when separating in three clusters an additional decrease of almost 10% is achieved. Increases in the number of clusters, until reaching values of 4 to 10, produce additional reductions of the total variance under 5% per additional cluster, and for a total number of clusters greater than 10, the reduction of the total variance per additional cluster is always under 1%.

Fig. 7 and 8 show the WP that are obtained with two and three clusters respectively, along with the mean annual cycle of the number of events per WP. Fig. 9 shows the EV distributions obtained in both cases, along with scatter plots of covariables: significant wave height and mean direction (D_m-H_{m0}), significant wave height and sea level (SL-H_{m0}) and deep water wave steepness and direction (D_m-s_0) Firstly, it is observed that the proposed methodology result in clearly differentiated WP. In the case of two clusters (Fig. 7), WP#1 is characterized by a low pressure to the N of the study zone and a high pressure to the S, both traveling Eastward, while WP#2 is characterized by a larger low pressure to the SE of the study zone, traveling Eastward, and a high pressure to the W, coming from the SW. Both WP are repeated in the case of three clusters (see Fig. 8), and a new one shows up (WP#3) that is characterized by a low pressure to the E of the study zone and a high pressure that span from SW to W of the study zone.

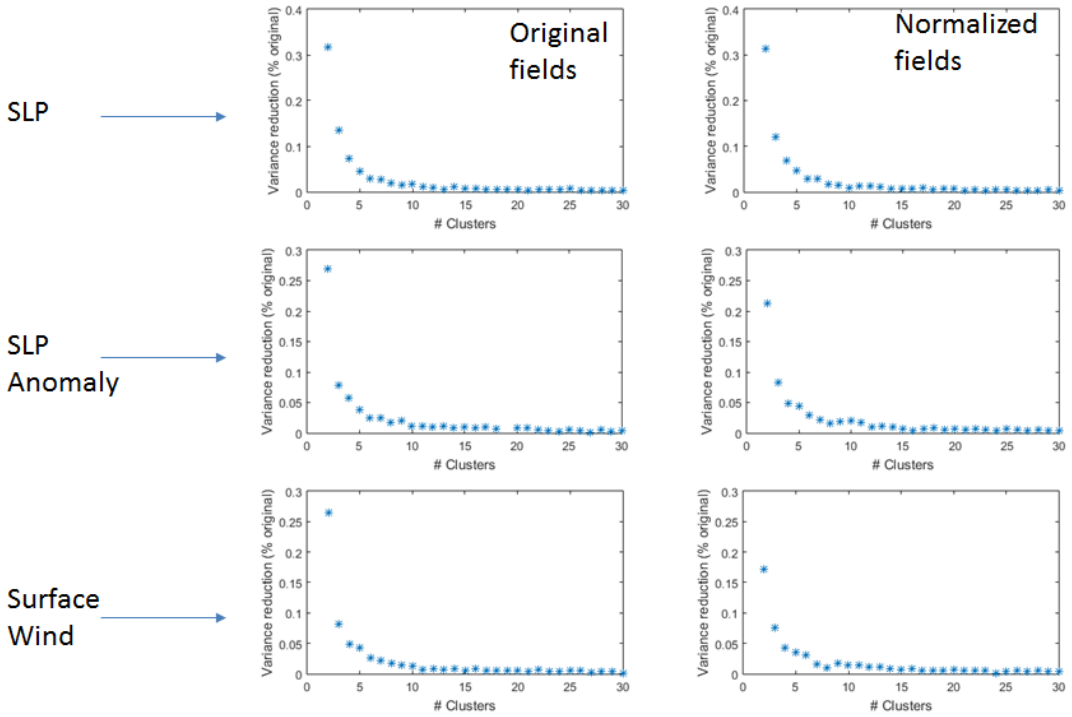

Figure 6. Variance reduction charts. Each dot shows the reduction in the total variance produced by increasing the number of cluster by one, relative to the variance obtained with no clusters (i.e. the whole data set assigned to a single cluster).

Regarding the quality of the fit of the GPD to the data, it is noted that a good fitting is obtained in all cases (see Fig. 9 top frames). In the case of two clusters (Fig. 9, top left frame), it is noted that both EV distributions are quite similar. However, in the case of three clusters (Fig. 9, top right frame) it is noted that WP#1 and WP#3 have the same trend, with values of WP#3 approx. 50 cm larger than values of WP#1, while WP#2 has a noticeable different trend, resulting in lower high return period quantiles than the others WPs.

The effect of including a third WP is particularly noticeable in the scatter plots of the covariables (Fig. 9, all frames but the two on top). Whether two or three clusters are used, WP#2 (similar in both classifications) produces a clearly distinguishable clustering of covariates: mean direction between 140° and 160°, tending to 140° as H_{m0} increases, sea level above average and approximately linear relationship between the deep water wave steepness and the mean direction (being refraction the possible link between them). In the case of two clusters, the data belonging to WP#1 do not present clear groupings in terms of covariates. On the contrary, in the case of three clusters, the data belonging to WP#1 and WP#3 clearly differ in both the mean direction and the sea level with which they occur; i.e. although the GPD obtained with data of WP#1 and WP#3 are very similar, with WP#3 events being 50 cm larger than those of WP#1 for the same return period, the extreme events of WP#1 are associated with below-average sea level and ESE mean direction, while extreme events of WP#3 are associated with above-average sea level and SSE mean direction.

At this point is interesting to note that the classification of the data that is obtained through WPS can not be reproduced by a directional analysis. In this sense, doubts arise about the ability of directional analysis to produce a classification of data that results in homogeneous populations, at least in this case study; see e.g. the scatter plots of D_m-s_0, where a directional classification would not be able to differentiate the behavior of WP#2.

Table 1 shows, for the case of three WPs, the linear correlation between the number of events per year in the different WPs, and between the number of events per year and different climatic indexes, namely: Antartic Oscillation (AAO or SAM), El Niño (NINO 3.4) y Tropical South Atlantic Index (TSA). In the table, only statistically significant correlations are shown, using significance level α=0.05. It is noted that there is negative correlation between the number of events per year of WP#1 and WP#3, i.e. years with a number of events in WP#1 above the average tend to have a number of events in WP#3 below average, and vice versa. The fact that we have correlation between the number of events per year in different WPs, highlights the importance of using a multivariate distribution for the simulation of the

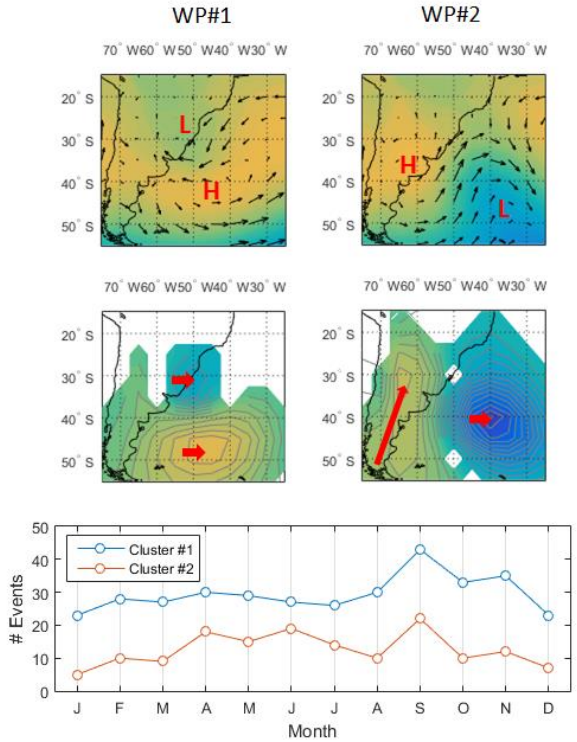

Figure 7. Weather-patterns obtained with two clusters. Upper panels: average mean sea level pressure and surface wind fields for each pattern. Central panels: average MSLP anomaly for each pattern and arrows showing the average travel of lows and highs in 12 hours. Lower panel: annual cycle of the mean number of events per cluster.

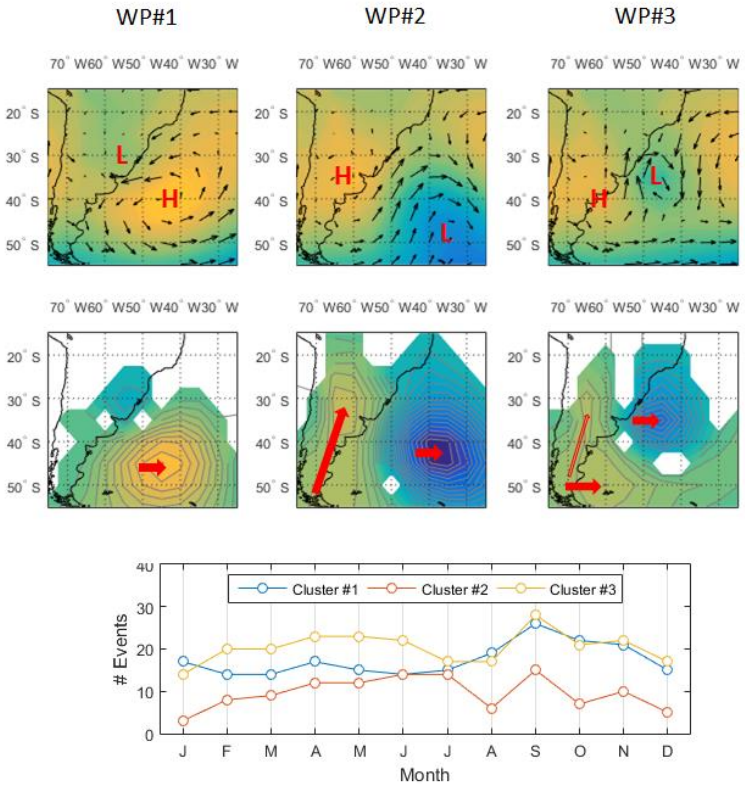

Figure 8. Weather-patterns obtained with three clusters. Upper panels: average mean sea level pressure and surface wind fields for each pattern. Central panels: average MSLP anomaly for each pattern and arrows showing the average travel of lows and highs in 12 hours. Lower panel: annual cycle of the mean number of events per cluster.

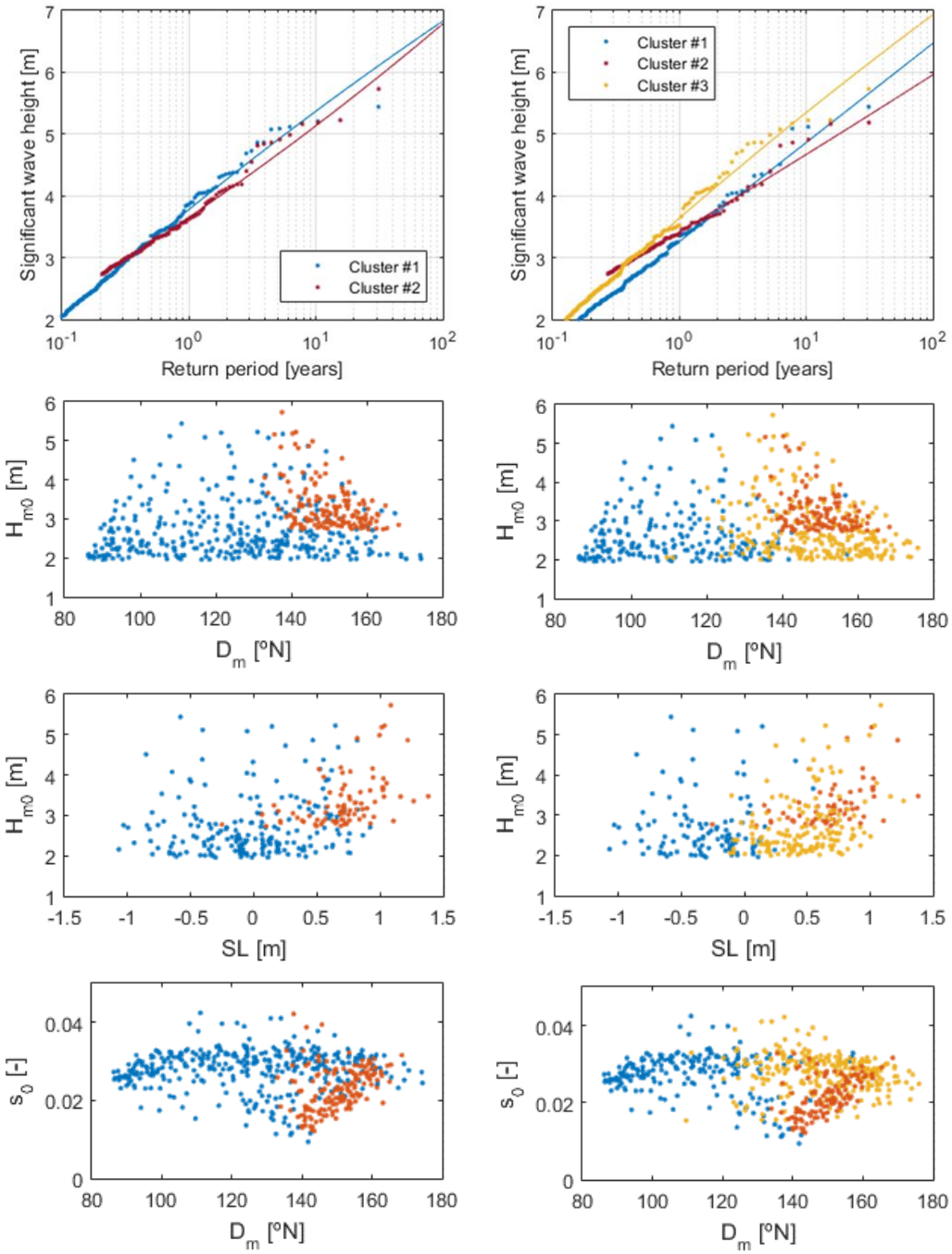

Figure 9. EV distribution and scatter plots of the covariates obtained with two WPs (left) and three WPs (right). Top frames: GPD fitted to data in each WP. Three lower frames: scatter plots of the covariates, with colors for differentiate each WP (same colors as top frames).

number of events per year when calculating the omni-WP distribution (a multivariate Possion distribution in this particular case). From Table 1 also arises that the only WP that have statistically significant linear correlation with the climatic indexes is WP#2, showing in this case that the number of events per year in WP#2 might be influenced by El Niño and TSA anomalies. Although not done here, these effects may be included in the simulation process outlines in Fig. 3

Lastly, Fig. 10 shows the comparison of the "omni-WP" distribution obtained directly by fitting a GPD to the data and by following the methodology outlined in Fig. 3. It is noted that both distributions agree, with a difference of approx. 5% for the 100-years return period quantile, and that the width of the confidence intervals is similar in both cases (there is no increase nor reduction of the uncertainty when using the proposed, more complex, statistical model). Also in Fig. 10 (lower frame) is included the "omni-WP" distribution obtained using 20 WPs; it is noted that the obtained EV distribution is in agreement with the one obtained with no WPs and with the one obtained with three WPs, and that in spite of the significant increase in the complexity of the model, there is no significant widening of the confidence intervals. In this regards, the proposed methodology seems to be insensitive to the choose of the number of WP to use in the analysis.

Table 1. Linear correlation between annual number of events in the different Weather Patterns, and between the annual number of event and different climate indexes.

	WP#1	WP#2	WP#3	AAO	Niño 3.4	TSA
WP#1	1	---	-0.46	---	---	---
WP#2		1	---	---	0.36	-0.45
WP#3			1	---	---	---

CONCLUSIONS

A methodology was introduces for performing EV analysis of met-ocean variables based on WP classification. The proposed methodology is in line with v in that is based on the same four basic steps, but differ in the ways this steps are implemented in practice.

From applying the proposed methodology to a case study in the Uruguayan South Atlantic coast, it is concluded that:

- The methodology seems to be able to identify homogeneous populations, not only in terms of the target variable whose extreme values are being analysed but also in terms of other covariables that are of utmost interest in coastal engineering
- It also properly reproduce the "omni-WP" distribution of the data. In this regards, the methodology seems to be unsensitised to the number of WP chosen for the analysis, which is an advantage since the choose of the number a WP remains subjective.

In addition, the proposed methodology provides a better insight of the physical processes that results in the observed extreme conditions and provides a way to reasonable generate combinations of the extreme variable and its covariables that might be use for design.

ACKNOWLEDGMENTS

NOAA National Centers for Environmental Prediction (NCEP) is acknowledge for providing CFSR surface wind and mean sea level pressure data. Uruguayan Ministry of Transport and Public Works (DNH-MTOP) is acknowledge for partial founding this research.

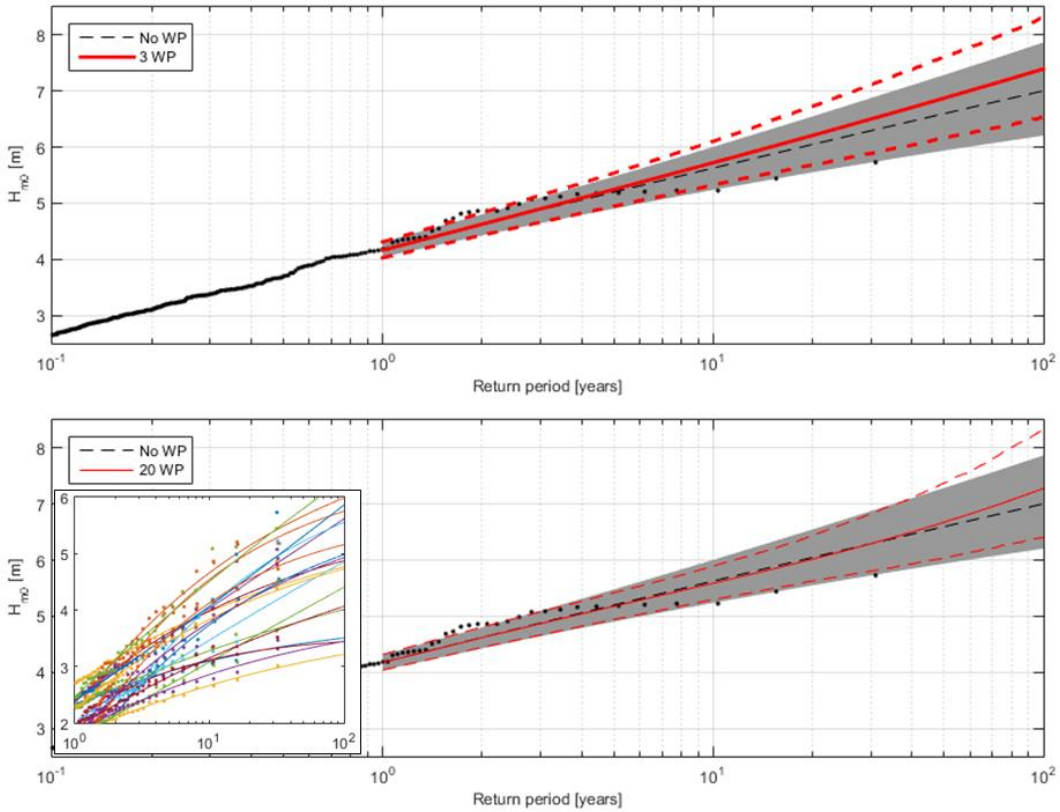

Figure 10. EV distribution of the whole peaks data set ("omni-WP" distribution) obtained by fitting a GPD to the original data (black dashed line and gray shadow) and from applying the simulation methodology outlined in Figure 3 (red continuous and dashed lines). Top panel: result obtained by using 3 WP. Low panel: result obtained by using 20 WP, along with the GPD fitted to the peaks in each WP.

REFERENCES

Alonso, A., S. Solari and L. Teixeira. 2015. Wave energy resource assessment in Uruguay. *Energy* 93, 683-696.

Bernardara, P., F. Mazas, X. Kergadallan and L. Hamm. 2014. A two-step framework for over-threshold modelling of environmental extremes. *Natural Hazards and Earth System Sciences*, 14, 635-647.

Camus, P., M. Menéndez, F.J. Méndez, C. Izaguirre, A. Espejo, V. Cánovas, J. Pérez, A. Rueda, I.J. Losada and R. Medina. 2014. A weather-type statistical downscaling framework for ocean wave climate. *Journal of Geophysical Research: Oceans* 119, 1-17.

Caravaglia, F., J. Gailhard, E. Paquet, M. Lang, R. Garçon and P. Bernardara. 2010. Introducing a rainfall compound distribution model based on weather patterns sub-sampling. *Hydrology and Earth System Sciences*, 14, 951-964.

Jonathan, P., K. Ewans, G. Forridstall. 2008. Statistical estimation of extreme ocean environments: The requirement for modelling directionality and other covariate effects. *Ocean Engineering* 35, 1211-1225.

Pringle, J., D.D. Stretch and A. Bárdossy. 2014. Automatic classification of the atmospheric circulation patters that drive regional wave climates. *Natural Hazards and Earth System Sciences*, 14, 2145-2155.

Pringle, J., D.D. Stretch and A. Bárdossy. 2014. On linking atmospheric circulation patterns to extreme wave events for coastal vulnerability assessments. *Natural Hazards* 79, 45-59.

Rueda, A., P. Camus, F.J. Méndez, A. Tomás and A. Luceño. 2016a. An extreme value model for maximum wave heights based on weather types. *Journal of Geophysical Research: Oceans* 121, 1-12.

Rueda, A., P. Camus, A. Tomás, S. Vitousek, F.J. Méndez. 2016b. A multivariate extreme wave and storm surge climate emulator based on weather patterns. *Ocean Modelling* 104, 242-251.

Saha, S., S. Moorthi, H-L Pan, X. Wu, J. Wang, S. Nadiga, et al. 2010. The NCEP Climate Forecast System Reanalysis. *Bulletin of the American Meteorological Society* 91, 1015-1057.

Santoro, P., M. Fossati, I. Piedra-Cueva. 2013. Study of the meteorological tide in the Rio de la Plata. *Continental Shelf Research* 60, 51-63.

Sartini, L., F. Cassola and G. Besio. 2015. Extreme waves seasonality analysis: An application in the Mediterranean Sea. *Journal of Geophysical Research:* Oceans 120, 6266-6288.

Solari, S., M. Egüen, M.J. Polo, M.A. Losada. 2017. Peaks Over Threshold (POT): a methodology for automatic threshold estimation using goodness-of-fit p-value. *Water Resources Research*. Accepted for publication.

Wilks, D.S. 2011.*Statistical Methods in the Atmospheric Sciences. Third Edition*. Academic Press.