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

Mixed extreme wave climate model for reanalysis data bases

R. Mínguez · A . Tomás · F. J. Méndez · R. Medina

Received: date / Accepted: date

Abstract Hindcast or Wave Reanalysis Data Bases (WRDB) constitute a powerful source with respect to instrumental records in the design of offshore and coastal structures, since they offer important advantages for the statistical characterization of wave climate variables, such as continuous long time records of significant wave heights, mean and peak periods, etc. However, reanalysis data is less accurate than instrumental records, making extreme data analysis derived from WRDB prone to under predict design return period values. This paper proposes a Mixed Extreme Value (MEV) model to deal with maxima, which takes full advantage of both i) hindcast or wave reanalysis and ii) instrumental records, reducing the uncertainty in its predictions. The resulting mixed model consistently merges the information given by both kinds of data sets, and it can be applied to any extreme value analysis distribution, such as GEV, POT (Peaks Over Threshold) or Pareto-Poisson. The methodology is illustrated using both synthetically generated and real data, the latter taken from a given location on the Northern Spanish coast.

Keywords design return periods · extreme waves · wave reanalysis

All authors
Environmental Hydraulics Institute "IH Cantabria", Universidad de Cantabria, Cantabria
Campus Internacional, Spain
Tel.: +34-942-201852
Fax: +34-942-201860
E-mail: roberto.minguez@unican.es

1 Introduction

Extreme value analysis is of paramount importance for the design process of coastal and offshore structures. The objective of the design is to verify that the structure satisfies the project requirements during its lifetime in terms of acceptable failure rates and costs. One of these project requirements is the ultimate limit state design, i.e. the structure must withstand the maximum stresses which are expected to occur during its lifetime. The appropriate definition of this ultimate limit state design relies on the correct evaluation of the wave climate producing the worst case scenario, i.e. on extreme wave climate analysis.

Over the last decade, in an attempt to improve the knowledge on wave climate, there has been 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. This fact has raised the attention of scientists and engineers, who have tried to use them for design purposes. However, several authors (Caires and Sterl (2005); Cavaleri and Sclavo (2006); Mínguez et al (2011, 2012)) have pointed out discrepancies when comparing reanalysis versus instrumental data. The reasons are multiple: numerical models are simplifications of reality, they have discrete spatio-temporal resolutions, temporal resolutions are too coarse (6 hours) to include high-frequency energy, the orography in certain regions is very complex, or bathymetric deficiencies among others. These discrepancies are especially relevant in shallow waters, and during the occurrence of hurricanes and typhoons, where the model does not reproduce the physics appropriately.

In order to reduce these discrepancies, several authors have attempted to combine reanalysis and instrumental observations, taking full advantage of the goodness of both types of information. For example, Caires and Sterl (2005) establish a nonparametric correction based on analogs, Cavaleri and Sclavo (2006) calibrate decadal time series over the Mediterranean Sea using buoys and satellites, Tomás et al (2008) propose a spatial calibration procedure based on empirical orthogonal functions and a non-linear transformation of the spatial-time modes, Mínguez et al

(2011) present a nonlinear regression model for directional calibration. Although these methods perform appropriately for most of the range of the probability density function of the reanalysis variables, they are not adequate for extremes. In fact, Mínguez et al (2012) demonstrate the importance of removing these extreme events from the calibration process, since on certain occasions they may distort the calibration procedure while still not solving the extreme event discrepancies. In addition, they introduce several regression models for automatic detection of these events, before removing them from the calibration process.

The statistical theory of extreme values (Castillo, 1988; Coles, 2001; Katz et al, 2002; Castillo et al, 2005) provides the mathematical framework to model the tail distribution, i.e. the extreme values, when maximum datasets are available. These models allow us to obtain useful information, such as return period values for certain variables. Several models and applications have been used in different climate studies to model block extremes, typically annual maxima or minima, both in observed and simulated data (Kharin et al, 2005; Goubanova and Li, 2007; Kioutsioukis et al, 2010; Nikulin et al, 2011). In addition, recent advances in extreme value theory allow introducing time-dependent variations in the extreme value models. In this kind of approach, parameters are replaced by different time dependent functions (Coles, 2001). In a simple setting, the parameters can include a trend term varying linearly with time (Cooley, 2009) or a forcing term varying with some external climatic indices, such as the Southern Oscillation Index or the North Atlantic Oscillation (NAO). There are also studies combining both approaches (Méndez et al, 2007). The most complex approaches consider harmonic functions reflecting the seasonality of the occurrence of maxima. For instance, Menéndez et al (2009); Izaguirre et al (2010); Mínguez et al (2010) developed a time-dependent model based on the GEV distribution which accounts for the seasonality and interannual variability of extreme wave height. A similar approach has been considered by Rust et al (2009) to model extreme precipitation in the UK on a seasonal basis. Galiatsatou and Prinos (2011) present a statistical model for extreme value analysis considering seasonality. They use a non-stationary point process approach estimated through the wavelet transform. Vanem (2011) presents a literature survey on time-dependent statistical modelling of extreme waves and sea states.

The main problem, from an engineering design perspective, is that neither i) any of the previous calibration/correction approaches, nor ii) the extreme value analysis models proposed in the literature, provide an answer on how to deal with extreme events appropriately in the case of reanalysis and instrumental data. The aim of this paper is to fill this niche by presenting a general method to deal with extremes that takes full advantage of both i) hindcast or wave reanalysis, and ii) instrumental records. The resulting model consistently merges the information given by both kinds of data sets, reducing the uncertainty of its predictions. In addition, it can be applied to any extreme value analysis distribution, such as GEV, POT or Pareto-Poisson.

The paper is organized as follows. Section 2 presents the proposed Mixed Extreme Value model, while in Section 3 several diagnostic tests are given to check the appropriateness of the model hypothesis. In Section 4, the functioning of the method is illustrated through two different simulation experiments. In contrast, Section 5 shows the performance on real data from a given location in the North of Spain. Finally, in Section 6 relevant conclusions are drawn.

2 Mixed Extreme Value Analysis Model

Extreme value analysis studies the occurrence of extreme events and their frequency, and a careful analysis requires the availability of data on such extremes (Castillo et al, 2008). The larger the size of the data record, the more accurate the statistical model for those extremes will be, which leads to better predictions with lower uncertainties. Vanem (2011) points out the importance of available wave data in order to develop adequate probabilistic models, and although buoy measurements are generally regarded as the most reliable, alternatives exist in satellite

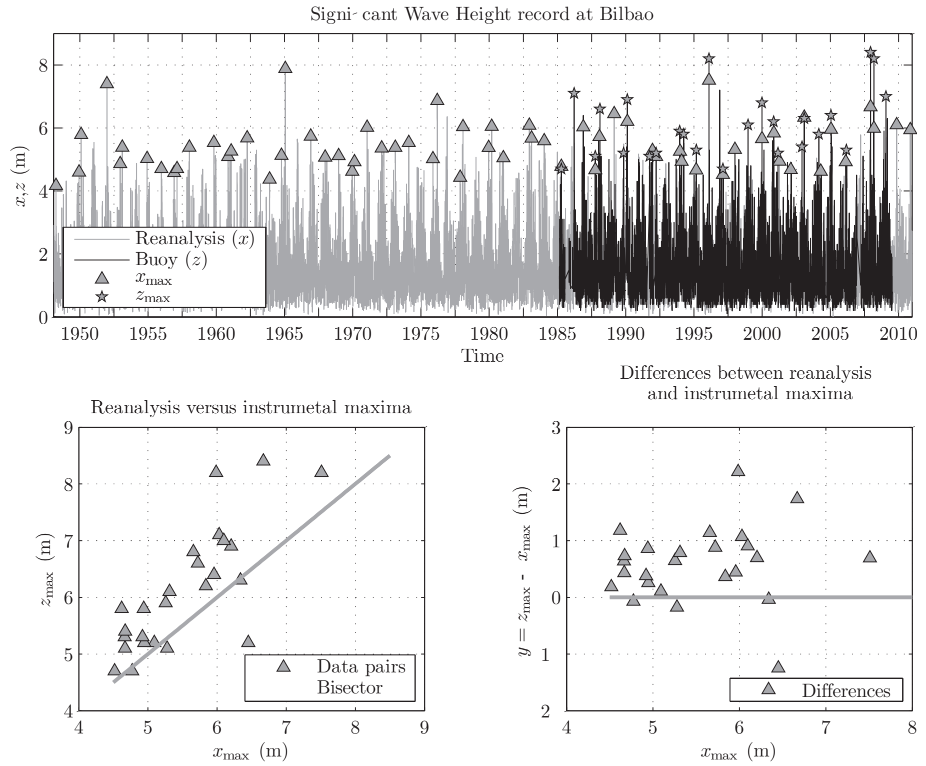

Fig. 1 Instrumental and reanalysis significant wave height records for Bilbao buoy location.

data and in reanalysis data obtained from wave models forced by various meteorological parameters. However, discrepancies between numerical and instrumental records must be accounted for within the analysis.

Figure 1 shows the significant wave height instrumental (black line, taken from Puertos del Estado (Spain) buoys database) and reanalysis (gray line) records for the Bilbao (Spain) buoy location, and their corresponding annual maxima (star and triangle dots, respectively). Reanalysis data is taken from the Downscaled Ocean Waves (DOW) database, a numerical wave database propagated to the Spanish coastal areas by the Environmental Hydraulics Institute (Spain). The DOW database is a hybrid downscaling (Camus et al, 2011) obtained using the GOW hindcast database (Global Ocean Waves, Reguero et al (2012)).

The most reliable and accurate maxima corresponds to instrumental data, in fact, for years where both maxima are available, the differences between them are shown in the panels below. These panels represent the reanalysis versus instrumental maxima, and reanalysis maxima versus the differences between them, respectively. If Extreme Value (EV) analysis were performed using instrumental and reanalysis data, respectively, the instrumental EV model would be more reliable with respect to expected return periods, however the uncertainty would increase compared to that of the reanalysis EV model, because the latter uses more data. The model presented in this paper allows using both instrumental and reanalysis data, which results in a more robust estimation of return period values, decreasing the uncertainty.

Let the maximum annual records from reanalysis ($\boldsymbol{x}_{\max}$) and instrumental ($\boldsymbol{z}_{\max}$) have lengths n_x and n_z, respectively. Note that we assume that $n_z << n_x$. For years where both maxima are available, we obtain the vector $\boldsymbol{y}_{\max}$ as the differences between instrumental and reanalysis maxima. The proposed mixed model relies on the following assumptions:

1. The annual maximum reanalysis random variable X is assigned a probability density function $f_X(x, \boldsymbol{\theta}_X)$ and a cumulative distribution function $F_X(x, \boldsymbol{\theta}_X)$. The distribution function may correspond to any type of distribution for maxima, such as, GEV, Pareto-Poisson, Gumbel, POT, etc.
2. The random variable Y corresponding to the difference between instrumental and reanalysis data conditioned to the reanalysis maximum data (X) follows a normal distribution, i.e. $f_{Y|X}(y) \sim N\left(\mu_{Y|X}, \sigma^2_{Y|X}\right)$. Note that $\mu_{Y|X}$ and $\sigma_{Y|X}$ correspond to the conditional mean and standard deviation parameters, which can be obtained using an heteroscedastic regression model.

The annual maximum instrumental random variable is equal to $Z = X + Y$, and their corresponding cumulative distribution function is equal to:

$$F_Z(z) = \text{Prob}(Z \leq z) = \int\limits_{x+y \leq z} f_{X,Y}(x, y) dy dx, \tag{1}$$

where $f_{X,Y}(x,y)$ is the joint probability density function of the random variables X and Y. Considering assumptions 1 and 2, expression (1) becomes:

$$F_Z(z) = \int_{-\infty}^{\infty} f_X(x, \boldsymbol{\theta}_X) \left[\int_{-\infty}^{z-x} f_{Y|X}(y) dy \right] dx, \tag{2}$$

and since the distribution of Y conditioned to X is normally distributed, expression (2) results in:

$$F_Z(z) = \int_{-\infty}^{\infty} f_X(x, \boldsymbol{\theta}_X) \Phi \left[\frac{z - x - \mu_{Y|X}}{\sigma_{Y|X}} \right] dx, \tag{3}$$

where $\Phi(\cdot)$ is the cumulative distribution of the standard normal random variable.

The corresponding probability density function is obtained differentiating (3) with respect to z:

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x, \boldsymbol{\theta}_X) \phi \left[\frac{z - x - \mu_{Y|X}}{\sigma_{Y|X}} \right] \frac{1}{\sigma_{Y|X}} dx, \tag{4}$$

where $\phi(\cdot)$ is the probability density function of the standard normal random variable. Note that the integration limits range from $-\infty$ to ∞, however, these limits may change depending on the type of X probability density function.

Both PDF (4) and CDF (3) require solving an integral over a varying domain. This task can be efficiently achieved using numerical quadrature methods. Numerical tests performed using different methods, indicate that 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.

The corresponding quantile z_q for a given probability q is obtained by solving the following implicit equation:

$$F_Z(z_q) = q, \tag{5}$$

which can be transformed into the problem of finding the root of the function $g(z_q) = q - F_Z(z_q)$. 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.

An important feature of any EV model corresponds to the calculation of the confidence intervals on both model parameters and estimates, i.e. uncertainty. Since the proposed mixed model depends on two independent distributions, prior to analyzing confidence intervals related to Z, we will briefly describe how to deal with $f_X(x)$ and $f_{Y|X}(y)$ distributions, and their confidence intervals. Finally, estimated parameters and quantile confidence intervals for the proposed model are given in Appendixes A and B respectively.

2.1 Reanalysis annual maxima distribution ($f_X(x)$)

One of the advantages of the proposed mixed model (MEV) is that annual maxima can be analyzed using an extreme value analysis distribution. In this paper, we only provide expressions and examples for the Generalized Extreme Value (GEV) distribution and the Pareto-Poisson model (Castillo, 1988).

2.1.1 GEV Distribution

Within this approach, annual maxima of successive years are assumed to be i) independent random variables and ii) identically distributed. Annual maximum X of the climate variable follows a GEV distribution with time-dependent location parameter μ, scale parameter ψ, and shape parameter ξ, with a cumulative distribution function (CDF) given by:

$$F_X(x;\mu,\psi,\xi) = \begin{cases} \exp\left\{ -\left[1+\xi\left(\dfrac{x-\mu}{\psi}\right)\right]_+^{-\dfrac{1}{\xi}} \right\}; \xi \neq 0, \\ \\ \exp\left\{ -\exp\left[-\left(\dfrac{x-\mu}{\psi}\right)\right]\right\}; \xi = 0, \end{cases} \tag{6}$$

where $[a]_+ = \max(0, a)$, and the support is $x \leq \mu - \psi/\xi$, if $\xi < 0$, or $x \geq \mu - \psi/\xi$, if $\xi > 0$. The GEV family includes three distributions corresponding to the different types of tail behavior: Gumbel ($\xi = 0$) with a light tail decaying exponentially; Fréchet distribution ($\xi > 0$) with a heavy tail decaying polynomially; and Weibull ($\xi < 0$) with a short tail.

From (6), the corresponding quantiles x_q can be straightforwardly calculated, where q is the corresponding probability.

Model parameters $\boldsymbol{\theta}_X = (\mu, \psi, \xi)^T$ may be estimated using the method of maximum likelihood.

2.1.2 Pareto-Poisson Distribution

This model combines the Generalized Pareto Distribution (GPD) for studying exceedances over a threshold u, and the Poisson distribution for occurrence of exceedances. It is based on the following assumptions:

1. The number of exceedances over the level u during the year has a Poisson distribution with parameter λ.
2. Those exceedances follow the GPD distribution.

Under these hypothesis, the cumulative probability distribution (CDF) of the annual maximum can be expressed as:

$$F_X(x; \lambda, \psi, \xi) = \begin{cases} \exp\left\{-\lambda\left[1+\xi\left(\dfrac{x-u}{\psi}\right)\right]_+^{-\dfrac{1}{\xi}}\right\}; \xi \neq 0, \\ \\ \exp\left\{-\lambda \exp\left[-\left(\dfrac{x-u}{\psi}\right)\right]\right\}; \xi = 0, \end{cases} \tag{7}$$

where $[a]_+ = \max(0, a)$, and the support is $x > u$, $x \leq \psi/|\xi|$ if $\xi < 0$, or $x \leq \infty$ if $\xi > 0$. The Pareto-Poisson family includes, as does the Pareto family, three distributions corresponding to the different types of tail behavior: Exponential ($\xi = 0$); traditional Pareto tail ($\xi > 0$), and an analogous Weibull distribution ($\xi < 0$) with a bounded tail.

Analogously to the GEV model, quantiles and parameter estimates can be obtained from (7) and the method of maximum likelihood, respectively.

2.2 Heteroscedastic regression model ($f_{Y|X}(y)$)

Consider the standard nonlinear regression model

$$\boldsymbol{y} = f_\mu(\boldsymbol{x}, \boldsymbol{\beta}_\mu) + \boldsymbol{\varepsilon}, \tag{8}$$

where $\boldsymbol{y} = (y_1, y_2, \dots, y_{n_y})^T$ is the $n_y \times 1$ response variable vector associated with the differences between instrumental and reanalysis maxima, $\boldsymbol{x}$ is a $n_y \times 1$ vector of predictor variables related to annual reanalysis maxima, the function f_μ is known and nonlinear in the parameter vector $\boldsymbol{\beta}_\mu$, and $\varepsilon_i; i = 1, \dots, n_y$ are jointly normally distributed $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{V})$ errors, where $\sigma^2\boldsymbol{V}$ is a positive definite variance-covariance matrix.

In the standard Nonlinear Least Square (NLS) method, the parameter estimation problem can be stated as:

$$\underset{\boldsymbol{\beta}}{\text{Minimize}} \; \boldsymbol{\varepsilon}^T(\sigma^2\boldsymbol{V})^{-1}\boldsymbol{\varepsilon}. \tag{9}$$

However, for the kind of data considered in this regression model, a simple scatter plot of differences (y) versus reanalysis data (x) allows observing how the variance of the regression model may change over the regression function (Mínguez et al, 2012). Consequently, we consider a nonlinear heteroscedastic regression model in which the standard deviation σ_i of the ith error is a function of the predictor variable (x_i):

$$\sigma_i = f_\sigma(x_i; \boldsymbol{\beta}_\sigma), \tag{10}$$

where $\boldsymbol{\beta}_\sigma$ is a vector of coefficients or parameters. The parameter vector $\boldsymbol{\beta} = [\boldsymbol{\beta}_\mu; \boldsymbol{\beta}_\sigma]$, of size $n_p \times 1$, can be estimated maximizing the log-likelihood function:

$$\ell(\boldsymbol{\beta}; \boldsymbol{x}, \boldsymbol{y}) = -\sum_{i=1}^{n_y} \log\left(f_\sigma(x_i; \boldsymbol{\beta}_\sigma)\right) - \frac{1}{2}\sum_{i=1}^{n_y}\left(\frac{y_i - f_\mu(x_i; \boldsymbol{\beta}_\mu)}{f_\sigma(x_i; \boldsymbol{\beta}_\sigma)}\right)^2. \tag{11}$$

The advantage of defining the regression model in a general manner is that it allows using different parameterizations for the mean and standard deviation. For instance, possible models are:

$$f_\mu(x_i, \boldsymbol{\beta}_\mu) = \beta_0 + x_i\beta_1; \;\; f_\sigma(x_i, \boldsymbol{\beta}_\sigma) = \beta_2 + x_i\beta_3, \tag{12}$$

$$f_\mu(x_i, \boldsymbol{\beta}_\mu) = \beta_0 x_i^{\beta_1}; \;\; f_\sigma(x_i, \boldsymbol{\beta}_\sigma) = \beta_2 x_i^{\beta_3}, \tag{13}$$

but different expressions for f_μ and f_σ could be used instead, for instance, f_μ be a linear and f_σ an exponential function. Note that both models in (12) and (13) include the classical homoscedastic linear regression model provided that $\beta_3 = 0$.

3 Hypothesis testing

The mixed model (MEV) proposed in this paper is based on several assumptions. For this reason, once the parameter estimation processes for both the EV model over x and the regression model over y conditional on x, are finished, it is very important to make and run different diagnostic plots and statistical hypothesis tests to check whether the selected distributions are appropriate or not.

3.1 EV model for X

Related to the EV model fitted to the reanalysis data, we use the following diagnostic plots and tests:

- Probability-probability (PP) and Quantile-quantile (QQ) plots. Points over the diagonal are indicative of a good quality fit.

- A one-sample Kolmogorov-Smirnov test (Massey (1951)). This test compares, for a given significance level α, the transformed values $\boldsymbol{x}^{\mathrm{N}}$ using transformation $\boldsymbol{x}^{\mathrm{N}} = \Phi^{-1}\left[F_X\left(\boldsymbol{x}_{\max}\right)\right]$ with respect to a standard normal distribution. The null hypothesis is that the transformed sample follows a standard normal distribution.
- Sample autocorrelation and partial autocorrelation functions related to the transformed sample $\boldsymbol{x}^{\mathrm{N}}$. These functions help checking the independence assumption between maxima. Their values should be within the confidence bounds.
- The Ljung-Box lack-of-fit hypothesis test (Brockwell and Davis, 1991) to further explore the independence hypothesis. This test is applied to study the model misidentification. It indicates the acceptance or not of the null hypothesis that the model fit is adequate (no serial correlation at the corresponding element of Lags).

3.2 Regression model for $Y|X$

Related to the heteroscedastic conditional regression model, the basic assumption for this model to be considered appropriate, is that studentized residuals are independent and normally distributed. Studentized residuals are computed as:

$$\hat{\varepsilon}_i^{\mathrm{N}} = \frac{\hat{\varepsilon}_i}{\sqrt{\Omega_{i,i}}} = \frac{y_i - f_\mu(x_i;\, \hat{\boldsymbol{\beta}}_\mu)}{\sqrt{\Omega_{i,i}}} \quad i = 1, \ldots, n_y, \tag{14}$$

where $\Omega_{i,i}$ is the ith diagonal element of the residual variance-covariance matrix $\boldsymbol{\Omega}$. Details on the derivation of this matrix can be found in Mínguez et al (2012).

According to this, we use the following diagnostic plots and tests:

- A one-sample Kolmogorov-Smirnov test (Massey (1951)), for a given significance level α, to check that studentized residuals follow a standard normal distribution.
- Sample autocorrelation and partial autocorrelation functions related to studentized residuals.

- To further explore the independence hypothesis, the Ljung-Box lack-of-fit hypothesis test (Brockwell and Davis, 1991) for model misidentification is applied.

In case any test allows rejecting the null hypothesis with a given significance level, the probability distribution assumptions must be revisited before accepting the model for return level predictions.

Additional or alternative tests to those selected above could be applied instead.

4 Simulation Case Study

Prior to the application of the proposed method to a realistic case, we will perform a simulation study to check whether the method provides consistent results when the data follow the required assumptions. This step is very important to increase the confidence in the model and compare the results achieved with those obtained using traditional methods. We consider two different simulated samples with the following characteristics:

Case 1: The reanalysis simulated sample follows a GEV distribution with parameters $\boldsymbol{\theta}_X^{\text{true}} = (10, \exp(0.5), -0.15)^T$, and the heteroscedastic model corresponds to that given in (12) with parameters $\boldsymbol{\beta}^{\text{true}} = (-0.5, 0.7, -0.3, 0.1)^T$. The samples $\boldsymbol{x}_1^{\max}$ and $\boldsymbol{y}_1$ have $n = 1000$ records each, corresponding to a simulated period of 1000 years.

Case 2: The reanalysis simulated sample follows a Pareto-Poisson distribution with parameters $\boldsymbol{\theta}_X^{\text{true}} = (25, \exp(-0.13), -0.05)^T$, where the first parameter corresponds to the expected number of exceedances per year, i.e. λ, and the remainder to the GPD distribution parameters. The threshold is equal to u $=$ 2.5. Regarding the heteroscedastic model, it also corresponds to that given in (12) with parameters $\boldsymbol{\beta}^{\text{true}} = (0.16, 0.04, 0.3, 0.06)^T$. In order to obtain 1000 years of maximum data, we sample $n = 1000 \times \lambda^{\text{true}} = 25000$ records of exceedances using the given GPD distribution, which constitutes the sample $\boldsymbol{x}_2$. The maximum of each year constitutes the sample $\boldsymbol{x}_2^{\max}$, and the associated difference with respect to instrumental data corresponds to $\boldsymbol{y}_2$.

With both sample records $(\boldsymbol{x}_1^{\max}, \boldsymbol{y}_1)$ and $(\boldsymbol{x}_2, \boldsymbol{x}_2^{\max}, \boldsymbol{y}_2)$, we proceed to perform the following steps:

1. We fit samples $\boldsymbol{x}_1^{\max}$ and $\boldsymbol{x}_2$ to GEV and Pareto distributions, respectively, by maximizing the log-likelihood function.
2. We fit samples $(\boldsymbol{x}_1^{\max}, \boldsymbol{y}_1)$ and $(\boldsymbol{x}_2^{\max}, \boldsymbol{y}_2)$ to the conditional regression model by maximizing the log-likelihood function (11). It is important to emphasize that, as expected, in all previous fits the true parameter values were all within the estimated 95% confidence bounds.
3. Additionally, we fit each instrumental sample, i.e. $\boldsymbol{z}_1^{\max} = \boldsymbol{x}_1^{\max} + \boldsymbol{y}_1$ and $\boldsymbol{z}_2^{\max} = \boldsymbol{x}_2^{\max} + \boldsymbol{y}_2$, to the GEV distribution. This allows performing the traditional analysis. Note that for the second case, related to the Pareto-Poisson model, we also use a GEV model because the shape parameter of the Pareto distribution fit is lower than $-1/2$, and therefore, even though the maximum likelihood estimators are generally obtainable, they do not have the standard asymptotic properties.
4. Using the three fitted models for each case, the quantiles and their 95% confidence intervals related to i) the reanalysis (X) data fit, ii) instrumental (Z) data fit using the proposed model, and iii) instrumental (Z) data fit using the GEV fit, are calculated.

Results are shown in Figures 2 and 3, where the following observations are pertinent:

1. The three models related to case 1 provide very good fits to simulated data, an expected result due to the length of the data samples. In case 2, both the Pareto-Poisson and the proposed models provide very good fits to simulated data, however, the GEV fit to instrumental maxima is not appropriate for return periods longer than 20 years.
2. For case 1, both the GEV and MEV fits provide the same return levels, however, the MEV fit is slightly better in terms of uncertainty, presenting lower confidence intervals.

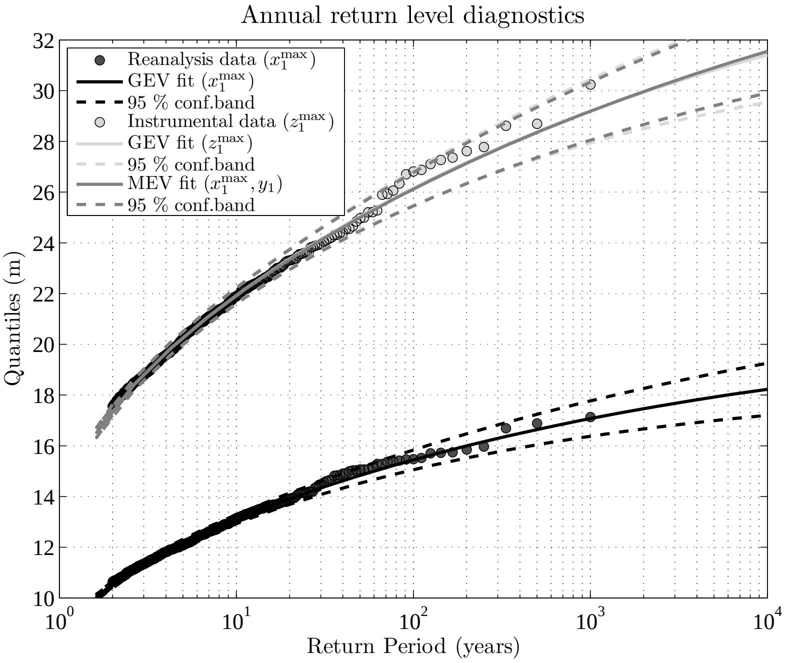

Fig. 2 Simulated data, fitted quantiles and their 95% confidence intervals related to i) the reanalysis (X) data using the GEV model, ii) instrumental (Z) data using the MEV model, and iii) instrumental (Z) data also using the GEV model.

Simulation results for both cases demonstrate the adequate functioning of the MEV model which, besides providing consistent results with respect to traditional Extreme Value analysis methods, decreases the uncertainty in model predictions. Nevertheless, the MEV will be further tested using real data.

5 Realistic Illustrative Example

In order to show the performance of the proposed methodology in a realistic case study, we have selected the record previously shown in Section 2, i.e. a 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 1st, 1948 up to January 1st, 2011, and ii) instrumental buoy records from February 21st, 1985 to July 13th, 2009. Both records are shown in Figure 1.

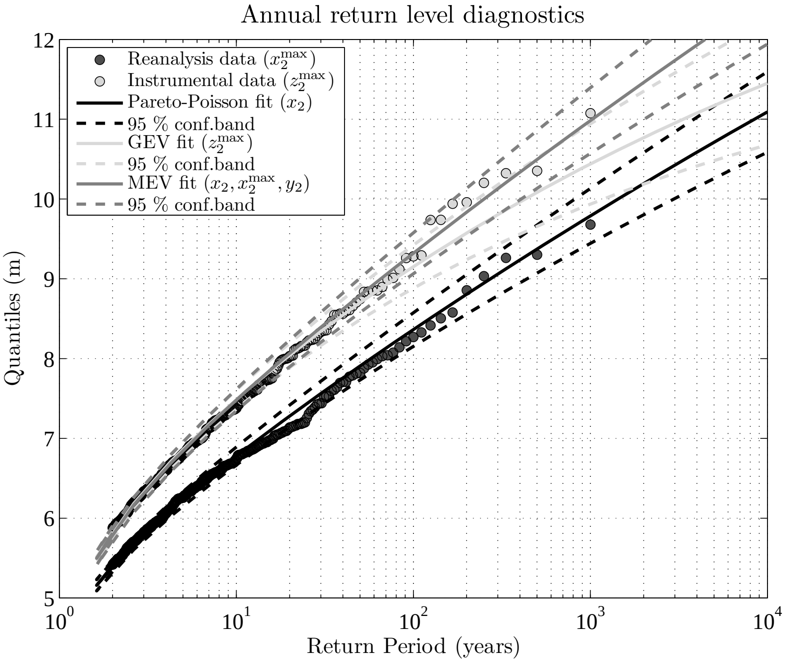

Fig. 3 Simulated data, fitted quantiles and their 95% confidence intervals related to i) the reanalysis (X) data using the Pareto-Poisson model, ii) instrumental (Z) data using MEV model, and iii) instrumental (Z) data using the GEV model.

We analyze in detail the Bilbao record using as EV model the GEV. Suppose the vectors $\boldsymbol{x}$, $\boldsymbol{x}^{\max}$, $\boldsymbol{z}^{\max}$, and $\boldsymbol{y}$ to be the reanalysis significant wave height records, the corresponding annual maxima, the instrumental annual maxima, and the differences between $\boldsymbol{x}^{\max}$ and $\boldsymbol{z}^{\max}$ for those years where we have both records. Using this information, we proceed to perform the following steps:

Step 1: Using the sample set $\boldsymbol{x}^{\max}$, we fit the GEV distribution using the maximum likelihood method. The following parameter estimates and 95% confidence bounds are obtained:

$$\begin{aligned}\hat{\mu}_x &= 5.1046\ (4.9497, 5.2596)\\ \hat{\psi}_x &= \exp(-0.5173)\ (\exp(-0.7125), \exp(-0.3222))\\ \hat{\xi}_x &= 0.\end{aligned} \tag{15}$$

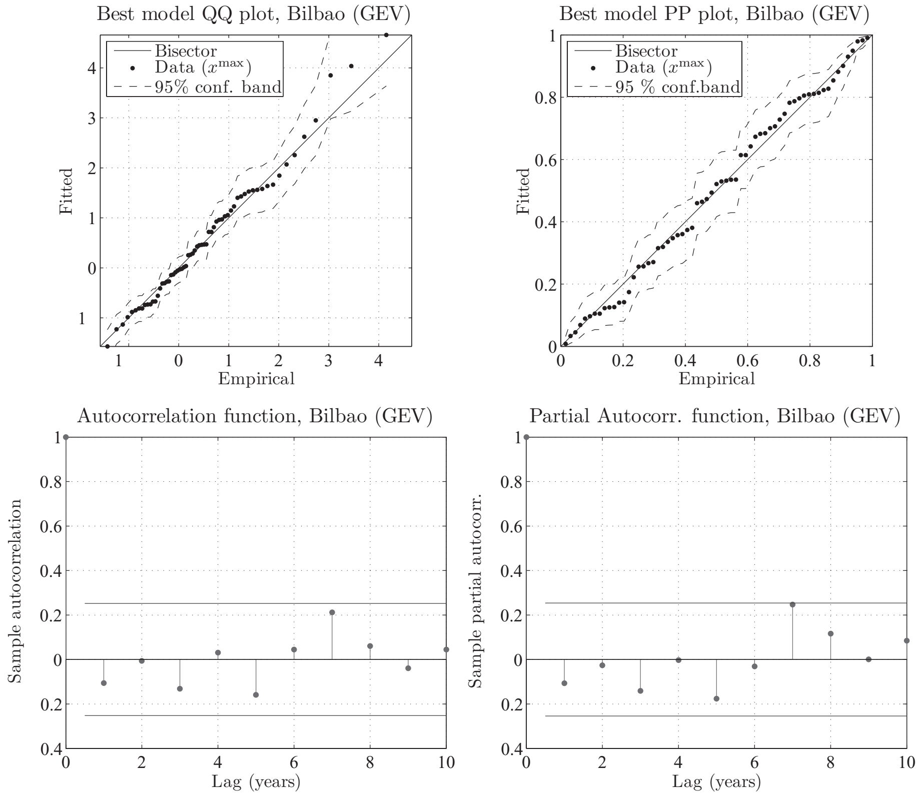

Fig. 4 Diagnostic plots for the GEV model fitted to Bilbao site reanalysis maxima ($\boldsymbol{x}^{\max}$): i) PP plot, ii) QQ plot, iii) autocorrelation and iv) partial autocorrelation functions.

Note that since the shape parameter $\hat{\xi}_x$ within the GEV model was not statistically significant, we fitted the data to the Gumbel distribution ($\hat{\xi}_x = 0$), and thus there are no confidence interval estimates. In Figure 4 several diagnostic plots of the fitting are shown. Note that PP and QQ plots (panels above) present good diagnostic statistics with points close to the diagonal. In addition, we apply the one-sample Kolmogorov-Smirnov test with 0.05 significance level for the transformed sample $\boldsymbol{x}^{\mathrm{N}} = \Phi^{-1}\left[\hat{F}_X(\boldsymbol{x}_{\max})\right]$. Note that the p-value obtained is 0.9410, and therefore accepting the null hypothesis by which the transformed sample follows a standard normal distribution. This implies that the Gumbel model is appropriate. In addition, Figures 4(iii, iv) show, respectively, the autocorrelation and partial autocorrelation functions of the transformed values. Note that in both cases the autocorrelation and partial autocorrelation functions for different time lags are within or close to the confidence bands, confirming that the values are uncorrelated. Finally, the Ljung-Box lack-of-fit hypothesis test, considering the null hypothesis that no serial correlation at the lags 1, 2, 3, 4, and 5 years exist, has been applied on the $\boldsymbol{x}^{\mathrm{N}}$ sample. The p-values obtained for a 5% significance level are (0.3896, 0.6897, 0.5897, 0.7386, 0.5840), respectively. Note that since the p-values are higher than the significance level in all the studied cases, the null hypothesis is accepted, hence confirming the independence assumption for reanalysis annual maxima.

Step 2: Using the samples $(\boldsymbol{x}^{\max}, \boldsymbol{y})$ and assuming model (12) for the conditional mean and standard deviations, we fit the regression model by maximizing the log-likelihood function (11), obtaining the following parameter estimates and 95% confidence bounds:

$$\begin{aligned}
\hat{\beta}_1 &= -0.0219\ (-1.8532, 1.8094)\\
\hat{\beta}_2 &= 0.1111\ (-0.2482, 0.4705)\\
\hat{\beta}_3 &= -0.9966\ (-2.1516, 0.1585)\\
\hat{\beta}_4 &= 0.2894\ (0.0608, 0.5179)
\end{aligned} \tag{16}$$

Figure 5 shows different diagnostic plots for the regression model fitted. Figure 5 (i) 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 given by (14), Figure 5 (ii) shows the studentized residuals on a normal probability plot. Note that data points are aligned with the normal fit, that is, they follow a standard normal distribution. To further reinforce this statement, we perform the one-sample Kolmogorov-Smirnov test with 5% significance level for the studentized residuals, obtaining a p-value equal to 0.9967, meaning that the sample comes from a standard normal distribution. Finally, Figures 5 (iii, iv) show, respectively, the autocorrelation and partial

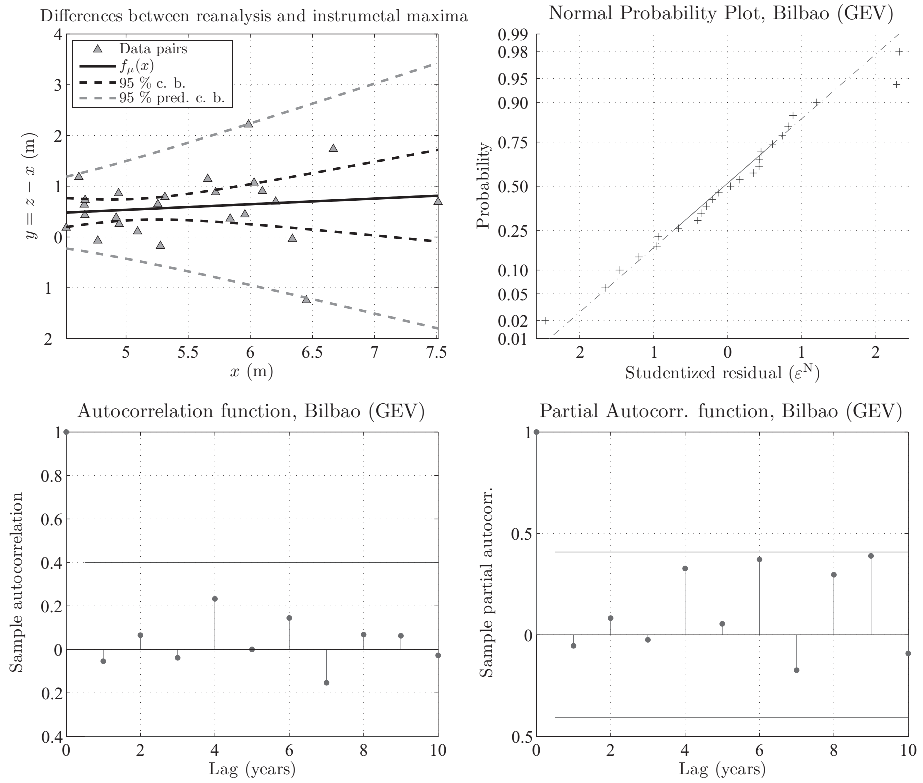

Fig. 5 Diagnostic plots for the regression model fitted to Bilbao site $(\boldsymbol{x}^{\max}, \boldsymbol{y})$: i) data pairs, mean values, upper and lower bounds for both expected values and predicted response, ii) normal probability plot of studentized residuals, iii) autocorrelation function of studentized residuals and iv) partial autocorrelation function of studentized residuals.

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 , 2, 3, 4, and 5 lag years. The p-values obtained for a 5% significance level are (0.7730, 0.9016, 0.9686, 0.7379, 0.8508) respectively, and the independence hypothesis of the data is accepted.

Step 3: Finally, for comparison purposes we fit the sample $\boldsymbol{z}^{\max}$ to the GEV distribution. The following parameter estimates and 95% confidence bounds are obtained:

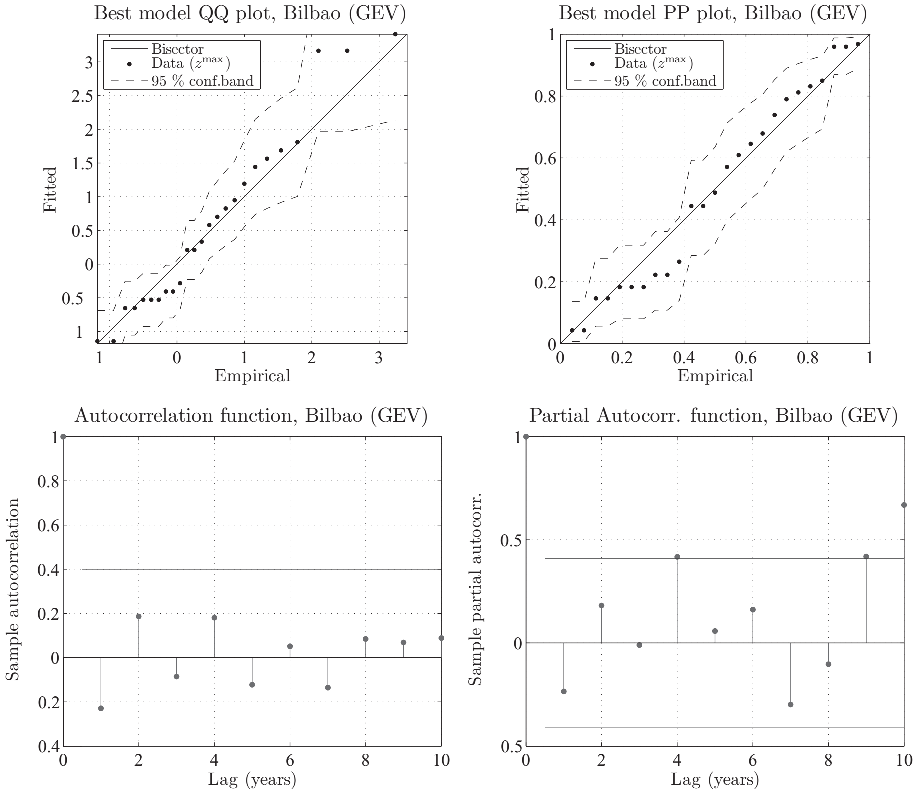

Fig. 6 Diagnostic plots for the GEV model fitted to Bilbao site instrumental maxima ($\boldsymbol{x}^{\max}$): i) PP plot, ii) QQ plot, iii) autocorrelation function and iv) partial autocorrelation function.

$$\begin{aligned}
\hat{\mu}_z &= 5.6301 \ (5.2956, 5.9646) \\
\hat{\psi}_z &= \exp(-0.2090) \ (\exp(-0.52745), \exp(0.1094)) \\
\hat{\xi}_z &= 0.
\end{aligned} \tag{17}$$

This fit also corresponds to the Gumbel case ($\hat{\xi}_z = 0$), as no confidence bands for the shape parameters exist. In Figure 6, analogous to the reanalysis maxima fit, several diagnostic plots of the fitting are shown. The fit is considered good because the one-sample Kolmogorov-Smirnov test with 5% significance level for the transformed sample $\boldsymbol{z}^{\mathrm{N}} = \Phi^{-1}\left[\hat{F}_Z(\boldsymbol{z}_{\max})\right]$ allows accepting the null hypothesis. The associated p-value is 0.6830. Analogously, the autocorrelation and partial autocorrelation functions of the transformed values, shown in Figures 6 (iii, iv), indicate that the values are reasonably uncorrelated. This result is confirmed using the Ljung-Box lack-of-fit hypothesis test considering the null hypothesis that no serial correlation exists at 1, 2, 3, 4, and 5 lag years. The p-values obtained for a 5% significance level are (0.2243, 0.2877, 0.4378, 0.4377, 0.5101), respectively, confirming the independence assumption for instrumental annual maxima.

Step 4: Using the information given by the three fitted models, we calculate the return levels using: i) reanalysis maxima information, ii) instrumental maxima information, iii) reanalysis and instrumental maxima through the method proposed in this paper.

Results are summarized in Figure 7, where return level estimates from the models and the data are shown. From this figure, the following comments are pertinent:

1. The reanalysis fit (GEV(x), black line) presents good agreement with the data, and the confidence bands are the narrowest among the three models. This result is obvious since the number of data values used for the fitting is the highest.
2. Performing extreme value analysis using reanalysis data may lead to under predictions of return level estimates. For the Bilbao case study, it varies between 0.5 and 2 meters depending on the return level, being 1 meter for the 10-year return period. This is not acceptable from an engineering design perspective.
3. Both the instrumental (GEV(z)) and the proposed model (MEV(z)) fits, are very close to each other, presenting slight differences. Most of the data are within the confidence bands. Note that both models present the same return level estimates for all return periods longer than 30 years.
4. Confidence bands for the proposed model (MEV(z)) are always narrower than those for the instrumental fit (GEV(z)), and are included between them. This proves that the proposed method decreases uncertainty in return level predictions.

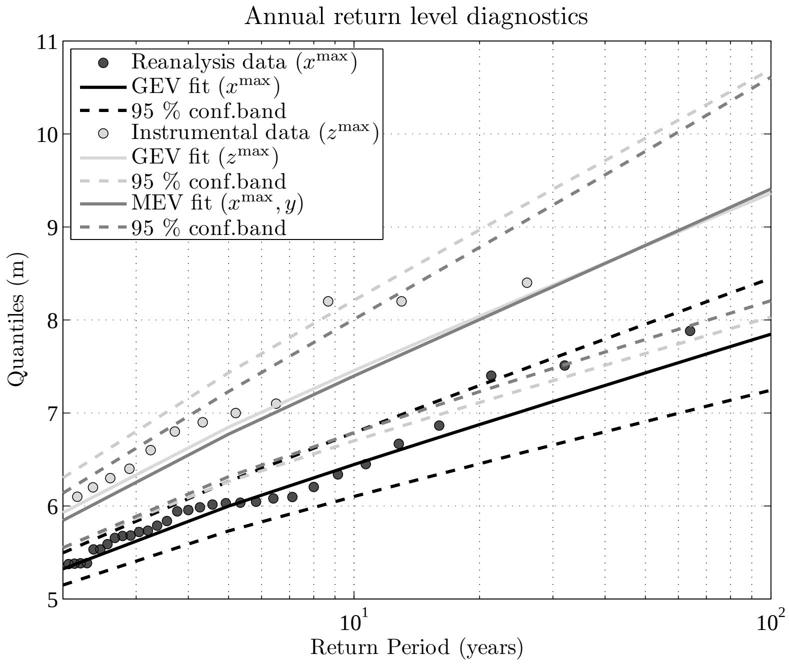

Fig. 7 Return level estimates from: i) reanalysis data, ii) instrumental data, iii) the GEV model fitted to reanalysis data, iv) the GEV model fitted to instrumental data and v) the MEV fitted model using reanalysis and instrumental data. For all models, 95% confidence bands are also plotted.

6 Conclusions

The model proposed in this paper allows performing extreme value analysis by merging together reanalysis and instrumental data. The proposed method has the following characteristics:

1. The model is supported by probability distribution and non-linear regression theory.
2. The hypothesis required to consider the proposed mixed model to be valid for EV analysis are properly established. In addition, several diagnostic plots and hypothesis tests are proposed to check whether these assumptions are certain by the data.

3. Numerical cases prove that the use of the proposed procedure provides more accurate return level estimates, thereby reducing uncertainty.
4. The model is very flexible, not only in terms of the marginal EV probability density function selected for the reanalysis maxima study, but also for the regression model, which may deal with homoscedastic, heteroscedastic, linear and nonlinear models.

Although the method is useful, especially for engineering design on specific locations, it would also be interesting to include spatial variability (Vanem et al, 2012a,b). This is a subject for further research.

Acknowledgements This work was partly funded by project "AMVAR" (CTM2010-15009) from Spanish Ministry MICINN, C3E (200800050084091) from the Spanish Ministry, MAMRM and project MARUCA (E17/08) and OCTOPOS from the Spanish Ministry MF. R. Mínguez is also indebted to the Spanish Ministry MICINN for the funding provided within the "Ramon y Cajal" program. The authors acknowledge Puertos del Estado for the availability of REDCOS coastal buoy network used in this study.

A Estimated parameter confidence intervals

The estimates $\hat{\boldsymbol{\theta}}_X$ and $\hat{\boldsymbol{\beta}}$ that maximize the log-likelihood functions of the selected EV distribution for reanalysis and (11), respectively, can be obtained using any of the available solvers for nonlinear programming. For specific details about the heteroscedastic model and algorithms to solve (11) see Mínguez et al (2012).

The estimated parameters $\hat{\boldsymbol{\theta}}_X$ and $\hat{\boldsymbol{\beta}}$ correspond to mean values, and assuming that observational errors are normally distributed, the estimated parameter vectors are distributed as follows:

$$\boldsymbol{\theta}_X \sim N\left(\hat{\boldsymbol{\theta}}_X, \Sigma_{\boldsymbol{\theta}_X}\right); \quad \boldsymbol{\beta} \sim N\left(\hat{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}}\right), \tag{18}$$

where N denotes the multivariate normal distribution, and $\Sigma_{\boldsymbol{\theta}_X}$ and $\Sigma_{\boldsymbol{\beta}}$ are the variance-covariance matrices of the parameter estimates. Using the method of maximum likelihood, if $\ell(\cdot)$ is twice differentiable with respect to estimated parameters, and under certain regularity conditions which are often satisfied in practice (Lehmann and Casella (1998)), then the parameter covariance matrices are equal to the inverses of the *Fisher information matrices* $(\boldsymbol{I}_{\boldsymbol{\theta}_X}, \boldsymbol{I}_{\boldsymbol{\beta}})$. Assuming that the log-likelihood is approximately quadratic in a neighborhood of the maximum, the *Fisher information matrices* are equal to the Hessian matrices of the log-likelihood functions with the sign changed:

$$\boldsymbol{I}_{\boldsymbol{\theta}_X} = -\frac{\partial^2 \ell(\boldsymbol{\theta}_X; \boldsymbol{x}_{\max})}{\partial^2 \boldsymbol{\theta}_X}; \quad \boldsymbol{I}_{\boldsymbol{\beta}} = -\frac{\partial^2 \ell(\boldsymbol{\beta}; \boldsymbol{x}, \boldsymbol{y})}{\partial^2 \boldsymbol{\beta}}. \tag{19}$$

The $(1-\alpha)$ confidence interval for each parameter is equal to:

$$\begin{array}{rcl} \theta^{\text{up}}_{X_j} &=& \hat{\theta}_{X_j} + t_{(1-\alpha/2, n_x - n_{p_x} - 1)} \hat{\sigma}_{X_j},\ j = 0, 1, \dots, n_{p_x} \\ \theta^{\text{lo}}_{X_j} &=& \hat{\theta}_{X_j} - t_{(1-\alpha/2, n_x - n_{p_x} - 1)} \hat{\sigma}_{X_j},\ j = 0, 1, \dots, n_{p_x}, \\ \beta^{\text{up}}_j &=& \hat{\beta}_j + t_{(1-\alpha/2, n_y - n_p - 1)} \hat{\sigma}_j,\ j = 0, 1, \dots, n_p \\ \beta^{\text{lo}}_j &=& \hat{\beta}_j - t_{(1-\alpha/2, n_y - n_p - 1)} \hat{\sigma}_j,\ j = 0, 1, \dots, n_p, \end{array} \tag{20}$$

where n_{p_x} is the number of components of vector $\boldsymbol{\theta}_X$, $t_{(1-\alpha/2, n_{df})}$ is the Student's t-distribution $(1-\alpha/2)$ quantile with n_{df} degrees of freedom and $\hat{\sigma}_{X_j}$ and $\hat{\sigma}_j$ are the corresponding estimated standard deviations for parameters j (square root of the corresponding diagonal term in $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_X}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$, respectively).

B Quantile confidence intervals

From an engineering design perspective, the calculation of return levels for different time spans (T_d) (usually in years) is of great interest. These return periods correspond, within the extreme value model selected, to quantiles associated with the following probability of not exceeding $q_{T_d} = 1 - 1/T_d$.

For the reanalysis case, these estimated quantiles $\hat{x}_q$ are calculated depending on the EV analysis model selected. For the proposed model, combining instrumental and reanalysis information, quantiles are obtained solving the implicit equation (5).

If we are interested in calculating the confidence bands for reanalysis quantiles x_q, it is known that for large sample sizes n_x, the quantile x_q is asymptotically normal, and thus, the delta method (Oehlert, 1992) can be applied as follows:

$$x_q \sim N\left(\hat{x}_q, \nabla^T_{\boldsymbol{\theta}_X} x_q \boldsymbol{\Sigma}_{\boldsymbol{\theta}_X} \nabla_{\boldsymbol{\theta}_X} x_q\right), \tag{21}$$

where $\nabla_{\boldsymbol{\theta}_X} x_q$ is the n_{p_x} vector of partial derivatives of quantile expressions with respect to $\boldsymbol{\theta}_X$.

Note that equation (21) allows obtaining the estimated variance $\hat{\sigma}^2_{x_q}$ of the quantile, and the confidence intervals then become:

$$\begin{aligned} x_q^{\mathrm{up}} &= \hat{x}_q + t_{(1-\alpha/2, n_x - n_{p_x} - 1)}\hat{\sigma}_{x_q}, \\ x_q^{\mathrm{lo}} &= \hat{x}_q - t_{(1-\alpha/2, n_x - p_x - 1)}\hat{\sigma}_{x_q}, \end{aligned} \tag{22}$$

For the proposed model, the process is analogous, we use the delta method in order to obtain the estimated variance of the corresponding quantile z_q:

$$z_q \sim N\left(\hat{z}_q, \nabla^T_{(\boldsymbol{\theta}_X;\boldsymbol{\beta})} z_q \boldsymbol{\Sigma}_{(\boldsymbol{\theta}_X;\boldsymbol{\beta})} \nabla_{(\boldsymbol{\theta}_X;\boldsymbol{\beta})} z_q\right), \tag{23}$$

where $\nabla_{(\boldsymbol{\theta}_X;\boldsymbol{\beta})} z_q$ is the $n_{p_x} + n_p$ vector of partial derivatives of quantiles from solving the implicit equation (5) with respect to $\boldsymbol{\theta}_X$ and $\boldsymbol{\beta}$. $\Sigma_{(\boldsymbol{\theta}_X;\boldsymbol{\beta})}$ is the variance-covariance matrix of all the estimated parameters, including the extreme value and the regression models. Since the models are independent by definition, it is equal to:

$$\Sigma_{(\boldsymbol{\theta}_X;\boldsymbol{\beta})} = \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\theta}_X} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \end{pmatrix}. \tag{24}$$

Note that the required derivatives for the reanalysis case are easily obtained analytically, however, for the composed model it is a challenge. For this reason, these are obtained numerically by finite differences:

$$\frac{\partial z_q}{\partial \gamma} = \frac{z_q(\gamma(1+\epsilon)) - z_q(\gamma(1-\epsilon))}{\epsilon\gamma}, \tag{25}$$

where γ represents the corresponding parameter and $\epsilon = 10^{-6}$.

References

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

Castillo E, Castillo C, Mínguez R (2008) Use of extreme value theory in engineering design. In: Martorel GSC S, Barnett J (eds) Proceedings of the European Safety and Reliability Conference 2008 (ESREL 2008), Safety, Reliability and Risk Analysis: Theory, Methods and Applications, vol 3, Taylor & Francis Group, Valencia, pp 2473–2488

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

Cooley D (2009) Extreme value analysis and the study of climate change. Climatic Change 97:77–83, DOI 10.1007/s10584-009-9627-x

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

Galiatsatou P, Prinos P (2011) Modeling non-stationary extreme waves using a point process approach and wavelets. Stochastic Environmental Research and Risk Assessment 25:165–183, DOI 10.1007/s00477-010-0448-2

Goubanova K, Li L (2007) Extremes in temperature and precipitation around the Mediterranean basin in an ensemble of future climate simulations. Global Planet Change 57:27–42

Izaguirre C, Méndez FJ, Menéndez M, Luceño A, , Losada IJ (2010) Extreme wave climate variability in Southern Europe using satellite data. Journal of Geophysical Research 115(–):–, DOI doi:10.1029/2009JC005802, to appear

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

Kharin VV, Zwiers FW, Zhang XB (2005) Intercomparison of near-surface temperature and precipitation extremes in AMIP-2 simulations, reanalyses and observations. J Climate 18:5201–5223

Kioutsioukis I, Melas D, Zerefos C (2010) Statistical assessment of changes in climate extremes over Greece (1955-2002). Int J Climatol 30:1723–1737

Lehmann EL, Casella G (1998) Theory of Point Estimation, 2nd edn. Springer Text in Statistics, Springer, New York

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

Méndez FJ, Menéndez M, Luceño A, Losada IJ (2007) Analyzing monthly extreme sea levels with a time-dependent GEV model. J Atmos Ocean Technol 24:894–911

Menéndez M, Méndez FJ, Izaguirre C, Losada IJ (2009) The influence of seasonality on estimating return values of significant wave height. Coastal Engineering 56(3):211–219

Mínguez R, Méndez FJ, Izaguirre C, Menéndez M, Losada IJ (2010) Pseudo-optimal parameter selection of non-stationary generalized extreme value models for environmental variables. Environmental Modelling & Software 25:1592–1607, DOI DOI: 10.1016/j.envsoft.2010.05.008

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

Nikulin G, Kjellstrom E, Hansson U, Strandberg G, Ullerstig A (2011) Evaluation and future projections of temperature, precipitation and wind extremes over Europe in an ensemble of regional climate simulations. TELLUS SERIES A-DYNAMIC METEOROLOGY AND OCEANOGRAPHY 63(1):41–55, DOI 10.1111/j.1600-0870.2010.00466.x

Oehlert GW (1992) A note on the Delta Method. The American Statistician 46(1):27–29

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

Rust HW, Maraun D, Osborn TJ (2009) Modelling seasonality in extreme precipitation. a UK case study. Eur Phys J Special Topics 174:99–111

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

Tomás A, Méndez FJ, Losada IJ (2008) A method for spatial calibration of wave hindcast data bases. Continental Shelf Research 28:391–398

Vanem E (2011) Long-term time-dependent stochastic modelling of extreme waves. Stochastic Environmental Research and Risk Assessment 25:185–209, DOI 10.1007/s00477-010-0431-y

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 pp 1–24, 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