Stochastic Environmental Research and Risk Assessment manuscript No.
(will be inserted by the editor)

Revisited mixed extreme wave climate model for reanalysis data bases

R. Mínguez · F . Del Jesus

Received: date / Accepted: date

Abstract Mixed Extreme Value (MEV) models (Mínguez et al, 2013b) have proved to be an appropriate tool for dealing with wave maxima because they take full advantage of upper tail information from both i) hindcast or wave reanalysis and ii) instrumental records, which reduces the uncertainty on return level estimates. However, in order to characterize stochastically the differences between instrumental and reanalysis maxima, the method developed in Mínguez et al (2013b) only uses information about annual maxima. This technical note revisits the MEV method so that those differences between instrumental and reanalysis maxima could be characterized using information on independent storm peaks, instead of annual extremes. This strategy increases the size of data sets during the estimation process, reducing uncertainty. The Revisited Mixed Extreme Value (RMEV) model is illustrated using data from the same location studied in Mínguez et al (2013b), and results are compared.

Keywords design return periods · extreme waves · wave reanalysis

1 Introduction

Extreme wave climate analysis is of paramount importance for the design of coastal and offshore structures. For this reason, the proper characterization of wave climate has been and still is an intensive area of research within the scientific and engineering community. Traditionally, all steps involved within the design process are based on instrumental records (mainly buoy), however, over the last decade, and in an attempt to improve the knowledge about wave climate, there has been

R. Mínguez
Independent Consultant, Ciudad Real, Spain
Tel.: +34-926-810046
E-mail: rominsol@gmail.com
F. Del Jesus
Environmental Hydraulics Institute, Universidad de Cantabria, Cantabria , Spain
E-mail: dejesusf@unican.es

an outstanding development of wave reanalysis models. These models allow a detailed description of wave climate in locations where long-term buoy records do not exist.

Scientists and engineers have started using these data bases for design purposes. However, as pointed out by several authors (Caires and Sterl (2005); Cavaleri and Sclavo (2006); Mínguez et al (2011, 2012); Reguero et al (2012)), there are discrepancies when comparing reanalysis versus instrumental data, which must be accounted for within the design process. These authors propose several calibration/correction techniques, which are valid for most of the range of the wave height probability distribution except for the upper tail. Note that there is an statistical theory of extreme values (EVT) (Castillo, 1988; Coles, 2001; Katz et al, 2002; Castillo et al, 2005) that provides the mathematical framework for modeling the tail distribution and none of those calibration/correction techniques is consistent with this EVT.

To fill this niche, Mínguez et al (2013b) proposes the mixed extreme value (MEV) climate model. This method allows correcting discrepancies between instrumental and reanalysis records in the upper tail and it is consistent with EVT. However, in order to characterize stochastically the differences between instrumental and reanalysis maxima, the method developed in Mínguez et al (2013b) only uses information about annual maxima, and it is not possible to apply methods such as Pareto-Poisson (Leadbetter et al, 1983) or Peaks Over Threshold (POT, Davidson and Smith (1990)), which are known to be more robust because they use more information during the estimation process. Note that within MEV it is possible to apply those methods only for the reanalysis data, however, the final convolution is performed in terms of annual maxima.

The aim of this paper is to revisit the MEV method so that those differences between instrumental and reanalysis maxima could be characterized using information on independent storm peaks, instead of annual maxima. This strategy increases the size of data sets during the estimation process, reducing uncertainty. Besides, this method is specially convenient in locations where short records of instrumental data are available, such as the case for under developing countries, or the analysis of design locations for the off-shore industry.

The rest of the technical note is organized as follows. Section 2 presents the proposed Revisited Mixed Extreme Value model. Section 3 shows the performance on real data from a given location in the North of Spain. Note that we use the same example as Mínguez et al (2013b). Finally, in Section 4 relevant conclusions are drawn.

2 Revisited Mixed Extreme Value Analysis Model

MEV model proposed in Mínguez et al (2013b) relies on the following assumptions:

1. The annual maximum reanalysis random variable X follows any distribution for maxima $F_X(x, \boldsymbol{\theta}_X)$.
2. The random variable Y corresponding to the difference between instrumental and reanalysis data conditioned to the annual reanalysis maximum data (X) follows a normal distribution, i.e. $f_{Y|X}(y) \sim N\left(\mu_{Y|X}, \sigma^2_{Y|X}\right)$.

According to these assumptions, the MEV extreme value model can only use annual maxima information to characterize the differences between instrumental and reanalysis data, and in case of working with Peak Over Threshold methods, this constraint does not allow to use all differences available among reanalysis storm peaks and their corresponding instrumental records. Alternatively, the Revisited Mixed Extreme Value model (RMEV) is based on the following alternative assumptions:

1. The number of independent storm peaks N exceeding threshold u in any one year follows a Poisson distribution with parameter λ.
2. The random variable X associated with independent reanalysis storm peaks follows a distribution with cumulative distribution functions $F_X(x, \boldsymbol{\theta}_X)$. According to Davidson and Smith (1990) this distribution function may correspond to Pareto, and this distribution is used in this paper, however the functioning of the proposed method is not limited to this parametric distribution.
3. The random variable Y corresponding to the difference between instrumental and reanalysis data conditioned to the reanalysis storm peak (X) follows a normal distribution, i.e. $f_{Y|X}(y) \sim N\left(\mu_{Y|X}, \sigma^2_{Y|X}\right)$.

Considering assumption (1) from RMEV, instead of dealing with the annual maximum random variable, we work with the random variable related to storm peaks $Z = X + Y$. Its corresponding cumulative distribution function, according to assumptions (2) and (3), is equal to:

$$F_Z(z) = \int_u^{\infty} f_X(x, \boldsymbol{\theta}_X) \Phi\left[\frac{z - x - \mu_{Y|X}}{\sigma_{Y|X}}\right] dx, \tag{1}$$

where $\Phi(\cdot)$ is the cumulative distribution of the standard normal random variable. Note that the integration limits range from u to ∞ since we are assuming the use of Pareto, however, these limits may change depending on the type of probability density function used for X.

The structure of the RMEV model is the same as the MEV, but the data used in the analysis is different. Regarding the numerical solution of the integral in (1), the same recommendations given for the MEV method still apply, i.e., the adaptive Gauss-Kronrod quadrature method (Shampine, 2008) is the most appropriate, since it supports infinite intervals and can handle moderate singularities at the endpoints.

However, the cumulative distribution function given by (1) does not correspond to annual maxima, which is usually the information required for engineering design. Considering assumptions (1)-(3) from RMEV, the probability of the annual maximum of the process to be lower than or equal to z is:

$$\begin{aligned} \text{Prob}\left(\max_{1 \le i \le N} Z_i \le z\right) &= \text{Prob}(N = 0) + \sum_{n=1}^{\infty} \text{Prob}(N = n) F_Z(z)^n \\ &= e^{-\lambda}\left[\sum_{n=1}^{\infty} \frac{e^{-\lambda}\lambda^n}{n!} F_Z(z)^n\right] = e^{-\lambda(1 - F_Z(z))}. \end{aligned} \tag{2}$$

Note that expression (2) allows calculating the annual maxima probability distribution function as a function of: i) the Poisson parameter associated with the

annual occurrence of storm peaks, and ii) the storm peak magnitude distribution $F_Z(z)$. Considering the asymptotic relationship between return period (T) and annual maxima given by Beran and Nozdryn-Plotnicki (1977):

$$T = -\frac{1}{\log\left(\text{Prob}\left[\max_{1 \leq i \leq N} Z_i \leq z\right]\right)}, \tag{3}$$

which improves estimates associated with returns periods lower than 10 years, and using (2), the following relationship is derived:

$$T = \frac{1}{\lambda(1 - F_Z(z_T))}. \tag{4}$$

Equation (4) allows using the RMEV model for annual return period estimation. Conversely, quantile z_T associated with given return period T is obtained by solving the following implicit equation:

$$F_Z(z_T) = 1 - \frac{1}{\lambda T}, \tag{5}$$

which can be transformed into the problem of finding the root of the function $g(z_T) = 1 - \frac{1}{\lambda T} - F_Z(z_T)$. Analogously to the MEV approach, numerical tests indicate that the algorithm proposed by Forsythe et al (1976), which uses a combination of bisection, secant, and inverse quadratic interpolation methods, is robust and efficient.

Note that MEV and RMEV methods are very similar in structure and share several characteristics:

1. The conditional regression model is the same, the only difference is the data used to perform the parameter estimation process.
2. Confidence intervals are obtained using the same methodology.
3. The same diagnostic tests and plots might be used to check the adequacy of fitted models and the hypotheses of independence, such as, Probability-probability (PP) and Quantile-quantile (QQ) plots, one-sample Kolmogorov-Smirnov test (Massey (1951)), or Ljung-Box lack-of-fit hypothesis test (Brockwell and Davis, 1991).

An important issue associated with the Poisson parameter is its estimation. Note that threshold selection is performed using reanalysis data and any of the methods proposed in the literature for this task, such as the mean residual life plot. However, instrumental data might be even below that threshold due to the discrepancies among both type of data. This fact might result in differences among Poisson parameter estimates using instrumental and reanalysis data, respectively, which could affect the return level estimates in (4). Our advice is to use the estimate given by instrumental data because this data is more reliable, however, we also recommend to use the estimate given by reanalysis data for comparative purposes, especially if the instrumental record length is below 5 years.

3 Realistic Illustrative Example

In order to show the functioning of the proposed methodology in a realistic case study, we have selected an specific location close to Bilbao Harbor (Northern coast of Spain). At this site, we have at our disposal i) hourly reanalysis significant wave height records from February 1, 1948 up to January 1, 2011, and ii) buoy instrumental records from February 21, 1985 to July 13, 2009. This example is also used in Mínguez et al (2013b). Reanalysis data is taken from Downscaled Ocean Waves (DOW) database, which constitutes a numerical wave database propagated to the Spanish coastal areas by the Environmental Hydraulics Institute "IH Cantabria" (Spain). The DOW database is a hybrid downscaling (Camus et al, 2011) from the GOW hindcast database (Global Ocean Waves, Reguero et al (2012)).

Let consider the vectors $\boldsymbol{x}$ and $\boldsymbol{x}^{\mathrm{POT}}$ to be, respectively, reanalysis significant wave heights and the corresponding storm peak exceedances over the threshold $u = 4.4915$, while $\boldsymbol{y}$ is the vector of differences between instrumental and reanalysis data for the same storms. Note that in order to ensure independence among storm peaks, we consider a minimum time span between consecutive storms of 3 days. In addition, in order to compare reanalysis and instrumental records during storms we also consider the possibility that both peaks (instrumental and reanalysis) occur within a time lag of one day. Threshold choice was informed using the mean residual life plot (Coles, 2001) and it corresponds to the percentile 99.4% of the reanalysis data.

We analyze in detail the Bilbao record using the following steps:

Step 1: Using the sample set ($\boldsymbol{x}^{\mathrm{POT}}$), we fit the Pareto distribution using the maximum likelihood method, i.e. by maximizing the log-likelihood function. The following parameter estimates and 95% confidence bounds are obtained:

$$\begin{array}{l} \hat{\psi}_x = 0.6407\ (0.5561, 0.7464) \\ \hat{\xi}_x = 0. \end{array} \tag{6}$$

This fit corresponds to the exponential case ($\hat{\xi}_x = 0$). By applying the one-sample Kolmogorov-Smirnov test with 0.05 significance level for the transformed sample $\boldsymbol{x}^{\mathrm{N}} = \Phi^{-1}\left[\hat{F}_X\left(\boldsymbol{x}_{\mathrm{POT}}\right)\right]$, the p-value obtained is 0.4542, so that the null hypothesis that the transformed sample follows a standard normal distribution is accepted. This implies that the exponential fit is appropriate. In addition, the Ljung-Box lack-of-fit hypothesis test considering the null hypothesis that no serial correlation at the lags 1, 2, and 3 storms exist has been applied on the $\boldsymbol{x}^{\mathrm{N}}$ sample. The p-values obtained for a 5% significance level are (0.6081, 0.5098, 0.2154), respectively. Note that since in all cases the p-values are higher than the significance level 0.05, the null hypothesis is accepted, which confirms the independence assumption between storm peaks.

Step 2: Using the samples ($\boldsymbol{x}^{\mathrm{POT}}, \boldsymbol{y}$) and an homoscedastic linear regression model for the conditional mean of the form $\boldsymbol{y} = \beta_1 + \beta_2 \boldsymbol{x}^{\mathrm{POT}}$ with constant standard deviation β_3, we estimate its parameters by maximizing the log-likelihood function, getting the following parameter estimates and 95% confidence bounds:

$$\begin{aligned}\hat{\beta}_1 &= -0.9406\ (-2.3922, 0.5110)\\ \hat{\beta}_2 &= 0.2050\ (-0.0756, 0.4857)\\ \hat{\beta}_3 &= 0.6512\ (0.5331, 0.7692).\end{aligned} \tag{7}$$

Figure 1 shows different diagnostic plots for the fitted regression model. Upper left panel presents the scatter plot (triangle dots), the conditional mean response (black line), 95% confidence bands for the mean response (dashed black line), and 95% confidence bands for the predicted values (dashed gray line). To check the normality assumption for studentized residuals, upper right panel shows the studentized residuals on a normal probability plot. Note that data points are aligned with the normal fit, i.e they follow a standard normal distribution. To further reinforce this statement, we perform the one-sample Kolmogorov-Smirnov test with 0.05 significance level for the studentized residuals, obtaining a p-value equal to 0.8452, i.e. the sample comes from a standard normal distribution. Finally, panels below of Figure 1 show, respectively, the autocorrelation and partial autocorrelation functions of the studentized residuals. Note that in both cases the autocorrelation and partial autocorrelation functions for different time lags are within the confidence bands, confirming that the values are uncorrelated. This is reinforced by performing the Ljung-Box lack-of-fit hypothesis test at the lags 1, 2, and 3 storms. The p-values obtained for a 5% significance level are (0.8771, 0.1171, 0.0614), respectively, and the independence hypothesis between data is accepted.

Note that the lineal term β_4 associated with the standard deviation from the heteroscedastic model has been removed because it was no statistically significant, however we did not remove parameter β_2 while also being no statistically significant at the 95% confidence level. The reason is that is is statistically significant at the 85% confidence level and the rest of diagnostic tests and plots diagnosed a good fit.

Step 3: Finally, for comparison purposes we fit the sample $\boldsymbol{z}^{\mathrm{POT}}$, which corresponds to the peaks over threshold but using the instrumental data. The following parameter estimates and 95% confidence bounds are obtained:

$$\begin{aligned}\hat{\psi}_z &= 0.8074\ (0.6662, 0.9992)\\ \hat{\xi}_z &= 0.\end{aligned} \tag{8}$$

The fit also corresponds to the exponential case ($\hat{\xi}_z = 0$). The fit is considered appropriate because the one-sample Kolmogorov-Smirnov test with 0.05 significance level, for the transformed sample $\boldsymbol{z}^{\mathrm{N}} = \Phi^{-1}\left[\hat{F}_Z(\boldsymbol{z}_{\mathrm{POT}})\right]$, allows accepting the null hypothesis. The associated p-value is 0.4001. Analogously, the Ljung-Box lack-of-fit hypothesis test considering the null hypothesis that no serial correlation at the lags 1, 2, and 3 storms exist. The p-values obtained for a 5% significance level are (0.0898, 0.19, 0.3323), respectively, confirming the independence assumption between instrumental storm peaks.

Step 4: An important issue within RMEV framework is the estimation of the number of storms per year, i.e. the Poisson parameter. It would be tempting to use the estimation from reanalysis storm peaks ($\boldsymbol{x}^{\mathrm{POT}}$), which corresponds to $\hat{\lambda} = 2.8292$, however, the method allows obtaining instrumental storm peaks

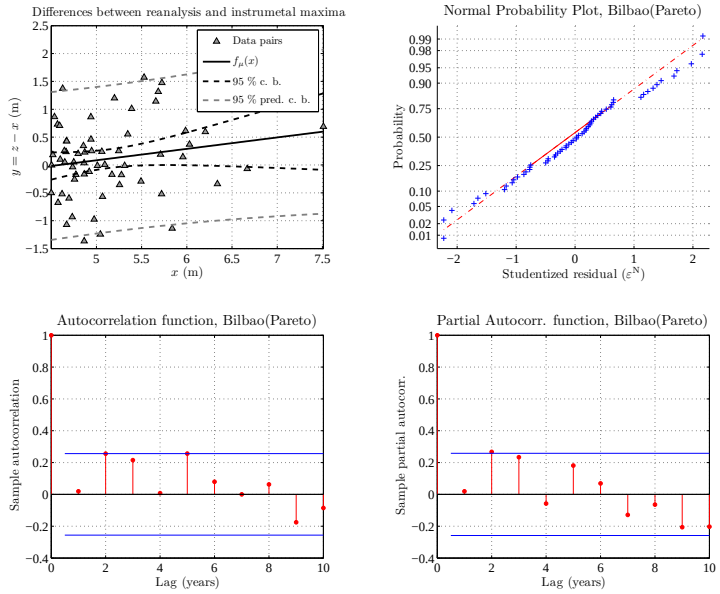

Fig. 1 Diagnostic plots for the regression fit related to Bilbao site $(\boldsymbol{x}^{\mathrm{POT}}, \boldsymbol{y})$: i) data pairs, mean values, upper and lower bounds for both expected values and predicted response, ii) normal probability plot of studentized residuals, iii) autocorrelation function of studentized residuals and iv) partial autocorrelation function of studentized residuals.

lower than the threshold used for reanalysis data, and for this reason it is preferable to use the Poisson estimation from the instrumental data, which is more realistic. In this particular case, $\hat{\lambda} = 3.8543$.

Step 5: Using the information given by the three model fitted on previous steps, we calculate the return period values using: i) reanalysis storm peaks, ii) instrumental storm peaks, and iii) reanalysis and instrumental storm peaks through the method proposed in this paper.

Results are summarized in Figure 2, where the annual return periods from the models and the data are shown using the graphical representation given by Mínguez et al (2013a). Note that we also include results from Mínguez et al (2013b) (MEV annual fit in white line and 95% confidence bands in medium gray shadow). From this figure, the following comments are pertinent:

1. The reanalysis Pareto fit (x^{POT}, medium gray line) presents good agreement with respect to data, and the confidence bands (light gray shadow) are the narrowest among all models. This result is obvious since the number of data values used for the fitting is the highest.
2. Making extreme value analysis using reanalysis data leads to under predictions of return period values of about a meter, which is not acceptable from the engineering design perspective.

3. The instrumental Pareto (z^{POT}, light gray line) underestimates wave heights associated with low return periods, while the RMEV model (black line) presents better agreement with respect to empirical data for those cases. These differences tend to decrease for wave heights related to higher return periods.
4. RMEV model provides higher significant wave heights for small return periods ($\leq$10 years) than MEV model based on annual maxima. However, they are very close to each other above 20 years return period. This result confirms that the proposed method performs better for a higher range of return periods.
5. Confidence bands for the proposed model RMEV (dark gray shadow) are always narrower than those for MEV (medium gray shadow), and are included between them. This proves that the proposed method decreases the uncertainty on return period predictions.

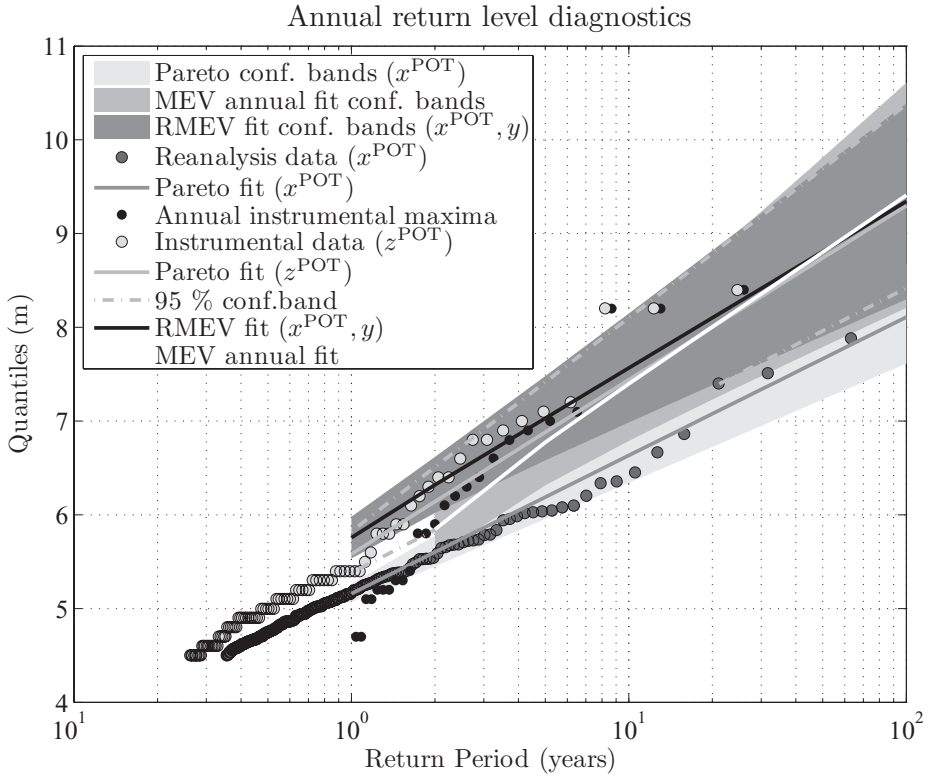

Fig. 2 Annual return period values from: i) reanalysis data (dark gray circle dots), ii) reanalysis fitted Pareto model (medium gray line), iii) instrumental annual maxima (black dots), iv) instrumental storm peaks data (light gray circle dots), v) instrumental fitted Pareto model (medium gray lines), vi) RMEV fitted model (black line), and MEV fitted model (white line) from Mínguez et al (2013b). For all models 95% confidence bands are also plotted in shadows and a dashed line for $\boldsymbol{z}^{\mathrm{POT}}$ Pareto fit.

4 Conclusions

The revisited mixed extreme value (RMEV) model has proved to be more effective than MEV, since it includes more information on the tail of the distribution. The method only requires an slightly modification of the approach presented by Mínguez et al (2013b) and the consideration of storms to follow an Poisson process during anyone year.

The major innovation of this contribution is that it allows to extend the MEV correction method for extreme wave heights to be used with storm peaks instead of annual maxima. The practical implications are clear because we can use more information about the tail distribution of wave heights, leading to more robust estimates of return periods. This fact is extremely usefull specially if short records (lower or equal to 10 years) of instrumental observations are available. Making an analogy with traditional Extreme Value Theory methods, this is equivalent to move from using the Generalized Extreme Value (GEV) distribution for annual maxima to the Peaks-Over-Threshold method, which is known among practitioners to be more convenient for extreme value analysis.

Although we have used the maximum likelihood method for estimation purposes, the Bayesian context is also perfectly applicable. The latter would modify and possibly improve results with respect to the estimation process but the method from the probability theory perspective remains unaltered. Another interesting issue is the extension of the univariate method presented in this paper for multivariate settings (multivariate data and/or multiple locations), which we believe it can be applied also on a Bayesin context (Vanem et al (2012a,b)). However, the specific application of this new method in these contexts is a subject for further research.

Acknowledgements The authors acknowledge to Puertos del Estado the availability REDCOS coastal buoy network and the Environmental Hydraulics Institute "IH Cantabria" for the reanalysis data used for this study. R. Mínguez was partly funded by the unemployment benefit of the Public Service of National Employment (SEPE) from the Spanish Ministry of Employment and Social Security. F. Del Jesus is funded by Fundación Iberdrola.

References

Beran MA, Nozdryn-Plotnicki MK (1977) Estimation of low return period floods. Bull Int Ass Hydrol Sci (2):275–282

Brockwell PJ, Davis RA (1991) Time series: Theory and methods, 2nd edn. Springer-Verlag, New York, NY

Caires S, Sterl A (2005) A new non-parametric method to correct model data: Application to significant wave height from the ERA-40 reanalysis. Journal of Atmospheric and Oceanic Technology 22:443–459

Camus P, Méndez FJ, Medina R (2011) A hybrid efficient method to downscale wave climate to coastal areas. Coastal Engineering DOI 10.1016/j.coastaleng.2011.05.007

Castillo E (1988) Extreme Value Theory in Engineering. Academic Press, New York

Castillo E, Hadi AS, Balakrishnan N, Sarabia JM (2005) Extreme Value and Related Models in Engineering and Science Applications. John Wiley & Sons, New York

Cavaleri L, Sclavo M (2006) The calibration of wind and wave model data in the mediterranean sea. Coastal Engineering 53:613–627

Coles S (2001) An introduction to statistical modeling of extreme values. Springer Series in Statistics

Davidson AC, Smith RL (1990) Models for exceedances over high thresholds. Journal of the Royal Statistical Society Series B (Methodological) 52(3):393–442

Forsythe GE, Malcolm MA, Moler CB (1976) Computer Methods for Mathematical Computations. Prentice-Hall

Katz RW, Parlange MB, Naveau P (2002) Statistics of extremes in hydrology. Advanced Water Resources 25:1287–1304

Leadbetter M, Lindgren G, Rootzén H (1983) Extremes and related properties of random sequences and processes. Springer-Verlag, New York

Massey FJ (1951) The Kolmogorov-Smirnov test for goodness of fit. Journal of the American Statistical Association 46(253):68–78

Mínguez R, Espejo A, Tomás A, Méndez FJ, Losada IJ (2011) Directional calibration of wave reanalysis databases using instrumental data. J Atmos Oceanic Technol 28:1466–1485, DOI 10.1175/JTECH-D-11-00008.1

Mínguez R, Reguero BG, Luceño A, Méndez FJ (2012) Regression models for outlier identification (hurricanes and typhoons) in wave hindcast databases. Journal of Atmospheric and Oceanic Technology 29:267–285, DOI 10.1175/JTECH-D-11-00059.1

Mínguez R, Guanche Y, Méndez FJ (2013a) Point-in-time and extreme-value probability simulation technique for engineering design. Structural Safety 41:29–36, DOI 10.1016/j.strusafe.2012.10.002

Mínguez R, Tomás A, Méndez FJ, Medina R (2013b) Mixed extreme wave climate model for reanalysis databases. Stochastic Environmental Research and Risk Assessment 27:757–768, DOI 10.1007/s00477-012-0604-y

Reguero BJ, Menéndez M, Méndez FJ, Mínguez R, Losada IJ (2012) A global ocean wave (GOW) calibrated reanalysis from 1948 onwards. Coastal Engineering 65:38–55, DOI 10.1016/j.coastaleng.2012.03.003

Shampine LF (2008) Vectorized adaptive quadrature in MATLAB. Journal of Computational and Applied Mathematics 211:131–140

Vanem E, Huseby A, Natvig B (2012a) A Bayesian hierarchical spatio-temporal model for significant wave height in the North Atlantic. Stochastic Environmental Research and Risk Assessment 26:609–632, DOI 10.1007/s00477-011-0522-4

Vanem E, Huseby A, Natvig B (2012b) Modelling ocean wave climate with a Bayesian hierarchical spacetime model and a log-transform of the data. Ocean Dynamics 62:355–375, DOI 10.1007/s10236-011-0505-5