ANALYZING EXTREME SEA STATE CONDITIONS BY TIME-SERIES SIMULATION ACCOUNTING FOR SEASONALITY

Erik Vanem
DNV Group Research & Development, Høvik, Norway
Department of Mathematics, University of Oslo, Oslo, Norway
E-mail: Erik.Vanem@dnv.com

ABSTRACT

This paper presents an extreme value analysis on data of significant wave height based on time-series simulation. A method to simulate time series with given marginal distribution and preserving the autocorrelation structure in the data is applied to significant wave height data. Then, extreme value analysis is performed by simulating from the fitted time-series model that preserves both the marginal probability distribution and the autocorrelation. In this way, the effect of serial correlation on the extreme values can be taken into account, without subsampling and de-clustering of the data. The effect of serial correlation on estimating extreme wave conditions have previously been highlighted, and failure to account for this effect will typically lead to an overestimation of extreme conditions. This is demonstrated by this study, that compares extreme value estimates from the simulated times-series model with estimates obtained directly from the marginal distribution assuming that 3-hourly significant wave heights are independent and identically distributed. A dataset of significant wave height provided as part of a second benchmark exercise on environmental extremes that was presented at OMAE 2021, has been analysed. This paper is an extension of a study presented at OMAE 2022 (OMAE2022-78795), and includes additional pre-processing of the data to account for seasonality and new results.

Keywords: Ocean environment; Extreme value analysis; Time series modelling; Significant wave height; Environmental loads; probabilistic wave models

1 INTRODUCTION

Probabilistic descriptions of extreme wave conditions are important for safe and reliable design and operation of marine structures. The most extreme environmental conditions the structure is expected to experience during its lifetime must be accounted for during design, and extreme value analysis of relevant metocean variables is crucial. One of the most relevant wave parameters for ocean engineering applications is the significant wave height, H_S, and several applications of extreme value analysis of this variable has been reported in the literature [1–8]. Indeed, a second benchmark exercise was announced at OMAE 2021 calling for estimates of extreme environmental conditions based on specific datasets of significant wave height that were made available [9]. This is a follow-up of a previous benchmarking exercise that, inter alia, highlighted the effect of serial correlation on extreme value estimates [10], as elaborated further in [11].

This paper presents extreme value estimates of significant wave height from the datasets presented in the second benchmark exercise, based on a time-series model that preserves the marginal distribution and the autocorrelation of the data. This is a way of accounting for serial correlation in the statistical analysis that would reduce the bias this has on extreme value estimates, but without the need for subsampling and de-clustering of the data. Such modelling could also provide a means for more realistic simulation of time-series of significant wave height that could be relevant for other applications than extreme value analysis, even though the focus of this paper is estimation of extreme conditions. A preliminary analysis of these types of models was presented at OMAE 2022 [12]. However, in this

paper the analysis presented in [12] is extended by more carefully handling the seasonality in the data. The systematic seasonal variation is removed from the data in a pre-processing step, and the time series model is then fitted to the pre-processed data. It is believed that this is a better way of accounting for seasonality. The seasonal component can then be added back in on simulated data from the time-series model in order to perform extreme value analysis and other inferences.

The extreme value analysis presented in this paper is based on simulating time-series from the fitted statistical model to obtain several samples of the extreme T-year significant wave height. This gives an empirical distribution of the maximum H_S in a T-year period, and the corresponding T-year return value can be estimated as the $1/e$-quantile of this distribution (see e.g. [9, 11]). However, the approach presented in this paper gives an estimate of the full T-year maximum distribution and other measures such as the mean and standard deviation of the T-year maximum can also be estimated. Moreover, a parametric distribution may be fitted to the empirical extreme value distribution. In this paper, the focus will be on the 25-year extreme values, but any other return value can be analysed in a similar way.

In the following, a brief description of the significant wave height data used in this study will be given, as well as a description of the pre-processing steps to account for seasonality, before the statistical model will be presented. Finally, results from applying such a model to the significant wave height data will be presented and discussed.

2 SIGNIFICANT WAVE HEIGHT DATA

Three different datasets of significant wave height, referred to as Site 1, Site 2 and Site 3, respectively, have been made available as part of the second benchmark exercise [9]. It is noted that these are synthetic data generated by the FIO-ESM v2.0 climate model [13,14] and not actual observations (see [9] and references therein for further details). In all datasets, values of significant wave height at 3-hourly intervals are provided for a period of 25 years. Trace plots of the data for the three sites are shown in Figure 1, where the index on the x-axis denotes the 3-hourly intervals. Some summary statistics of the data are provided in Table 1, including some L-moments and L-moment ratios (see. e.g. [15, 16] for an introduction to L-moments). The 25- and 50-year return value estimates presented in this table are from a standard block-maxima approach on annual maximum data assuming a generalized extreme value distribution, see e.g. [17], and are provided for reference. The estimated values of the shape parameters (and their standard errors) are -0.148 (0.165) for Site 1, -0.0412 (0.173) for Site 2 and -0.103 (0.176) for Site 3, respectively[1]. It is noted that these estimates are somewhat different from the baseline estimates provided in [9], which were based on the block maximum approach on annual maxima but assuming a Gumbel distribution that tacitly assumes that the shape parameter is 0. The Gumbel assumption seems reasonable from the estimated shape parameters, but these differences illustrate the large uncertainties associated with extreme value analysis. Note also that although the most severe conditions in the baseline results presented in [9] are associated with Site 1, the summary statistics in Table 1 suggest that the most severe conditions occur at Site 3. Hence, it is possible that the naming of the datasets have been mixed up and that what is referred to as Site 1 in [9] is indeed dataset Site 3 in this paper. However, as long as one do not intend to compare results, the naming of the datasets has no relevance. Note also that all datasets have positive skewness, indicating right-skewed distributions, and that the kurtoses are greater than 3, indicating heavier tails than the normal distribution.

2.1 DATA PRE-PROCESSING AND REMOVAL OF SEASONAL COMPONENTS

It is clearly observed from the trace plots of the significant wave height data that there are strong seasonal variations in the data. The mean values changes over the year, and the data is also heteroscedastic, i.e. the variability varies throughout the year. In [12], this was not explicitly accounted for, and the time series models were fitted to these raw data without any pre-processing to account for the seasonality. In this paper, however, the systematic seasonal component is removed, and the time-series model is fitted to the residuals, i.e. the seasonally normalized data. This is believed to be a more correct way of handling seasonality and other deterministic components in observed time-series data, and will yield updated results.

The same approach for estimating and removing the seasonal component of the observed time-series as proposed in [18], and also utilized in [19], is taken. That is, seasonally standardized data are obtained by subtracting the seasonal mean and dividing by the seasonal standard deviation. Finally, the the overall mean value is added back in. If $Y_{i,j}$ is a sample belonging to seasonal time point $j, j = 1,\ldots,J_y$, then the pre-processed sample X_i is calculated as follows, where μ_j is the seasonal average for seasonal time j, σ_j is the

[1] Note that a negative shape parameter of the GEV distribution indicates a finite upper bound on the variable. The upper bounds corresponding to the fitted GEV models are 16 m (Site 1), 33 m (Site 2) and 27 m (Site 3), respectively

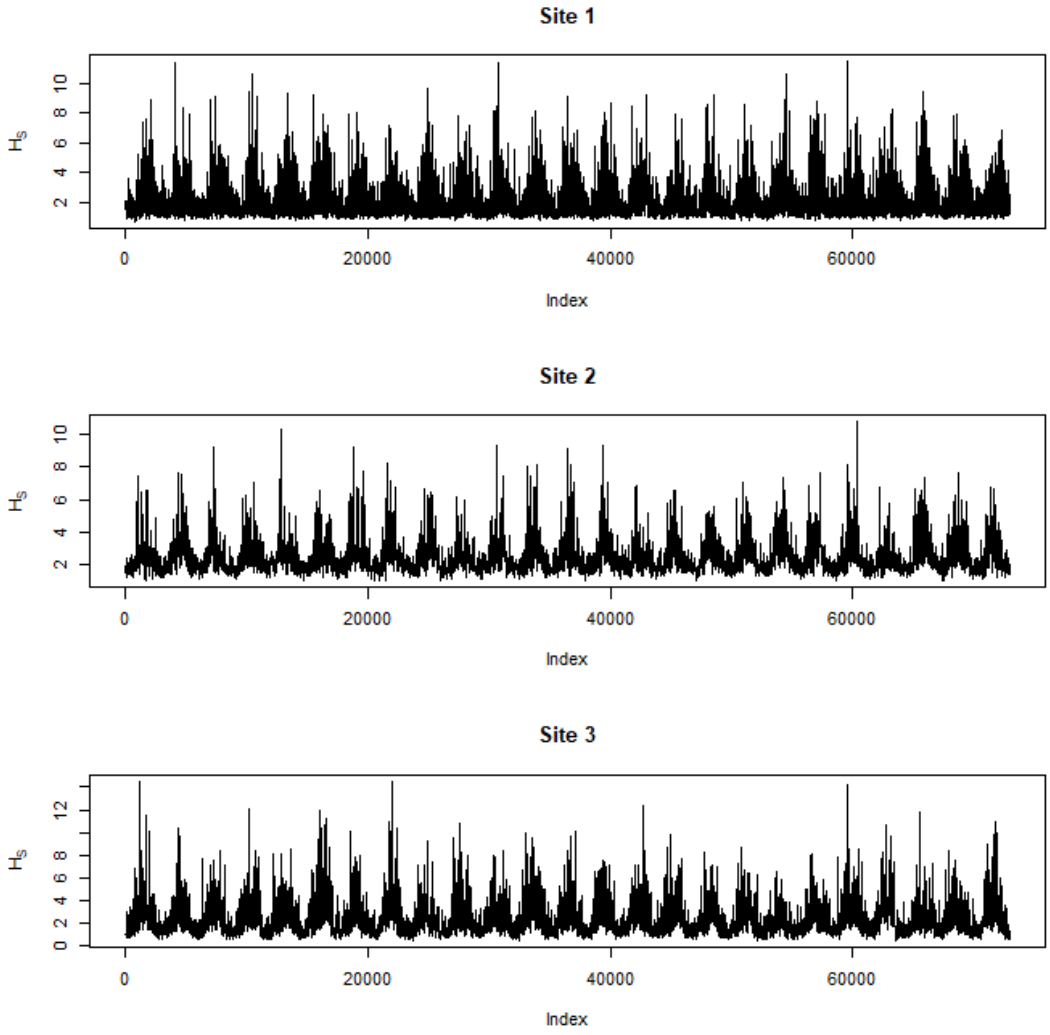

FIGURE 1. Trace plots of the significant wave height data

seasonal standard deviation for seasonal time j and M is the overall mean,

$$X_i = \frac{Y_{i,j} - \mu_j}{\sigma_j} + M. \tag{1}$$

It is noted that seasonality may influence other moments and other features in the data than the mean and the variance, but it is tacitly assumed that correcting for the seasonal variation in the mean and the variability will account for the most important seasonal contribution. It is possible to also account for seasonal changes in higher order moments, but this will complicate things and is left for further study.

In this study, 3-hourly data is available, and hence $J_y = 8 \times 365 = 2920$ since these are simulated data without any leap years. Hence, a seasonal mean and standard deviation is estimated for each 3-hour period throughout the year. Traceplots of the seasonally standardized data are shown in Figure 2, and the estimated seasonal means and standard deviations are shown in Figure 3. The seasonally normalized data, which do not exhibit any seasonal effects will be modelled by the time series model in such a way that it preserves the marginal and the autocorrelation structure in the data. The seasonal effects can be added back in by inverting eq. (1).

TABLE 1. Summary statistics for the significant wave height data (in meters).

	Site 1	Site 2	Site 3
minimum	0.72	1.01	0.41
median	1.64	2.16	1.85
mean	1.95	2.41	2.20
90%tile	3.19	3.58	3.91
99%tile	5.85	5.60	6.87
maximum	11.44	10.79	14.53
standard deviation	1.01	0.90	1.32
skewness	2.38	1.78	1.81
kurtosis	11.42	7.79	8.35
L-scale	0.49	0.46	0.68
L-skewness	0.36	0.28	0.28
L-kurtosis	0.21	0.18	0.16
25-year return value	11.39	10.44	14.38
50-year return value	11.84	11.09	15.28

The effect of removing the seasonal component on the distribution of the data is illustrated in Figure 4, where the empirical densities are shown for both the raw data and the seasonally corrected data for all three sites. It is noted that one unwanted effect of the seasonal standardization is that you may end up with negative values for Y_i for certain i. Since negative significant wave heights are un-physical, and one typically want to fit a probability distribution with only positive support to these data, these may be removed from the dataset before fitting a time-series model. If the ratio of such points is low, this is not deemed to be of any practical relevance, and for the three datasets analysed in this study, negative values of Y_i occurred 3 (site 1), 2 (Site 2) and 5 (Site 3) times, respectively. In this analysis, those points were simply replaced by the minimum values of the other standardized data points. Other options could be to simply remove them or to set them to 0. Anyway, the ratio of such points is extremely low, so this is not believed to influence the analysis and results.

3 STATISTICAL MODEL FOR TIME-SERIES SIMULATION

Typically, in extreme value analysis, one may fit statistical models to the peaks in the data only, e.g. model peaks over a threshold or block maxima, which may be assumed to be independent and identically distributed (iid) after sensible de-clustering. In this way, serial correlations are removed in the data that are analysed, and the effect of serial correlation will be negligible. However, one drawback of this approach is that it is wasteful, and the amount of data available for the analysis is significantly reduced. Typically, with an annual maximum approach, the amount of data is reduced to 25 from 25 years of observations. Even though the bias due to serial correlation is avoided, the uncertainty due to sampling variability will be rather large.

In this paper, a different approach is suggested that explicitly takes the serial correlation into account in the statistical modelling. That is, rather than merely fitting a distribution function to the data and making inference based on that, a statistical model that both preserves the marginal distribution and the autocorrelation in the data will be established, and estimates of extreme values will be based on that model. There are different approaches that could be used in order to establish such time-series models [20]. For example, [21,22] propose diffusion-type models with given marginal distributions and autocorrelation functions, Weibull and gamma autoregressive processes are presented in [23], and Gaussian and Non-Gaussian autoregressive models are discussed in [24]. However, in this paper, an approach based on transforming a Gaussian parent autoregressive time series to have the desired marginal distribution and temporal correlation structure will be taken, as suggested in [25]; see also [26,27].

The time-series of interest, $X(t)$, is specified by a marginal distribution, $F_X(x)$ and an autocorrelation structure, $\rho_X(\tau)$ and it is assumed that this has a parent Gaussian time-series, $Z(t)$ with a standard Gaussian marginal $\Phi_Z(z)$ and another autocorrelation structure $\rho_Z(\tau)$. Hence, modelling and simulation of the time-series $X(t)$ correspond to modelling and simulating the parent Gaussian time-series and finding the transformations that transform the parent time-series to the desired one, i.e. determine functions g and $\mathscr{T}$ such that

$$\begin{aligned} X(t) &= g(Z(t)) \\ \rho_Z(\tau) &= \mathscr{T}(\rho_X(\tau)) \end{aligned} \tag{2}$$

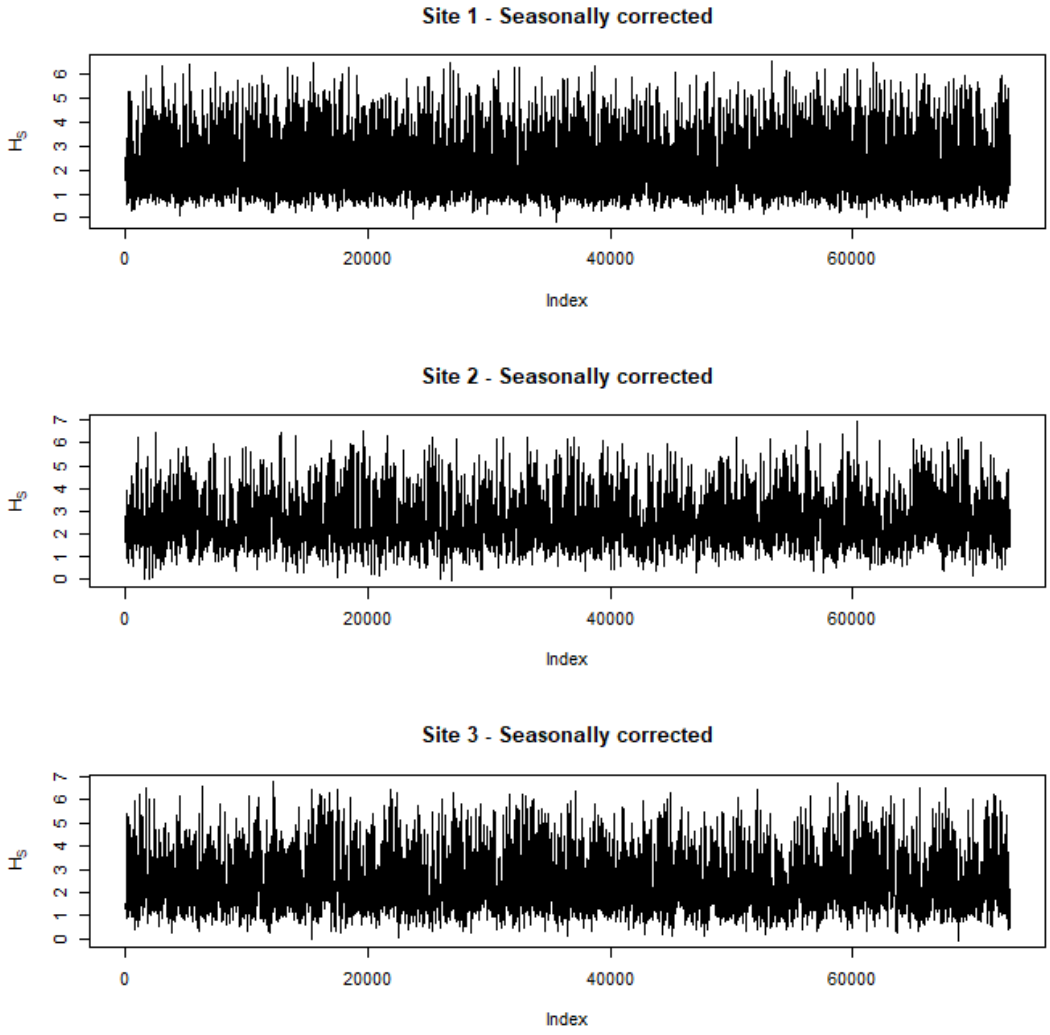

FIGURE 2. Trace plots of the seasonally standardized data

Simulating from a Gaussian time-series is not difficult, and the normal distribution have several well-know properties. For example, if $Z(t)$ is a standard normal time-series, then any linear combination of $Z(t)$'s will be jointly normally distributed and the marginal will be standard normal. Hence, if one can find a parent Gaussian time-series that can be transformed into the time-series $X(t)$, simulation of $X(t)$ is straightforward.

As pointed out in [25], g is easily identified as $g(Z) = F_X^{-1}(\Phi_Z(Z))$, which will give the desired marginal distribution. However, this transformation of the marginal distribution does not preserve the autocorrelation structure, which depends on the marginal distribution $F_X(x)$. Hence, one needs to establish the autocorrelation structure of the parent Gaussian time-series, $\rho_Z(\tau)$ that yields the required autocorrelation structure for the target time-series $X(t)$ after transformation.

In the bivariate case, it can be shown that for an arbitrary transformation, the correlation between the transformed variables will be smaller than the correlation between the initial standard normal variables (see e.g. [25]), i.e.

$$\rho_X = Cor[g(Z(t_1)), g(Z(t_2))] \leq Cor[Z(t_1), Z(t_2)] = \rho_Z, \tag{3}$$

meaning that the ρ_Z values needs to be inflated to obtain the target ρ_X. Moreover, the correlation coefficient of two lagged variables

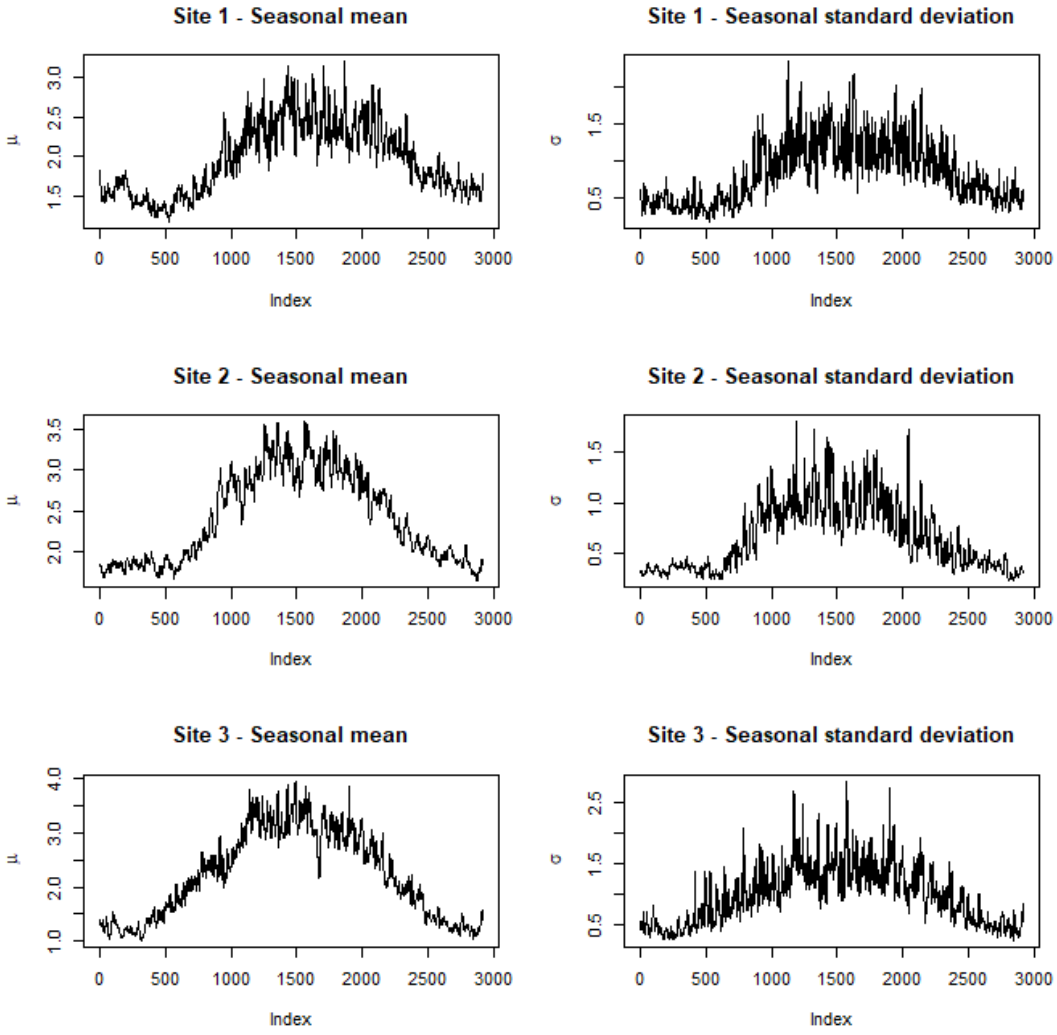

FIGURE 3. Estimated seasonal effects; seasonal means and standard deviations

$X(t_1)$ and $X(t_2)$ can be expressed as

$$\rho_X(\tau) = \frac{E\left(X(t_1)X(t_2)\right) - \mu_X^2}{\sigma_X^2}, \tag{4}$$

where E is the expectation operator, $\tau = t_2 - t_1$ and μ_x and σ_X are the mean and the standard deviation of X. Now, assuming that $X(t) = g(Z(t)) = F_X^{-1}\left(\Phi_Z(Z(t))\right)$, one has that

$$\begin{aligned} E\left(X(t_1)X(t_2)\right) &= E\left(F_X^{-1}(\Phi_Z(Z(t_1)))F_X^{-1}(\Phi_Z(Z(t_2)))\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_X^{-1}(\Phi_Z(u))F_X^{-1}(\Phi_Z(v))\phi_{u,v}(u,v;\rho_Z(t_2-t_1))dudv, \end{aligned} \tag{5}$$

where $\phi(u,v;\rho)$ denotes the density function of the bivariate standard normal distribution with correlation ρ. This gives a relationship

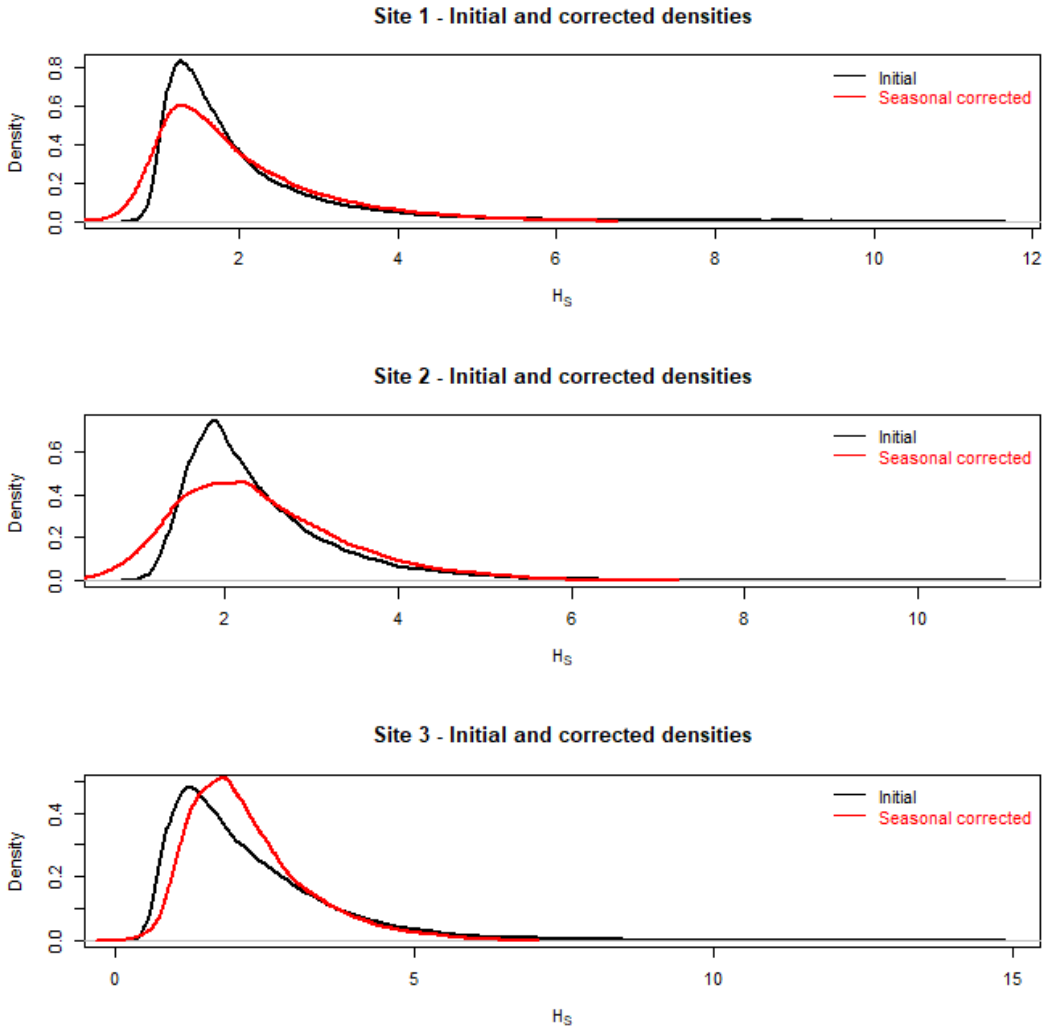

FIGURE 4. Empirical densities of the significant wave height data before and after seasonal normalization

between $\rho_Z(\tau)$ and $\rho_X(\tau)$ and can be used to calculate $\rho_X(\tau)$ from known $\rho_Z(\tau)$, i.e.

$$\rho_X(\tau) = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_X^{-1}(\Phi_Z(u))F_X^{-1}(\Phi_Z(v))\phi_{u,v}(u,v;\rho_Z(\tau))dudv - \mu_X^2}{\sigma_X^2}. \tag{6}$$

Note however, that in order to model the time series with known $\rho_X(t)$, one needs the inverse transformation to find $\rho_Z(t)$ from $\rho_X(t)$. Note also that even though the double integral above does not have an analytical solution in general, its numerical estimation is straightforward.

Hence, the modelling problem is solved by defining a correlation transformation function to estimate the autocorrelation of the parent Gaussian time-series from a given target autocorrelation structure that may be estimated from the data. The following parametric

form of such an autocorrelation transformation function is proposed in [25]:

$$\rho_Z = \mathscr{T}(\rho_X) = \frac{(1+\alpha\rho_X)^{(1-\beta)}-1}{(1+\alpha)^{(1-\beta)}-1}. \tag{7}$$

This can be used to estimate a parametric autocorrelation function of the parent Gaussian time-series, $\rho_Z(\tau)$ based on a parametric autocorrelation function for the target time-series, $\rho_X(\tau)$. The parameters of eq. (7) can be fitted by calculating ρ_X for $\rho_Z = (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95)$ according to eq. (6) and fit the function to the set of (ρ_X, ρ_Z)-points (see [25] for details).

With this approach, modelling a time series with desired marginal and correlation structure involves fitting the marginal distribution and estimating the empirical autocorrelation structure from the time-series, fitting the parametric autocorrelation transformation function to obtain the autocorrelation structure of the parent Gaussian time-series, fit an autoregressive model to generate the Gaussian parent time-series with standard normal marginals and transforming the Gaussian time-series using the transformation $X(t) = g(Z(t)) = F_X^{-1}(\Phi(Z(t))$. In the following, such a model will be fitted to the three datasets of significant wave height by estimating the marginal distributions and the autocorrelation functions. All analyses will be done on the seasonally corrected data.

3.1 Marginal distributions

The first step in the modelling approach presented above is to fit a marginal distribution to the data. For significant wave height data, the 3-parameter Weibull distribution is often assumed to give a good fit [28], and this distribution will be assumed also in this study. Hence, the 3-parameter Weibull distribution, with density function given in Eq. (8), will be fitted to the three datasets. Two fitting techniques will be tried, i.e. maximum likelihood (ML) and fitting based on the second order Anderson-Darling statistic (AD), see e.g. [29]. The resulting fits are shown in Figure 5, and estimated model parameters are shown in Table 2. It is observed that the marginal fits are not too good, but AD-fits are considerably better for the upper tail compared to the ML-fits, although completely missing the lower values, and these marginal distributions will be assumed in the further analysis.

$$f_{H_s}(h_s) = \frac{\beta}{\alpha}\left(\frac{h_s-\gamma}{\alpha}\right)^{\beta-1} \exp\left[-\left(\frac{h_s-\gamma}{\alpha}\right)^{\beta}\right]. \tag{8}$$

It is noted that improved fits could possibly be obtained by using other fitting methods, e.g. the method of moments, or by assuming other parametric distribution functions, for example a hybrid model with a Weibull distribution for the body and a generalized Pareto distribution for the tails. However, for the purpose of this study, the 3-parameter Weibull distribution fitted by minimizing the second order Anderson-Darling statistic will be assumed, and this is found to capture the upper tail quite well. It should also be noted that the maximum likelihood estimator is known to be a poor choice for fitting the 3-parameter Weibull distribution, and it may not even exist [30]. it is included in this study merely as a reference. For other fitting methods of the 3-parameter Weibull distribution, see e.g. [31–39].

TABLE 2. Fitted model parameters for the marginal distributions.

		shape (β)	Scale (α)	location (γ)
Site 1	ML fit	2.117	2.208	0.00487
	AD fit	**1.289**	**1.360**	**0.696**
Site 2	ML fit	2.594	2.707	0.0047
	AD fit	**1.680**	**1.778**	**0.825**
Site 3	ML fit	2.345	2.457	0.0317
	AD fit	**1.409**	**1.489**	**0.855**

The parametric function assumed for the autocorrelation transformation function in eq. (7) is found to fit the autocorrelation transformation points very well, as demonstrated in Figure 6 for the three marginal distributions fitted to the different datasets in this study.

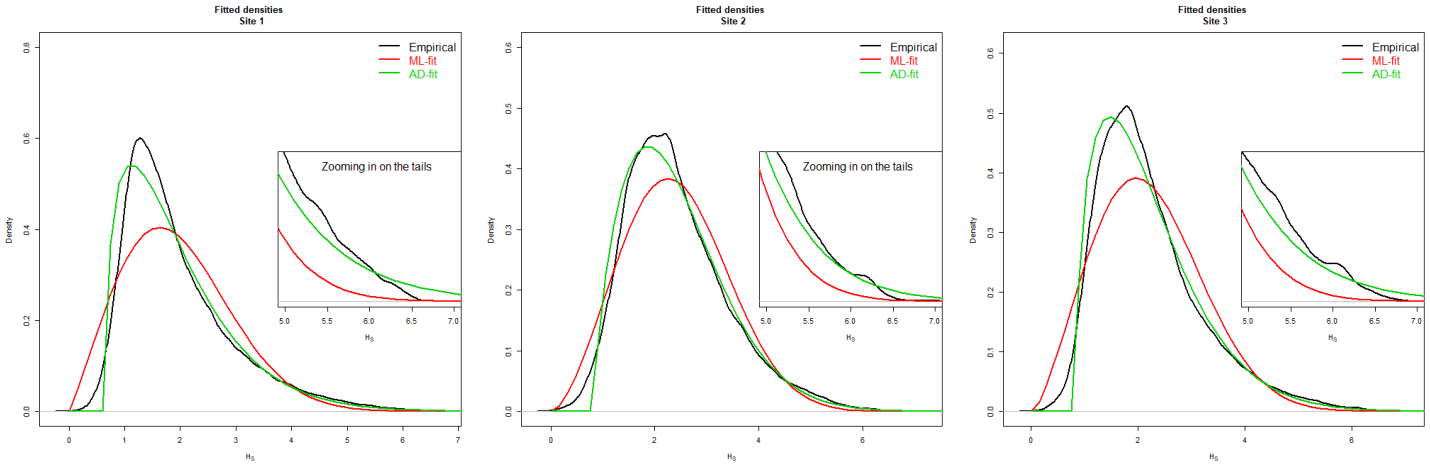

FIGURE 5. Fitted marginal distributions

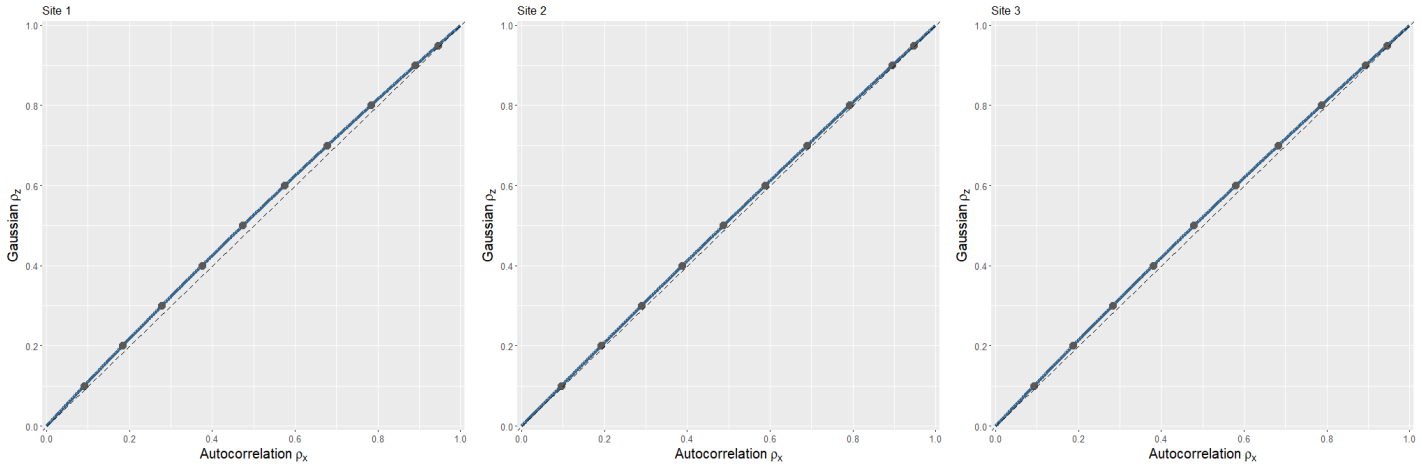

FIGURE 6. Fitted autocorrelation transformation functions according to eq. (7)

3.2 Autocorrelation functions

In order to specify the autocorrelation of the parent Gaussian autoregressive model, one must estimate the autocorrelation function for the target time-series, and this is done by fitting a parametric function to the empirical autocorrelation function (ACF). There are many parametric autocorrelation functions to choose from, see e.g. [25] and in this study, six different parametric functions are tried out, i.e. the Weibull, the Burr, the logarithmic, the fGn, the generalised fGn and the Pareto ACF. It is argued in [40] that it might be sufficient to consider only a few ACFs that cover different asymptotic cases, but in this study, as many as 6 ACFs are explored. The empirical autocorrelation functions are shown together with the fitted parametric candidate models in Figure 7. The functions are fitted by least squares, and the residual sum of squares (RSS) are indicated for each candidate model in the plots. It is observed that the 3-parameter Burr-type autocorrelation function fits best to the data for two of the three sites, but that the 2-parameter Weibull ACF fits better for Site 1. However, except for the fGn and gfGn types of ACF, the other parametric ACFs all fit the empirical ACFs quite well. It is also possible to use the non-parametric empirical ACF in the time-series model. Note also that the 3-parameter Burr-type ACF is known to sometimes yield not positive definite covariances, see e.g. [41] and hence might not always be suitable. This occurred for some of the sites on the raw data (see [12]), but not for the seasonally normalized data in this study. Note also that the empirical autocorrelation functions shown in Figure 7 look very different from the autocorrelation functions of the raw data, as shown in [12], and that they now more clearly tend to zero for increasing lags. This is obviously because the seasonal components have been removed, and the pre-processed data are much better suited for the time-series modelling proposed in this paper. The parametric forms of the different autocorrelation functions are

given in Eq. (9).

$$
\begin{aligned}
\rho_{Burr}(\tau;\zeta,\eta,\theta) &= \left(1+\theta\left(\frac{\tau}{\zeta}\right)^{\eta}\right)^{-\frac{1}{\eta\theta}} \\
\rho_{Weibull}(\tau;\zeta,\eta) &= e^{\left(-\left(\frac{\tau}{\zeta}\right)^{\eta}\right)} \\
\rho_{Log}(\tau;\zeta,\eta) &= \left(1+\ln\left(1+\eta\frac{\tau}{\zeta}\right)\right)^{-\frac{1}{\eta}} \\
\rho_{fGn}(\tau;\zeta) &= \frac{1}{2}\left(|\tau-1|^{2\zeta}-2|\tau|^{2\zeta}+|\tau+1|^{2\zeta}\right) \sim \tau^{-2(1-\zeta)} \\
\rho_{gfGn}(\tau;\rho_1,\zeta) &= \zeta(2\zeta-1)\left\{\tau-1+\left[\frac{\rho_1}{\zeta(2\zeta-1)}\right]^{-\frac{1}{2(1-\zeta)}}\right\}^{-2(1-\zeta)} \sim \tau^{-2(1-\zeta)} \\
\rho_{Pareto}(\tau;\zeta,\eta) &= \left(1+\eta\frac{\tau}{\zeta}\right)^{-\frac{1}{\eta}}
\end{aligned}
\tag{9}
$$

Having established the statistical model, it is possible to simulate from it to generate simulated data with the desired marginal distribution and autocorrelation function. One would then first simulate data from the time-series model to obtain seasonally normalized data, and one could then add back in the seasonal component that was identified in the pre-processing step. Examples of such simulated data are shown in Figure 8 for all the three sites. The top plots show 8 years of observed data from the various sites and the bottom plots show first 4 years of data simulated from the fitted marginal distributions assuming iid and then four years of data simulated from the time-series models assuming the marginal distribution and an empirical autocorrelation function for the seasonally normalized data. It is observed that the data simulated from the time-series model better resemble the observed data than the data simulated assuming iid.

4 EXTREME VALUE ANALYSIS RESULTS

Having established the statistical models, it is possible to simulate synthetic time series of significant wave height for the three sites in order to do extreme value analysis. In this study, $M = 500$ 25-year time series will be simulated from the statistical models. For each simulated time-series the largest value is taken to represent a sample of the 25-year maximum significant wave height. This yields $M = 500$ samples from the distribution and can be used to make inference of the 25-year maximum significant wave height. One may estimate the entire distribution as well as for example the mean 25-year maximum significant wave height and the 25-year return value of significant wave height.

The simulated distributions of 25-year maximum significant wave height are shown in Figure 9. Results are shown for time-series models assuming the parametric Burr, Weibull, Logarithmic and Pareto autocorrelation functions, as well as for a model using the non-parametric empirical autocorrelation function. In addition, results are shown for data simulated assuming iid from the marginal distributions. The plots also include the corresponding estimates of the 25-year return values for significant wave height based on annual maxima, as well as the quantile of the marginal distribution corresponding to the 25-year return value if serial correlation is ignored.

Some general observations can be made from these results. Firstly, it is observed that the assumed autocorrelation function somewhat influences the return value estimates. For all of the sites, assuming the same marginal distribution, return value estimates vary according to what parametric autocorrelation function was fitted to the data. Furthermore, it is observed that all estimates obtained from a time-series model are significantly lower than the return value estimates obtained from assuming iid and hence ignoring serial correlation. Hence, even though the best time series model is not determined, it seems clear that accounting for serial dependence in some way reduces or even eliminates the positive bias in return value estimates obtained when neglecting serial correlation. It appears that this effect is most significant for Site 2, and this makes sense since these data are also the data with strongest autocorrelation, as can be seen from the autocorrelation functions shown in Figure 7.

For the three sites, the largest differences in the 25-year return value estimates based on time series models with the some autocorrelation function and the estimates ignoring serial correlation are 2.04 m (Site 1 with Weibull ACF), 1.72 m (Site 2 with empirical ACF) and 1.59 m (Site 3 with Burr ACF), respectively. This may be practically significant, suggesting that serial correlations should be

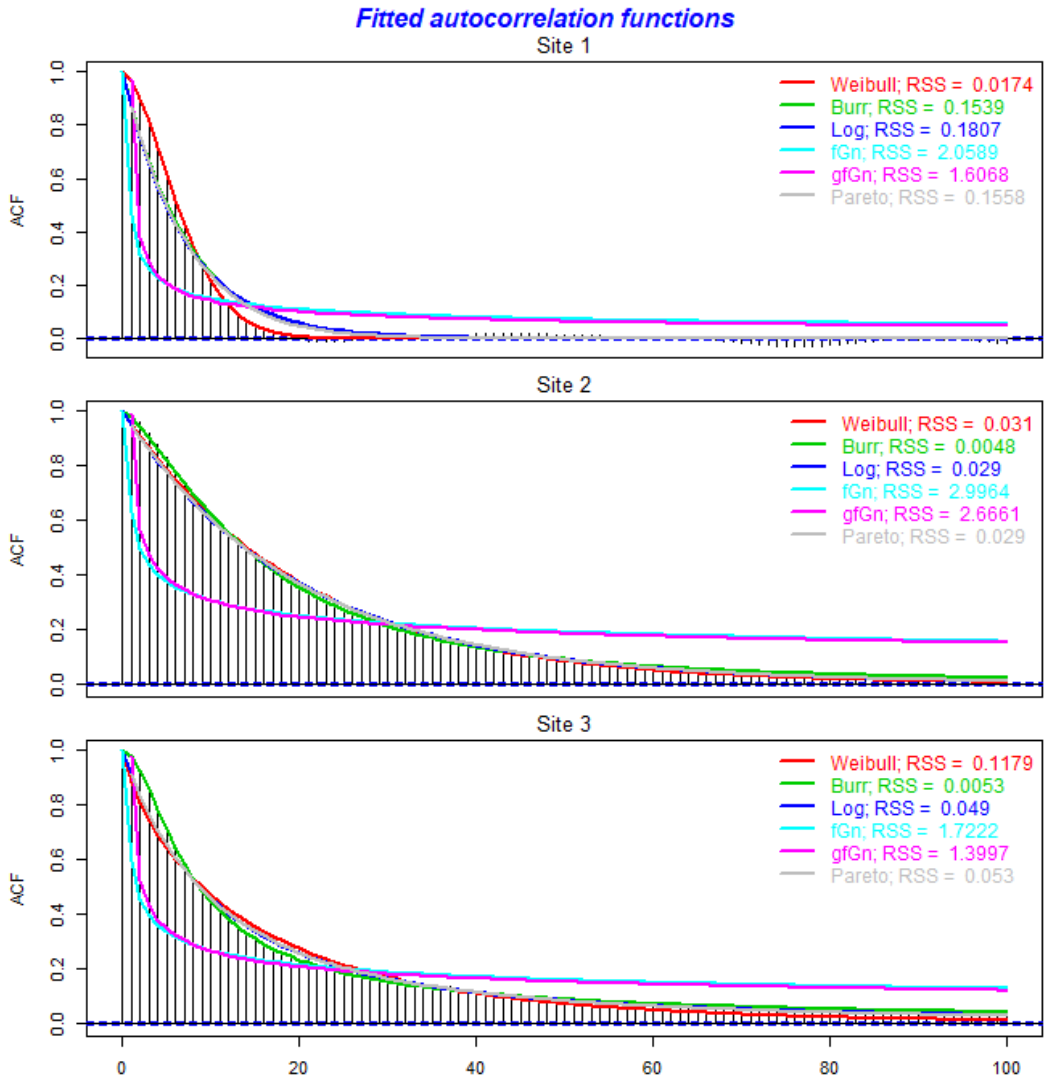

FIGURE 7. Autocorrelation functions

taken into account in extreme value analysis for ocean engineering applications. On the other hand, the choice of autocorrelation function influences the results, and the smallest differences between the 25-year return value estimates from the time-series models and the models assuming IID are 1.33 m (Site 1 with logarithmic ACF), 0.53 m (Site 2 with Pareto ACF) and 0.43 m (Site 3 with Weibull ACF), respectively. This may still be significant from a practical point of view, but illustrates the uncertainty due to modelling assumptions on the autocorrelation function.

Results may also be compared to the reference values in Table 1 obtained from rather standard GEV-fitting to annual maxima, and it is observed that there are no recognizable patterns. For Site 1, the return value estimates obtained from fitting GEV to annual maxima are considerably lower than all estimates obtained using the time series models, and also obviously much lower than the return values obtained when ignoring serial correlation. For Sites 2 and 3, the annual-maxima based return value estimates fall between the estimates obtained from the time series models and the estimates obtained when ignoring time-series models. If nothing else, these observations serve to remind about the large uncertainties associated with extreme value analysis [29].

It is also possible to fit parametric distributions to the simulated T-year extreme data, and the Gumbel distribution can be assumed as a reasonable extreme value model. Hence, the resulting parametric distributions fitted to the simulated 25-year maximum data for the various sites are shown in Figure 10. The empirical distributions are shown with dashed lines and the solid lines are the fitted Gumbel distributions. The fits are good and slightly different return value estimates can also be obtained from these parametric distributions,

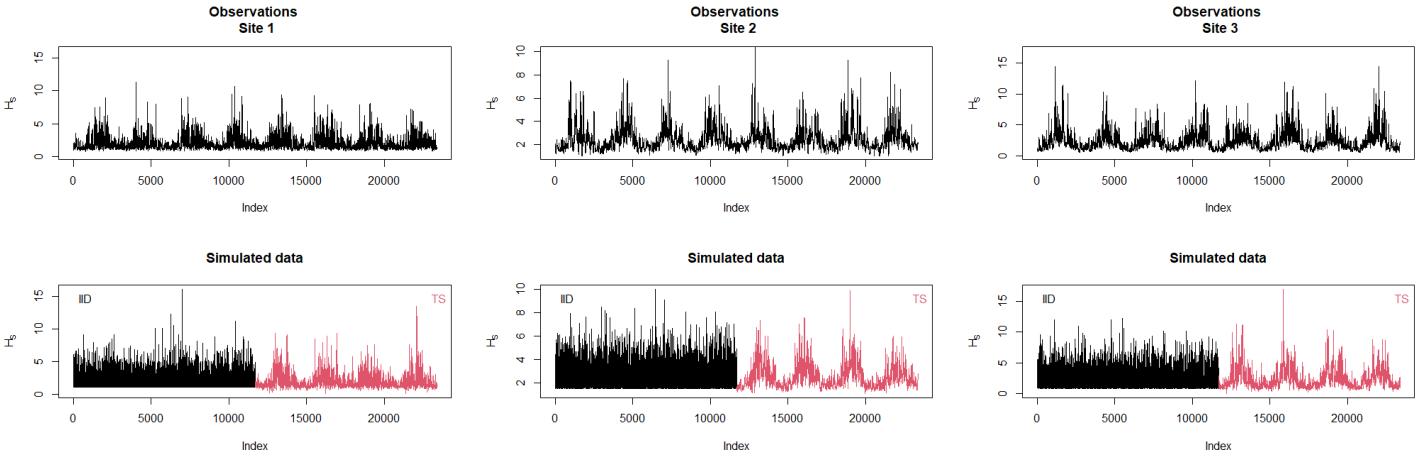

FIGURE 8. Comparing observed (top) and simulated (bottom) data; simulated data are from the marginal distribution assuming iid and from the time-series model with desired marginal and autocorrelation function, respectively

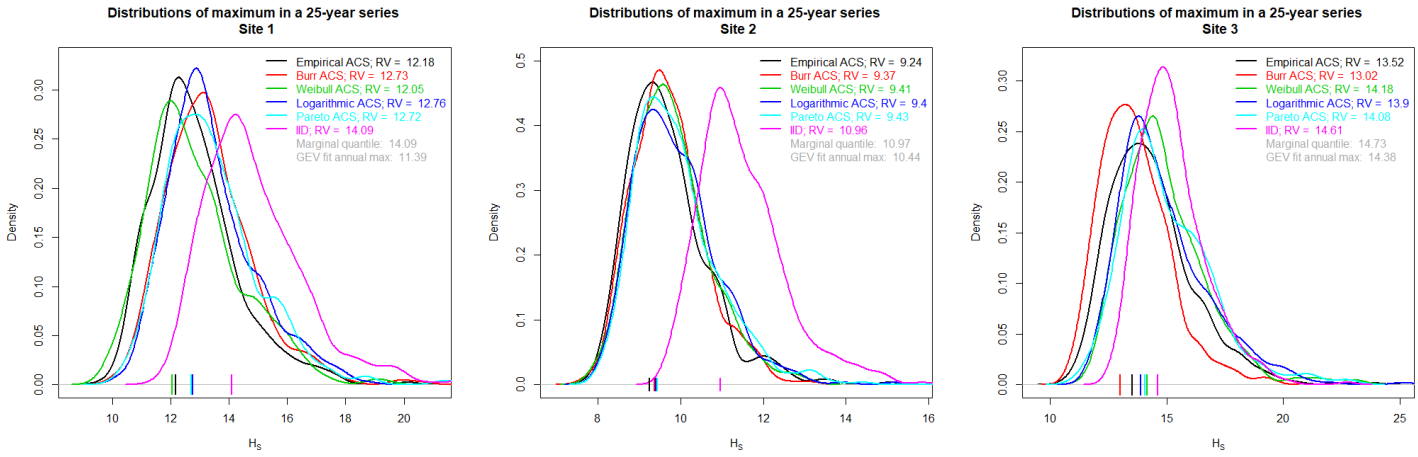

FIGURE 9. Distributions of 25-year maximum significant wave height from the time-series models

as indicated in the figure. Note, however, that in real engineering applications, the Gumbel assumption should be carefully checked, and alternatively the full GEV model could be fitted. However, for the purpose of this exercise, the Gumbel distribution is assumed reasonable.

Having simulated data from a fitted model, this may obviously be used for extreme value analysis and to estimate return values corresponding to any return period. In Table 3, simulated data from the time series model assuming the empirical autocorrelation function are used to estimate 50-, 500- and 5000-year return values. The estimates are based on the $M = 500$ simulated 25-years of data and then fitting a Gumbel distribution to the 25-year maxima, and the table also includes estimated 95% confidence intervals of the return value estimates. These are quite narrow, since a rather large sample size (M = 500) was used to fit the Gumbel distributions.

These estimates can be compared to estimates of the same return values based on fitting a generalized extreme value directly to the annual maxima, as shown in Table 4. With this classical approach, serial correlation is accounted for by sub-sampling, but this means that only 25 samples are available for the fitting and subsequent inference, resulting in significantly wider confidence intervals, particularly for increasing return periods.

Comparing these return value estimates, it is clear that the point estimates based on the time-series models are sometimes higher and sometime lower compared to the annual maxima fitting. However, the confidence intervals are considerably tighter since more data

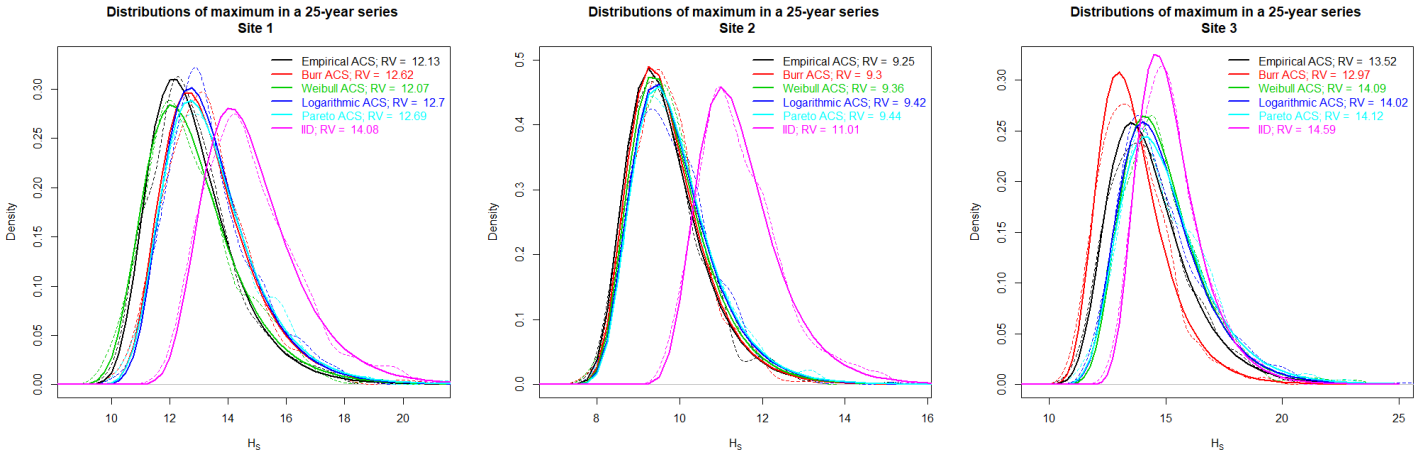

FIGURE 10. Fitting Gumbel distributions for the 25-year maximum data

TABLE 3. Return value estimates based on the time-series model

	Return period	Estimate	95% c.i.
Site 1	50 y	12.562	(12.442, 12.686)
	500 y	15.640	(15.360, 15.945)
	5000 y	18.390	(17.940, 18.883)
Site 2	50 y	9.529	(9.452, 9.608)
	500 y	11.499	(11.319, 11.696)
	5000 y	13.259	(12.969, 13.578)
Site 3	50 y	14.044	(13.900, 14.193)
	500 y	17.751	(17.412, 18.118)
	5000 y	21.062	(20.516, 21.658)

TABLE 4. Return value estimates based on fitting a generalized extreme value distribution to annual maxima

	Return period	Estimate	95% c.i.
Site 1	50 y	11.844	(10.963, 15.517)
	500 y	13.056	(11.751, 23.081)
	5000 y	13.915	(11.827, 27.829)
Site 2	50 y	11.094	(9.879, 16.483)
	500 y	13.123	(10.749, 22.965)
	5000 y	14.961	(11.968, 33.661)
Site 3	50 y	15.283	(13.528, 23.097)
	500 y	17.828	(14.872, 39.545)
	5000 y	19.826	(15.861, 49.566)

have been used to fit the models. It could be noted that the confidence bounds for the time-series results might be unrealistically tight, since they do not include the model uncertainty, but they are still believed to be reasonable estimates. Results based on annual maxima, on the other hand, are associated with very large confidence bounds, most likely due to sampling variability and the fact that only 25 annual maxima are available for the analysis.

Finally, the results can be compared with return value estimates using a similar approach but ignoring serial correlation. That is, simulate the same number of 25-years of data assuming IID from a marginal distribution, and then fitting a Gumbel distribution to the simulated $M = 500$ 25-year maxima. This would presumable give equally tight confidence intervals, but with a positive bias from

ignoring the serial correlation. Indeed, this is what is generally observed, as presented in Table 5. Hence, the proposed time-series model has an advantage over sub-sampling techniques such as block maxima in that it can provide tighter confidence intervals, and it can account for the positive bias in approaches based on all data that ignores serial correlation.

TABLE 5. Return value estimates based on simulating IID

	Return period	Estimate	95% c.i.
Site 1	50 y	14.558	(14.426, 14.696)
	500 y	17.963	(17.650, 18.304)
	5000 y	21.004	(20.500, 21.559)
Site 2	50 y	11.305	(11.224, 11.390)
	500 y	13.396	(13.205, 13.605)
	5000 y	15.263	(14.956, 15.603)
Site 3	50 y	15.005	(14.891, 15.123)
	500 y	17.938	(17.671, 18.230)
	5000 y	20.558	(20.127, 21.031)

5 DISCUSSION

The extreme value analysis presented in this paper is a way of accounting for serial correlation by simulating from a time series model that preserves the marginal distribution and the autocorrelation in the observed data. It is demonstrated to give lower extreme value estimates than what one would get if serial correlation is ignored and hence to reduce the positive bias known to occur [11]. It also has an advantage compared to standard extreme value analysis techniques in that there is no need for sub-sampling and de-clustering, making it less wasteful. However, results will be sensitive to the actual time-series model fitted to the data, in terms of both the marginal and the autocorrelation function. Indeed, different autocorrelation functions were used in this study and were found to give different results. Hence, efforts should be made in proper model validation and selection.

In this study, the rather crude models presented in [12] have been improved by carefully removing the systematic seasonal component before fitting the time-series models. This is believed to be important, so that the seasonal effect is not a part of the autocorrelation function. In this paper, the approach for modelling and removing the seasonal components proposed in [18] has been adopted. A similar approach was suggested in [42]. It could also be possible to fit parametric functions to the estimated seasonal effects, for example different trigonometric functions fitted to the seasonal means and standard deviations shown in Figure 3, but this would probably not improve the models much. The study in [43] suggested that an annual components would be sufficient and that e.g. adding a semi-annual part would not make much difference for a similar dataset. Other approaches could be to bin the data into different seasons and fit season-specific models to each bin, similarly to the approach taken by [7, 44], or to fit non-stationary extreme value models by introducing covariates, analogue to the analyses presented in e.g. [45–49], see also [17]. However, it should be kept in mind that non-stationary models might introduce additional sources of uncertainty, as pointed out in [50].

When the interest is in the extremes, one could argue that it is not necessarily the standard autocorrelation structure in the data that is most relevant, but rather the level-dependent serial correlation expressed by the conditional probability $\chi(u,\tau) = Pr(X(t+\tau) > u | X(t) > u)$, which may be illustrated by extremograms [51, 52]. The correlation structure for high levels, u, should also be preserved in the time series to ensure the effect of serial correlation on extremes are properly accounted for. It is possible to compare the sample extremogram of the data and the simulated time-series, as shown in Figure 11 with levels corresponding to quantiles $q = 0.99$, 0.999 and 0.9995. As can be seen from these plots, the simulated time series preserves the extremal correlation reasonably well, and much better than the simulated IID samples (where also the extremes will be independent). The TS-extremograms shown in Figure 11 correspond to the time series simulated with the non-parametric empirical autocorrelation structure.

In this paper, the time-series models have been applied to univariate time-series of significant wave height. In practical engineering applications, one is often interested in the joint behaviour of several variables [19, 53]. For example, environmental contours may be constructed in 2 or higher dimensions and used in design of offshore structures [54–57]. Hence, it would be useful to extend the time-series models presented in this paper to be able to model multivariate time-series. Environmental contours based on this approach are discussed in [58], but there the time-series model was only applied to the primary variable, and a conditional model was assumed for the

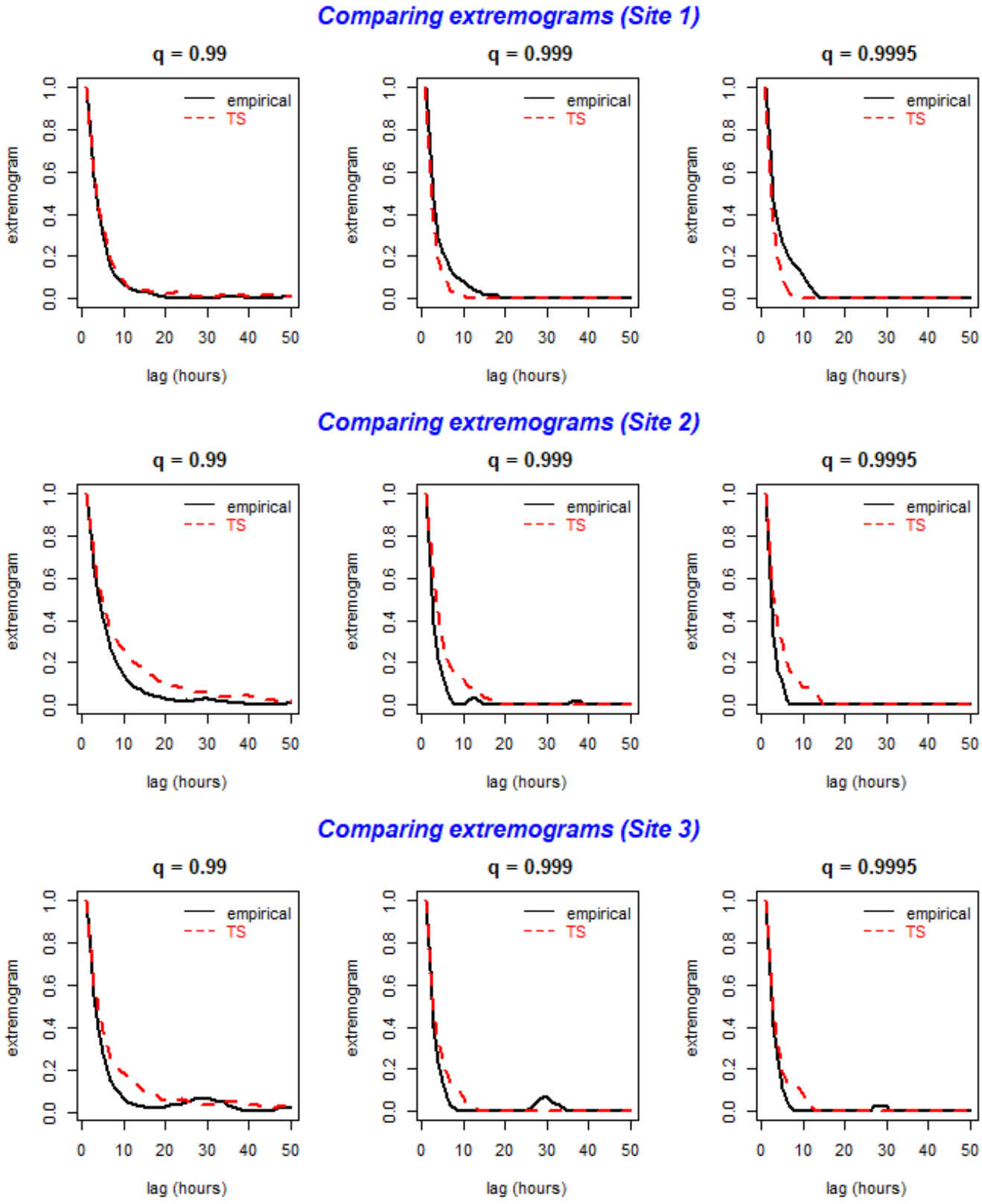

FIGURE 11. Comparing extremograms for the empirical data and simulated time series

secondary. Hence, the serial correlation was only partly accounted for. Even though this is deemed to be an improvement, further work on extending this model to higher dimensions, to yield models with desired marginals for all variables as well as properly modelling the autocorrelations and cross-correlations between all variables is recommended for further research.

6 SUMMARY AND CONCLUSIONS

This paper has presented an approach to fitting statistical models to metoean data that preserves the marginal distribution and the autocorrelation in the observed time series. Such models may be established from fitted marginal distributions and either the empirical autocorrelation function or a parametric function fitted to this. These models can then be used for extreme value analysis by simulating from the fitted models in order to obtain distributions of T-year maximum data. Return values can then be estimated from these distributions, which then takes serial correlation into account. It is also demonstrated that this reduces or eliminates the positive bias that is known to occur in extreme value analysis when ignoring the effect of serial correlation. Hence, the presented approach is presented as an alternative to standard extreme value analysis approaches based on peak values, and one advantage with the proposed method is that it is not wasteful. There is no need for de-clustering or peak picking of the data. Moreover, such statistical models may have other uses than extreme value analysis, and provide a means to simulate more realistic time series for both normal and extreme conditions.

Extreme value estimates based on the proposed approach are found to be reasonable. In particular, estimates are lower than estimates obtained when ignoring serial correlation. However, results also demonstrate that return value estimates will be sensitive to the assumed time series model. In particular, different assumptions of the autocorrelation function will give different results. Hence, particular focus should be put on obtaining the best possible model, and objective model selection criteria will be important. Furthermore, careful pre-processing of the data to remove any deterministic components related to for example seasonality should be carried out, and this paper has demonstrated a simple approach for taking this into account. Further work on extending this model to higher dimensions is recommended.

Acknowledgement

The work presented in this paper has been partly carried out within the research project HIPERWIND, with support from EU H2020 (grant agreement number 101006689), and has also been supported by the center for research-based innovation, Big Insight, with support from the Research Council of Norway (grant number 237718).

References

[1] Muir, L. R., and El-Shaarawi, A., 1986. "On the calculation of extreme wave heights: A review". *Ocean Engineering,* ***13***, pp. 93–118.

[2] Jonathan, P., and Ewans, K., 2013. "Statistical modelling of extreme ocean environments for marine design: a review". *Ocean Engineering,* ***62***, pp. 91–109.

[3] Vanem, E., 2011. "Long-term time-dependent stochastic modelling of extreme waves". *Stochastic Environmental Research and Risk Assessment,* ***25***, pp. 185–209.

[4] Stefanakos, C. N., and Athanassoulis, G. A., 2006. "Extreme value predictions based on nonstationary time series of wave data". *Environmetrics,* ***17***, pp. 25–46.

[5] Baxevani, A., Caires, S., and Rychlik, I., 2009. "Spatio-temporal statistical modelling of significant wave height". *Environmetrics,* ***20***, pp. 14–31.

[6] Vanem, E., Huseby, A. B., and Natvig, B., 2012. "A Bayesian hierarchical spatio-temporal model for significant wave height in the North Atlantic". *Stochastic Environmental Research and Risk Assessment,* ***26***, pp. 609–632.

[7] Randell, D., Feld, G., Ewans, K., and Jonathan, P., 2015. "Distributions of return values for the ocean wave characteristics in the South China Sea using directional-seasonal extreme value analysis". *Environmetrics,* ***26***, pp. 442–450.

[8] Vanem, E., Fazeres-Ferradosa, T., Rosa-Santos, P., and Taveira-Pinto, F., 2019. "Statistical description and modelling of extreme ocean wave conditions". *Proceedings of the Institution of Civil Engineers - Maritime Engineering,* ***172***, pp. 124–132.

[9] Mackay, E., Haselsteiner, A. F., Coe, R. G., and Manuel, L., 2021. "A second benchmarking exercise on estimating extreme environmental conditions: Methodology & baseline results". In Proc. 40th International Conference on Ocean, Offshore and Arctic Engineering (OMAE 2021), American Society of Mechanical Engineers (ASME).

[10] Haselsteiner, A. F., Coe, R. G., Manuel, L., Chai, W., Leira, B., Clarindo, G., Guedes Soares, C., Hannesdóttir, Á., Dimitrov, N., Sander, A., Ohlendorf, J.-H., Thoben, K.-D., de Hauteclocque, G., Mackay, E., Jonathan, P., Qiao, C., Myers, A., Rode, A.,

Hildebrandt, A., Schmidt, B., Vanem, E., and Huseby, A. B., 2021. "A benchmarking exercise for environmental contours". *Ocean Engineering,* ***236***, p. 109504.
[11] Mackay, E., de Hauteclocque, G., Vanem, E., and Jonathan, P., 2021. "The effect of serial correlation in environmental conditions on estimates of extreme events". *Ocean Engineering,* ***242***, p. 110092.
[12] Vanem, E., 2022. "Analyzing extreme sea state conditions by time-series simulation". In Proc. 41st International Conference on Ocean, Offshore and Arctic Engineering (OMAE 2022), American Society of Mechanical Engineers (ASME).
[13] Bao, Y., Sing, Z., and Qiao, F., 2020. "FIO-ESM version 2.0: Model description and evaluation". *Jouarnal of Geophysical Research: Oceans,* ***125***(6), p. e2019JC016036.
[14] Song, Z., Bao, Y., Zhang, D., Shu, Q., Song, Y., and Qiao, F., 2020. "Centuries of monthly and 3-hourly global ocean wave data for past, present, and future climate research". *Scientific Data,* ***7***, pp. 226:1–11.
[15] Vanem, E., 2017. "A regional extreme value analysis of ocean waves in a changing climate". *Ocean Engineering,* ***144***, pp. 277–295.
[16] Hoskins, J. R. M., and Wallis, J. R., 1997. *Regional Frequency Analysis*. Cambridge University Press.
[17] Coles, S., 2001. *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag.
[18] Vanem, E., 2018. "A simple approach to account for seasonality in the description of extreme ocean environments". *Marine Systems & Ocean Technology,* ***13***, pp. 63–73.
[19] Vanem, E., 2016. "Joint statistical models for significant wave height and wave period in a changing climate". *Marine Structures,* ***49***, pp. 180–205.
[20] Monbet, V., Ailliot, P., and Prevesto, M., 2007. "Survey of stochastic models for wind and sea state time series". *Probabilistic Engineering Mechanics,* ***22***, pp. 113 – 126.
[21] Bibby, B. M., Skovgaard, I. M., and Sørensen, M., 2005. "Diffusion-type models with given marginal distribution and autocorrelation function". *Bernoulli,* ***11***(2), pp. 191–220.
[22] Bensoussan, A., and Brouste, A., 2018. "Marginal Weibull diffusion model for wind speed modeling and short-term forecasting". In *Renewable Energy: Forecasting and Risk Management*, P. Drobinski, M. Mougeot, D. Picard, R. Plougonven, and P. Tankov, eds., Vol. 254 of *Springer Proceesings in Mathematics & Statistics*. Springer Nature, Switzerland, pp. 3–22.
[23] Sim, C.-H., 1986. "Simulation of weibull and gamma autoregressive stationary process". *Communications in Statistics - Simulation and Computation,* ***15***(4), pp. 1141–1146.
[24] Kaur, S., and Rakshit, M., 2019. "Gaussian and non-Gaussian autoregressive time series models with rainfall data". *International Journal of Engineering and Advanced Technology,* ***9***(1), pp. 6699–6704.
[25] Papalexiou, S. M., 2018. "Unified theory for stochastic modelling of hydroclimatic processes: Preserving marginal distributions, correlation structures, and intermittency". *Advances in Water Resources,* ***115***, pp. 234–252.
[26] Papalexiou, S. M., Markonis, Y., Lombardo, F., AghaKouchak, A., and Foufoula-Georgiou, E., 2018. "Precise temporal disaggregation preserving marginals and correlations (DiPMaC) for stationary and nonstationary processes". *Water Resources Research,* ***54***, pp. 7435–7458.
[27] Carpena, P., Bernaola-Galván, P. A., Gómez-Extremera, M., and Coronado, A. V., 2020. "Transforming Gaussian correlations: Applications to generating long-rang power-law correlated time series with arbitrary distribution". *Chaos,* ***30***, p. 083140.
[28] DNV, 2021. *Environmental Conditions and Environmental Loads*, september 2019 ed. DNV. DNV-RP-C205.
[29] Vanem, E., 2015. "Uncertainties in extreme value modeling of wave data in a climate change perspective". *Journal of Ocean Engineering and Marine Energy,* ***1***, pp. 339–359.
[30] Cheng, R., and Amin, N., 1983. "Estimating parameters in continuous univariate distributions with a shifted origin". *Journal of the Royal Statistical Society, Series B,* ***45***, pp. 394–403.
[31] Cousineau, D., 2009. "Fitting the three-parameter Weibull distribution: Review and evaluation of existing and new methods". *IEEE Transactions on Dielectrics and Electrical Insulation,* ***16***, pp. 281–288.
[32] Nagatsuka, H., Kamakura, T., and Balakrishnan, N., 2013. "A consistent method of estimation for the three-parameter Weibull distribution". *Computational Statistics & Data Analysis,* ***58***, pp. 210–226.
[33] Cousineau, D., 2009. "Nearly unbiased estimators for the three-parameter Weibull distribution with greater efficiency than the iterative likelihood method". *British Journal of Mathematical and Statistical Psychology,* ***62***, pp. 167–191.
[34] Ng, H., Luo, L., Hu, Y., and Duan, F., 2012. "Parameter estimation of the three-parameter Weibull distribution based on progressively Type-II censored samples". *Journal of Statistical Computation and Simulation,* ***82***, pp. 1661–1678.
[35] Teimouri, M., Hoseini, S. M., and Nadarajah, S., 2013. "Comparison of estimation methods for the Weibull distribution". *Statistics,* ***47***, pp. 93–109.
[36] Nwobi, F. N., and Ugomma, C. A., 2014. "A comparison of methods for the estimation of Weibull distribution parameters".

Metodološki zvezki, ***11***, pp. 65–78.
[37] Örkcü, H. H., Özsoy, V. S., Aksoy, E., and Dogan, M. I., 2015. "Estimating the parameters of 3-p Weibull distribution using particle swarm optimization: A comprehensive experimental comparison". *Applied Mathematics and Computation,* ***268***, pp. 201–226.
[38] Li, T., Griffiths, W., and Chen, J., 2017. "Weibull modulus estimation by the non-linear least squares method: A solution to deviation occuring in traditional *w*eibull estimation". *Metallurgical and Materials Transactions A,* ***48***, pp. 5516–5528.
[39] Vanem, E., Gramstad, O., and Bitner-Gregersen, E. M., 2019. "A simulation study on the uncertainty of environmental contours due to sampling variability for different estimation methods". *Applied Ocean Research,* ***91***, p. 101870.
[40] Papalexiou, S. M., 2022. "Rainfall generation revisited: Introducing CoSMoS-2s and advancing copula-based intermittent time series modelling". *Water Resources Reseearch,* ***58***, p. e2021WR031641.
[41] Papalexiou, S. M., Serinaldi, F., and Porcu, E., 2021. "Advancing space-time simulation of random fields: From storms to cyclones and beyond". *Water Resources Research,* ***57***, p. e2020WR029466.
[42] Stefanakos, C. N., and Vanem, E., 2018. "Nonstationary fuzzy forecasting of wind and wave climate in very long-term scales". *Journal of Ocean Engineering and Science,* ***3***, pp. 144–155.
[43] Vanem, E., Huseby, A. B., and Natvig, B., 2012. "Modeling ocean wave climate with a Bayesian hierarchical space-time model and a log-transform of the data". *Ocean Dynamics,* ***62***, pp. 355–375.
[44] Zanini, E., Eastoe, E., Jones, M. J., Randell, D., and Jonathan, P., 2020. "Flexible covariate representations for extremes". *Environmetrics,* ***31***, pp. e2624:1–20.
[45] Menéndez, M., Méndez, F. J., Izaguirre, C., Luceño, A., and Losada, I. J., 2009. "The influence of seasonality on estimating return values of significant wave height". *Coastal Engineering,* ***56***, pp. 211–219.
[46] Calderón-Vega, F., Vázquez-Hernández, A., and García-Soto, A., 2013. "Analysis of extreme waves with seasonal variation in the gulf of Mexico using a time-dependend GEV model". *Ocean Engineering,* ***73***, pp. 68–82.
[47] Jonathan, P., and Ewans, K., 2011. "Modeling the seasonality of extreme waves in the gulf of Mexico". *Journal of Offshore Mechanics and Arctic Engineering,* ***133***(2), pp. 021104:1–9.
[48] Vanem, E., 2015. "Non-stationary extreme value models to account for trends and shifts in the extreme wave climate due to climate change". *Applied Ocean Research,* ***52***, pp. 201–211.
[49] De Leo, F., Besio, G., Briganti, R., and Vanem, E., 2021. "Non-stationary extreme value analysis of sea states based on linear trends. analysis of annual maxima series of significant wave height and peak period in the Mediterannean sea". *Coastal Engineering,* ***167***, p. 103896.
[50] Serinaldi, F., and Kilsby, C. G., 2015. "Stationarity is undead: Uncertainty dominates the distribution of extremes". *Advances in Water Resources,* ***77***, pp. 17–36.
[51] Davis, R. A., and Mikosch, T., 2009. "The extremogram: A correlogram for extreme events". *Bernoulli,* ***15***(4), pp. 977–1009.
[52] Davis, R. A., Mikosch, T., and Cribben, I., 2012. "Towards estimating extremal serial dependence via the bootstrapped extremogram". *Journal of Econometrics,* ***170***, pp. 142–152.
[53] Horn, J.-T., Bitner-Gregersen, E., Krokstad, J., Leira, B. J., and Amdahl, J., 2018. "A new combination of conditional environmental distributions". *Applied Ocean Research,* ***73***, pp. 17–26.
[54] Haver, S., and Winterstein, S., 2009. "Environmental contour lines: A method for estimating long term extremes by a short term analysis". *Transactions of the Society of Naval Architects and Marine Engineers,* ***116***, pp. 116–127.
[55] Huseby, A. B., Vanem, E., and Natvig, B., 2013. "A new approach to environmental contours for ocean engineering applications based on direct Monte Carlo simulations". *Ocean Engineering,* ***60***, pp. 124–135.
[56] Huseby, A. B., Vanem, E., and Natvig, B., 2015. "Alternative environmental contours for structural reliability analysis". *Structural Safety,* ***54***, pp. 32–45.
[57] Vanem, E., 2018. "3-dimensional environmental contours based on a direct sampling method for structural reliability analysis of ships and offshore structures". *Ships and Offshore Structures,* ***14***, pp. 74–85.
[58] Vanem, E., 2023. "Analysing multivariate extreme conditions using environmental contours and accounting for serial dependence". *Renewable Energy,* ***202***, pp. 470–482.