Wave Extremes in the Northeast Atlantic

OLE JOHAN AARNES AND ØYVIND BREIVIK

Norwegian Meteorological Institute, and Geophysical Institute, University of Bergen, Bergen, Norway

MAGNAR REISTAD

Norwegian Meteorological Institute, Bergen, Norway

(Manuscript received 2 March 2011, in final form 12 July 2011)

ABSTRACT

The objective of this study is to compute 100-yr return value estimates of significant wave height using a new hindcast developed by the Norwegian Meteorological Institute. This regional hindcast covers the northeast Atlantic and spans the period 1958–2009.

The return value estimates are based upon three different stationary models commonly applied in extreme value statistics: the generalized extreme value (GEV) distribution, the joint GEV distribution for the *r* largest-order statistic (*r*LOS), and the generalized Pareto (GP) distribution. Here, the qualitative differences between the models and their corresponding confidence intervals are investigated.

1. Introduction

Wind-generated ocean waves are, in many cases, the most critical factor in design of offshore structures and coastal development. The everyday strain inflicted by waves weakens most of the local construction and shapes the coastal landscape. However, the biggest concern is often related to storm events when wave loads may become catastrophic. As presented by Alves and Young (2003), Caires and Sterl (2005), and Semedo et al. (2011), the highest wave conditions globally are found in the northeast Atlantic, making this area particularly interesting for extreme value analyses.

Considerable effort has been made to obtain accurate return value estimates of significant wave height, both locally and globally; see Soukissian and Kalantz (2006) for a review of earlier works. These estimates are typically based upon one of two closely related families of extreme value distributions, either the generalized extreme value (GEV) family or the generalized Pareto (GP) family (see Coles 2001), where the type of data extraction determines the family to be applied. With the GEV distribution, only block maxima are retained from the initial dataset (e.g., the annual maximum). This is quite a wasteful approach and therefore requires a substantial dataset. Alternatively, a fixed number *r* of the highest peaks within each block can be extracted to utilize the joint GEV distribution for the *r* largest-order statistic (see, e.g., Soares and Scotto 2004). A third option is the so-called peaks-over-threshold (POT) approach, where all uncorrelated peaks above some predetermined threshold are retained. The data subset will vary in size according to the level of the threshold, but should conform to the GP distribution. In the end, there is no consensus on a method that is superior in all cases. Sometimes the choice is dictated by the data at hand. However, more often the discrepancy or agreement between the different approaches is investigated as a measure of confidence in the return value estimates.

Although in situ wave measurements rely on different techniques and instruments, they still represent the best estimates of the ground truth. Unfortunately, they are also sparsely distributed geographically, contain gaps, and often span short periods of time, and are therefore not always adequate for extreme value analysis. Satellite altimetry is really the only other source of wave height data, besides numerical models, that offers satisfactory spatial resolution and global coverage. However, as polar-orbiting satellites revisit the same location once every 10–35 days, the temporal resolution is very poor for wave measurement. Within the context of extreme value statistics, this problem is addressed by Cooper and

Corresponding author address: Ole Johan Aarnes, Norwegian Meteorological Institute, Allégaten 70, 5007 Bergen, Norway.
E-mail: ole.aarnes@met.no

DOI: 10.1175/JCLI-D-11-00132.1

Forristall (1997), Panchang et al. (1999), and Anderson et al. (2001). A common approach has been to bin data into larger geographical areas, (e.g., $2° \times 2°$), with assumed statistical homogeneous wave conditions. This is illustrated in the work of Carter (1993), who fitted the Fisher–Tippet-type 1 (FT-1) distribution to all data (i.e., the initial distribution method, IDM), to obtain 50-yr return value estimates covering the northeast Atlantic. Wimmer et al. (2006) focused on the same area and reported up to a 37% reduction in their corresponding estimates going from the IDM–FT-1 to the POT–GP combination. Two different approaches were applied in Alves and Young (2003) to obtain global 100-yr return value estimates of significant wave height, H_{s100}. They found the IDM–FT-1 more suitable than the POT–three-parametric Weibull distribution (3PW) when working with satellite data.

Attractive alternatives to in situ and remote sensing data are modeled reanalyses or hindcasts as these datasets offer regularity both in time and space. Caires and Sterl (2005) based their extreme value analysis on the 45-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40; Uppala et al. 2005) to obtain global H_{s100} estimates. By utilizing the POT approach with the threshold set at the 93th percentile of the initial data, they assumed the retained data to conform to the exponential distribution—one of the three forms incorporated in the GP family. The final result was adjusted by a linear relation found between the return value estimates based on the reanalysis and the available buoy observations. In areas where the assumption of exponentiality was found to be inapplicable, estimates were usually found to be conservative (i.e., too high). Williams (2005) performed a similar analysis based on the NEXT Re-Analysis (NEXTRA) hindcast, a revised and updated version of the North European Storm Study (NESS) hindcast (Peters et al. 1993). This dataset covers U.K. waters and spans the period 1964–98, but omits several summer months during the period. The H_{s100} estimates were based on the 3PW distribution fitted to all data above the 95th percentile and calibrated to observations and altimeter data. Compared to the results of Caires and Sterl (2005), the H_{s100} estimates are significantly smaller, especially within the North Sea.

The main objective of this study is to compute H_{s100} estimates covering the northeast Atlantic using a new hindcast developed by the Norwegian Meteorological Institute. The hindcast is presented in section 2. The H_{s100} estimates will be based on three different statistical models utilizing different subsets of the initial data, that is, the annual maximum, the r largest-order statistic, and the peaks over threshold, which we present in section 3. In section 4 we present the results of the different approaches, while the discrepancies between the estimates are discussed in section 5. Finally, conclusions are given in section 6.

2. Data

Norwegian Reanalyses 10 km (NORA10)

The 10-km Norwegian Reanalyses (NORA10) make up the latest contribution to a series of wave hindcasts developed by the Norwegian Meteorological Institute (see Reistad et al. 2007, 2011). This regional hindcast is a dynamical downscaling of the ERA-40 dataset (Uppala et al. 2005), producing 3-hourly wave fields at 10–11-km grid spacing. The atmospheric forcing is obtained with the 10-km High-Resolution Limited Area Model (HIRLAM10; Undén et al. 2002). Temperature, wind velocity, specific humidity, and liquid water in the boundary zone are relaxed toward ERA-40, while some of the large-scale features are maintained using a digital filter. Sea surface temperatures are interpolated from the ERA-40 dataset or the ice data archive at the Norwegian Meteorological Institute. For wave simulations a modified version of the wave modelling (WAM) cycle 4 (Komen et al. 1994), is run on the same grid as the HIRLAM10, nested inside a WAM model at 50-km resolution forced by ERA-40 winds (Fig. 1). NORA10 covers the northeast Atlantic, including the North Sea, the Norwegian Sea, and the Barents Sea.

The ERA-40 dataset spans the period September 1957 to August 2002. However, NORA10 is continually being extended using operational analyses from the ECMWF as boundary and initial conditions. In the following, we focus on the period 1958–2009, a total of 52 yr. To ensure that NORA10 does not possess a discontinuity of significance, we have made a statistical comparison of the modeled and observed H_s at three locations in the North Sea and the Norwegian Sea: Ekofisk, Gullfaks C, and Draugen (see Fig. 1). The Ekofisk data are obtained with a Waverider, while the latter two are obtained with a platform-mounted radar (Miros). Poor quality observations have been coarsely filtered by excluding all data outside the interval $0.5H_s$–$2H_s$ of NORA10, and only data obtained at corresponding hours have been used. Figure 2 presents the annual bias and the annual discrepancy in the 95th percentile between 1994 and 2008. There are no abrupt changes in the statistics during this period. Therefore, any discontinuity present is to be insignificant in the course of this work and should not affect the final results.

The added value of running a regional hindcast is clearly manifested in the superior performance of NORA10 relative to ERA-40. The model validation presented by Reistad et al. (2011) indicates no need for calibration similar to what was performed by Caires and

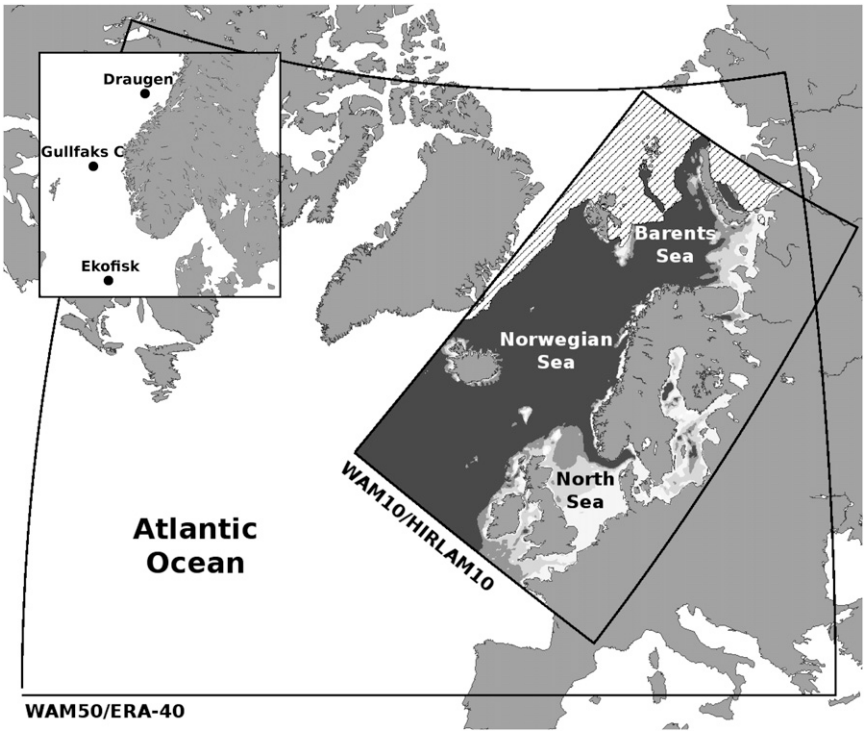

FIG. 1. Model setup and nesting. Outer boundaries represent the model domain of the coarser WAM model, forced by ERA-40 winds. Inner boundaries enclose the model domain of NORA10, forced by HIRLAM10 winds. Filled contours represent the bathymetry: 0–50 m, white; 50–100 m, light gray; 100–150 m, gray; and below 150 m, dark gray. The hatched area represents the ice coverage by 1 June 2011. In the top-left corner, the oil rigs Ekofisk, Gullfaks C and Draugen are shown offshore of Norway.

Sterl (2005). However, there is a need to establish what the following H_{s100} estimates represent in terms of duration at a point specific location. With a finite model resolution of 10–11 km, the crossing of a grid cell will vary approximately with the mean wave period of the wave system. Generally, the higher the H_s, the longer the mean wave period and hence the faster the crossing, governed by the dispersion relation. As the wave climate will vary substantially within the model domain, it is not straightforward to give a universal duration estimate.

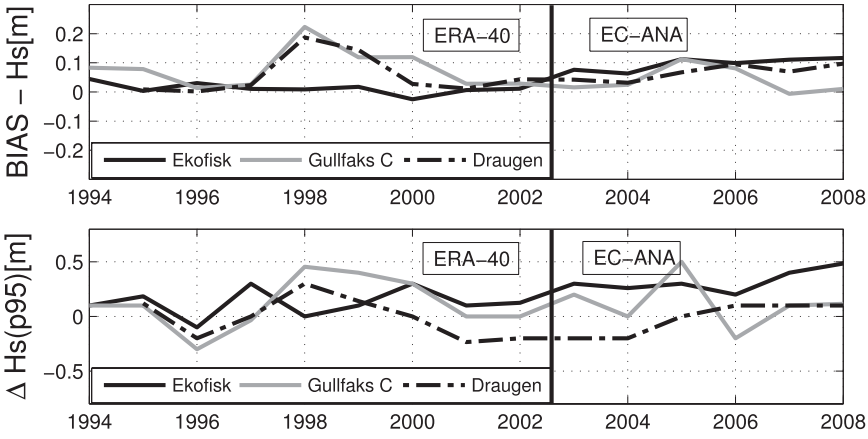

FIG. 2. (top) Time series of annual bias in H_s (NORA10-OBS). (bottom) Annual discrepancy in the 95th percentile of H_s (NORA10-OBS). The statistics are based on 3-hourly data at corresponding hours. August 2002 marks the transition in NORA10, going from initial and boundary conditions obtained with the ERA-40 to the ECMWF analysis.

TABLE 1. Statistical comparison of modeled and observed H_s for $H_s > 0$, $H_s > H_s(p_{95})$ and $H_s > H_s(p_{99})$ at Ekofisk (2001–2009), Gullfaks C (1999–2009), and Draugen (1996–2009). The 3-hourly NORA10 data are validated against the maximum observed H_s(20 min) within ±1.5 h of NORA10, the mean observed H_s over periods of 20 min and 1, 3, and 6 h, centered at the time of NORA10. Statistical measures shown are the scatter index (SI, %), the NORA10-OBS bias (m), the correlation coefficient (R), and the regression line of NORA10 = $a + b \times$ obs.

		Ekofisk				Gullfaks C				Draugen			
	Obs period	SI	Bias	R	$a + bx$	SI	Bias	R	$a + bx$	SI	BIAS	R	$a + bx$
$H_s > 0$	max 20 min	16.50	−0.17	0.97	$0.07 + 0.89x$	16.29	−0.24	0.95	$-0.01 + 0.92x$	20.83	−0.18	0.94	$0.15 + 0.88x$
	20 min	17.54	0.07	0.96	$0.03 + 1.02x$	17.26	0.01	0.95	$0.08 + 0.97x$	21.17	0.05	0.94	$0.23 + 0.93x$
	1-h mean	16.03	0.06	0.97	$0.00 + 1.03x$	16.60	0.01	0.96	$0.06 + 0.98x$	20.42	0.06	0.94	$0.23 + 0.94x$
	3-h mean	15.48	0.06	0.97	$-0.01 + 1.04x$	15.86	0.01	0.96	$0.04 + 0.99x$	19.60	0.05	0.95	$0.22 + 0.94x$
	6-h mean	15.23	0.06	0.97	$-0.04 + 1.05x$	15.07	0.01	0.96	$-0.01 + 1.01x$	18.31	0.05	0.95	$0.18 + 0.95x$
$H_s > H_s(p_{95})$	max 20 min	11.33	−0.56	0.82	$1.05 + 0.73x$	11.21	−0.43	0.79	$-0.18 + 0.96x$	12.30	−0.56	0.78	$-0.24 + 0.96x$
	20 min	12.44	0.10	0.78	$0.48 + 0.93x$	12.76	0.02	0.76	$-0.06 + 1.01x$	13.92	−0.10	0.75	$-0.35 + 1.04x$
	1-h mean	10.47	0.16	0.85	$0.22 + 0.99x$	12.22	0.05	0.79	$-0.35 + 1.06x$	13.29	−0.07	0.78	$-0.56 + 1.07x$
	3-h mean	9.76	0.18	0.87	$0.19 + 1.00x$	11.70	0.09	0.82	$-0.56 + 1.10x$	12.84	−0.06	0.80	$-0.60 + 1.08x$
	6-h mean	9.88	0.25	0.86	$0.23 + 1.00x$	11.75	0.20	0.82	$-0.61 + 1.13x$	12.85	0.02	0.80	$-0.59 + 1.09x$
$H_s > H_s(p_{99})$	max 20 min	11.18	−0.80	0.73	$2.20 + 0.61x$	9.86	−0.45	0.71	$0.56 + 0.88x$	10.48	−0.53	0.73	$0.35 + 0.90x$
	20 min	11.93	0.02	0.81	$0.97 + 0.86x$	11.17	0.04	0.65	$0.65 + 0.93x$	11.82	0.02	0.69	$0.00 + 1.00x$
	1-h mean	9.43	0.13	0.78	$0.17 + 0.99x$	10.59	0.13	0.70	$0.19 + 0.99x$	11.54	0.08	0.73	$-0.08 + 1.02x$
	3-h mean	8.68	0.16	0.82	$0.37 + 0.97x$	9.91	0.22	0.73	$0.34 + 0.99x$	11.29	0.14	0.73	$0.32 + 0.98x$
	6-h mean	9.27	0.26	0.79	$0.29 + 1.00x$	10.36	0.42	0.71	$0.97 + 0.93x$	11.48	0.35	0.71	$0.74 + 0.95x$

However, in Table 1 we present a model validation of NORA10 at the same locations as are presented above. At this stage we include all available observations (20-min averages) recorded every 20 min at Ekofisk and every 10 min at Gullfaks C and Draugen. The 3-hourly NORA10 data have been validated against the maximum observed H_s within a time window of ±1.5 h, H_s at the corresponding hour, and the 1-, 3-, and the 6-h means. Similar statistical comparisons have been conducted using all available data ($H_s > 0$) and cases where the observed H_s is higher than the corresponding 95th [$H_s > H_s(p_{95})$] and 99th percentiles [$H_s > H_s(p_{99})$]. In general, we see that NORA10 is biased low against the maximum H_s in all three cases. For all data, the NORA10 validates the best result being somewhere between the 3- and 6-h means. For H_s above the 95th percentile, the best result is achieved closer to the 3-h mean. This becomes even more evident when using data exclusively above the 99th percentile, which underlines the remarks made above. Now, as the H_{s100} estimates will be well above the 99th percentile of the data, it is very likely that these events will validate better against means taken over even shorter time windows. We therefore assume that the following H_{s100} estimates will represent approximately a 1-h-mean sea state.

3. Method

The following analysis is primarily based on the extreme value theory presented in Coles (2001). Here, we use three closely related statistical models to obtain H_{s100} estimates. A common denominator for all approaches is the assumption of an independent and identical distributed (IID) time series. While the criterion of independence is discussed in the following, the violation of the homogeneity principle is addressed in section 5.

Each initial 3-hourly time series needs to be reduced to a subset of uncorrelated data entries, or peak events. A pragmatic way of dealing with this issue is to require a minimum time interval between each entry. Here, we use 48 h to decluster the dataset (Caires and Sterl 2005; Lopatoukhin et al. 2000). This exceeds the average time scale of the passing of an extratropical cyclone and should prevent data representing the same weather system. The subset is further reduced by only retaining the highest events. This is traditionally done either by retaining a constant number (r) of the highest entries per block, where a block refers to a year, or by retaining all peaks above some predefined threshold, the POT approach. With $r = 1$, the former reduces to the annual maximum (AM), while $r > 1$ is known as the r largest-order statistic (rLOS). These three data subsets (AM, rLOS, and POT) descend from the same initial dataset, but conform to different distributions and require different attention.

a. The AM model

Let $X_i = X_1, \ldots, X_n$ represent a random sequence of independent variables with the common distribution function F (IID), then the distribution of the block maximum $M_n = \max(X_i)$ can be expressed by

$$\Pr(M_n \le z) = \Pr(X_1 \le z) \times \cdots \times \Pr(X_n \le z) = [F(z)]^n. \tag{1}$$

It follows that if z_+ represents the smallest value of z, where $F(z) = 1$, then $F^n(z) \to 0$ for all $z < z_+$ when $n \to \infty$; that is, the distribution of M_n degenerates to a point mass on z_+. To avoid this difficulty, we renormalize M_n by

$$M_n^* = \frac{M_n - \mu_n}{\sigma_n} \tag{2}$$

and seek a combination of constants ($\sigma_n > 0$) and μ_n that stabilizes the distribution of M_n^* such that

$$\Pr(M_n^* < z) \to G(z). \tag{3}$$

Independent of the form of F, it can be shown (Coles 2001) that $G(z)$ must take the form

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z - \mu_n}{\sigma_n}\right)\right]^{-1/\xi}\right\}, \tag{4}$$

where σ_n and μ_n represents the scale and location parameter of G, respectively, while ξ is known as the shape parameter. The distribution of G is categorized into three classes of distributions depending on ξ: Fréchet ($\xi > 0$), Gumbel ($\xi = 0$), or reversed Weibull ($\xi < 0$). While the latter is approaching an asymptotic limit on z, the former two are unbounded. Together, they are known as the family of the generalized extreme value distribution (GEV).

The parameter estimates are obtained by maximizing the likelihood function L or equivalently the log-likelihood function ℓ defined by

$$\ell(z_i;\theta) = \log L(z_i;\theta) = \sum_{i=1}^{n} \log f(z_i;\theta), \tag{5}$$

where $\theta(\xi, \sigma, \mu)$ represents the parameter vector, f represents the probability density function of the statistical model and z_i are realizations of the same model (i.e., observed block maxima). This is solved iteratively and is known as the maximum likelihood approach.

b. The rLOS model

Again, we consider n instances of the IID variable $X_i = X_1, \ldots, X_n$, but now the extracted data subset is expanded to contain the r largest-order statistic. Even though the expression for the joint density function for M_n^r and the log-likelihood function differ somewhat from the GEV, it can be shown that the parameter estimates correspond to those of the GEV distribution (i.e., ξ, σ, and μ). See Coles (2001) for further details.

By including more data, we hope to improve the fit between the data and the statistical model. However, the choice of r will be a trade-off between variance and bias. For small r [e.g., $r = 1$ (AM)], the variance between the data and model is expected to be high, while larger values of r are subject to increased bias. Here, the choice of r is based on the likelihood ratio test (Soares and Scotto 2004; Coles 2001), defined by

$$D = 2[\ell(M_1) - \ell(M_0)] \sim \chi_1^2, \tag{6}$$

where $\ell(M_1)$ and $\ell(M_0)$ represent the maximized log-likelihood function for the rLOS model of $r + 1$ and r, respectively. We increase r until the model of M_0 is a valid representation of the model M_1, that is, when $D < c_\alpha$, where c_α is the $(1-\alpha)$ quantile of the χ_1^2-distribution($\alpha = 0.05$).

c. The POT model

If the block maximum $M_n = \max(X_i)$ has an approximate distribution of the GEV, given that Eq. (3) is satisfied, it can be shown that for large enough u the cumulative distribution function of $y = X_i - u$ for $y > 0$ is approximately given by

$$H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi} \quad \text{and} \tag{7}$$

$$\tilde{\sigma} = \sigma + \xi(u - \mu). \tag{8}$$

This is known as the generalized Pareto distribution. Similar to the GEV, the GP takes three forms depending on ξ. For $\xi = 0$, the GP reduces to the exponential distribution.

We have explored several options on how to set the threshold in the POT model, which will be revisited in section 5. However, our final choice is primarily based on the Anderson–Darling goodness-of-fit test. This test quantifies the fit between a statistical distribution and a set of data, and is especially suited for extreme value distributions as the conformity at the tail is heavily weighted. The test statistic is defined as

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i - 1)\{\log[q_{(i)}] + \log[1 - q_{(n+1-i)}]\}, \tag{9}$$

where $q_{(i)} = H(y_{(i)})$. The null hypothesis H_0 states that $y_1, \ldots, y_n$ originates from Eq. (7). For H_0 to hold, the upper-tail asymptotic percentage points or the critical values of A^2 need to exceed the test statistic given in Eq. (9). For the GP distribution, these critical values were first established by Choulakian and Stephens (2001) and vary with ξ when both ξ and σ are unknown. In Table 2,

TABLE 2. Critical value $c_{0.05}(\xi)$ for the Anderson Darling test statistic at the 5% significance level $\Pr[A^2 \geq c_{0.05}(\xi)]$ for the GP, taken from Choulakian and Stephens (2001).

ξ	−0.5	−0.4	−0.3	−0.2	−0.1	0.0	0.1	0.2	0.5	0.9
$c_{0.05}(\xi)$	1.321	1.221	1.140	1.074	1.020	0.974	0.935	0.903	0.830	0.771

the critical values $[c_{0.05}(\xi)]$ are presented at the 5% significance level.

Confidence intervals of the H_{s100} estimates are obtained with the profile likelihood approach, defined by Eq. (6). By solving Eq. (4) on behalf of μ, the corresponding log-likelihood equations can be reformulated as a function of $\theta_i(z, \sigma, \xi)$. By fixing z at different wave heights or return levels, the log-likelihood function can be maximized in the usual way. If $\ell(M_1)$ represents the maximum log-likelihood function and $\ell(M_0)$ represents

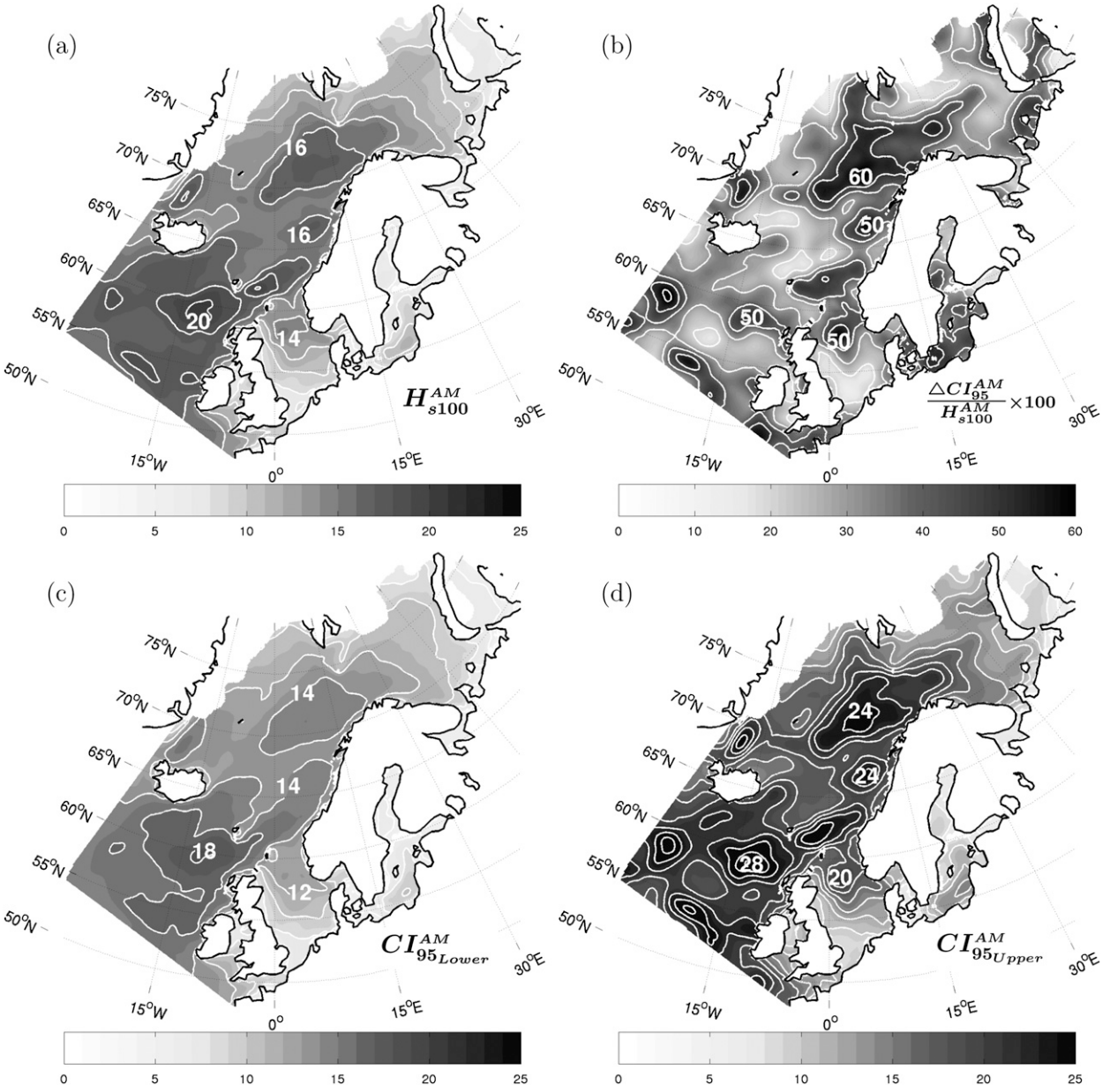

FIG. 3. Results of the AM model: (a) H_{s100} (m), (b) the width of the 95% confidence interval relative to the H_{s100} estimate given in percent, (c) the 2.5% confidence limit, and (d) 97.5% confidence limit of H_{s100}. Notice that some areas slightly exceed the gray scale.

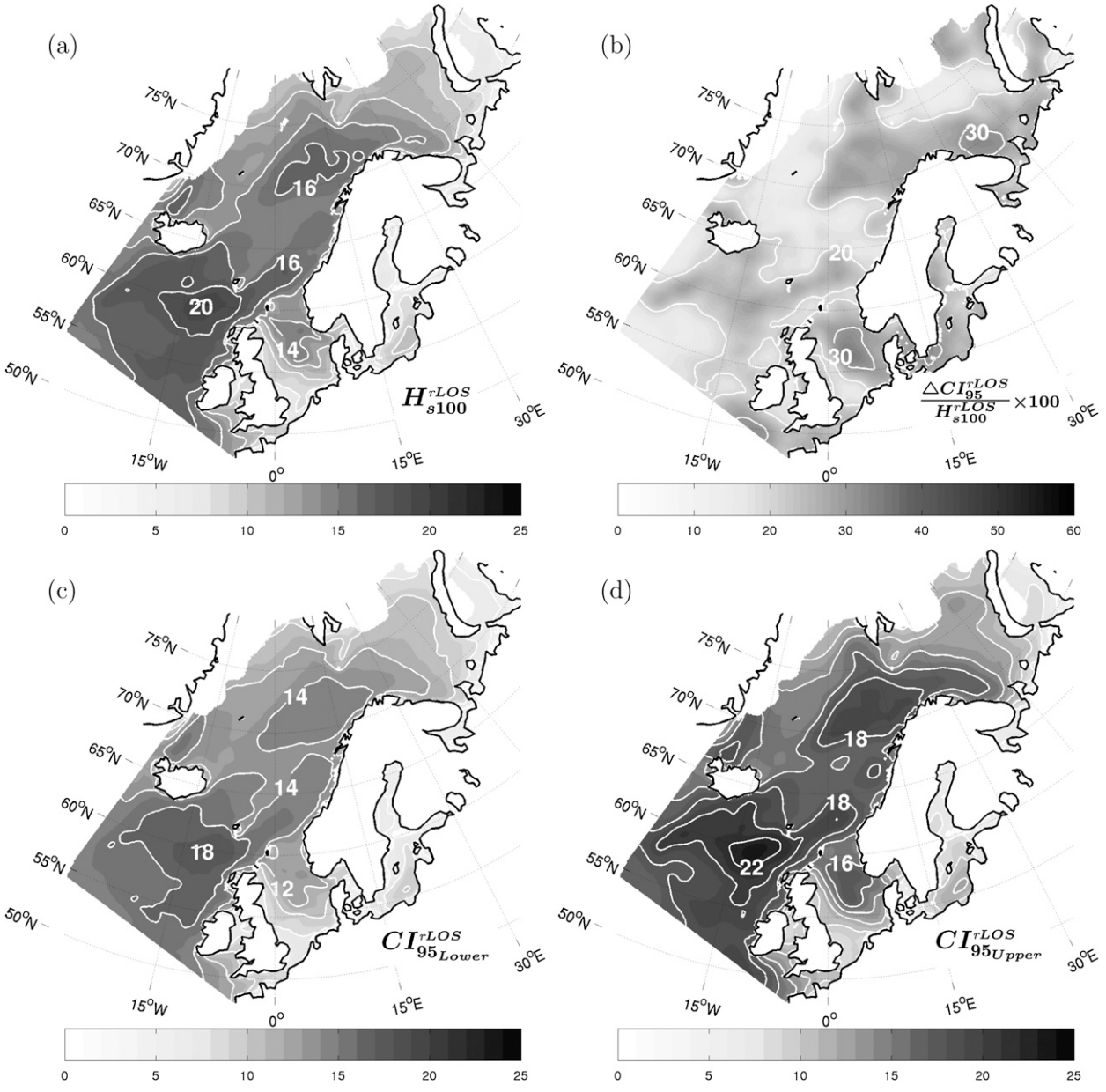

FIG. 4. As in Fig. 3, but for the *r*LOS model (except no areas exceed the gray scale).

the reformulated log-likelihood function, then the $1 - \alpha$ confidence interval at z is defined as $\mathrm{CI}_\alpha = [\theta_i: D(\theta_i) \leq c_\alpha]$, where $c_\alpha = 3.84146$ at the 95% quantile ($\alpha = 0.05$) of the χ_1^2 distribution. An equivalent procedure applies for the GP distribution. For more details, see Coles (2001).

4. Results

Here, we present the H_{s100} estimates and the corresponding 95% confidence intervals based on the three different approaches. All gray scales are held constant and span the interval 0–25 m, with white isolines given every 2 m. We also include a plot of the relative uncertainty, defined by the width of the confidence interval relative to the H_{s100} estimate, presented as a percentage and with white isolines every 10%. All plots have been smoothed with a mean filter, that is, assigning the mean value of a $(2N + 1)$-by-$(2N + 1)$ grid matrix to each center grid point. In panels b and d of Figs. 3–5 $N = 6$; otherwise, $N = 3$.

The results of the AM model are presented in Fig. 3 and will be used as a benchmark in the following. The main features of the H_{s100} estimate are the global maximum located southwest of the Faroe Islands, peaking just above 21 m; a branch of the global maximum extending through

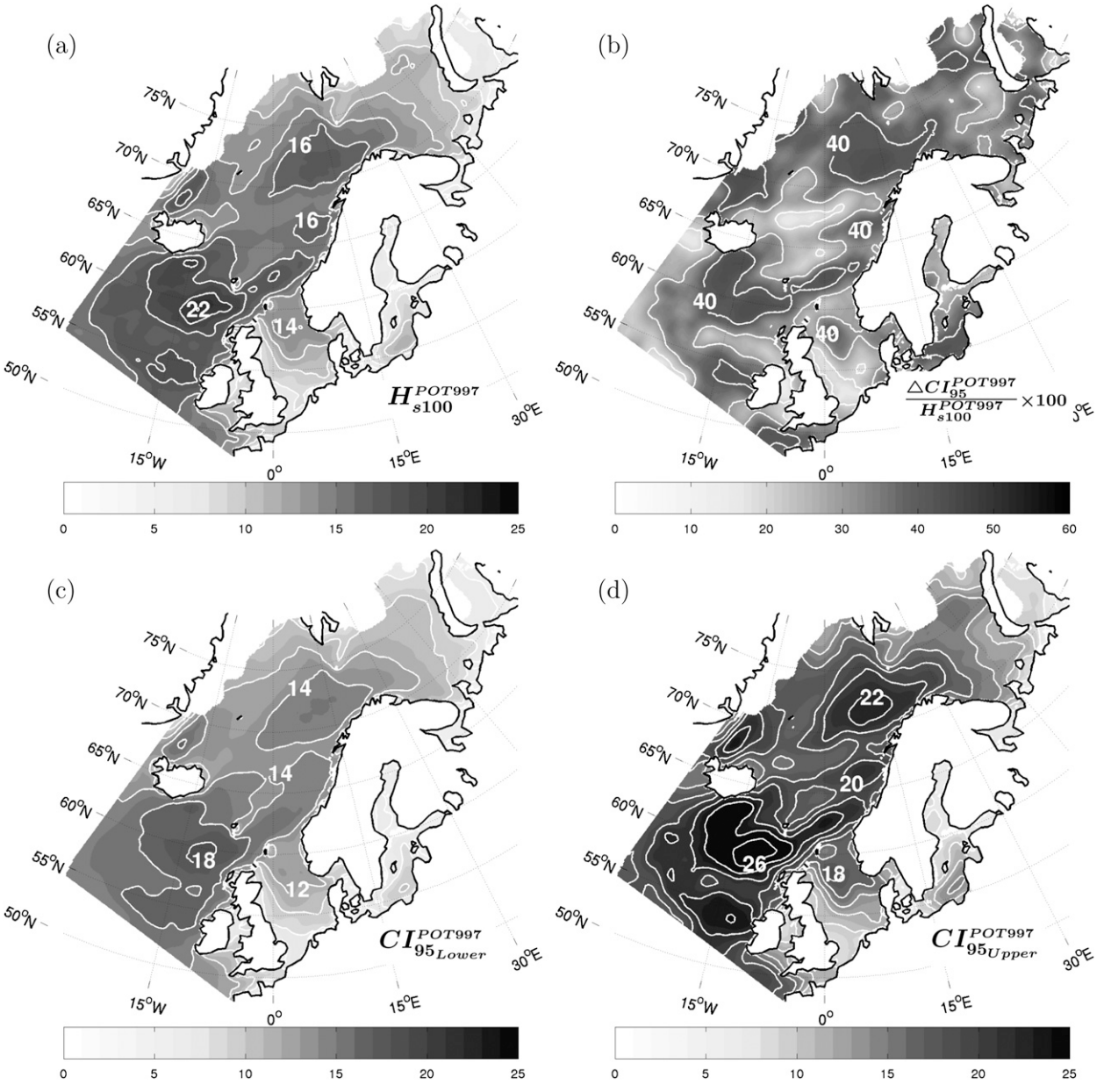

FIG. 5. As in Fig. 3, but for the POT997 model.

the Faroe–Shetland Channel toward the Norwegian coast; a local maximum in the central North Sea; and another maximum in the northern Norwegian Sea. Somewhat unexpectedly, we find a local minimum in the central parts of the Norwegian Sea, a feature that contradicts the general wave climate (i.e., a wave field gradually decreasing going north-northeast). In about 77% of all grid points the shape parameter is negative ($\xi < 0$; i.e., a Weibull type), and H_s is bounded above.

The uncertainty associated with the AM model is substantial. At most, the width of the confidence interval is 60% of the best estimate. The majority of the uncertainty is affiliated with the upper level of the confidence interval, which is a result of the positively skewed χ^2-distributed profile likelihood. Notice that the main features of the relative uncertainty are highly comparable to the pattern we see in the H_{s100} estimate.

In Fig. 4 we present the results of the *r*LOS model. These estimates are based on a constant number *r* of the highest-order statistic per year per grid point. Our choice of *r* is determined with the likelihood ratio test and lies between 2 and 3 in the North Sea–Barents Sea, 3 and 5 in

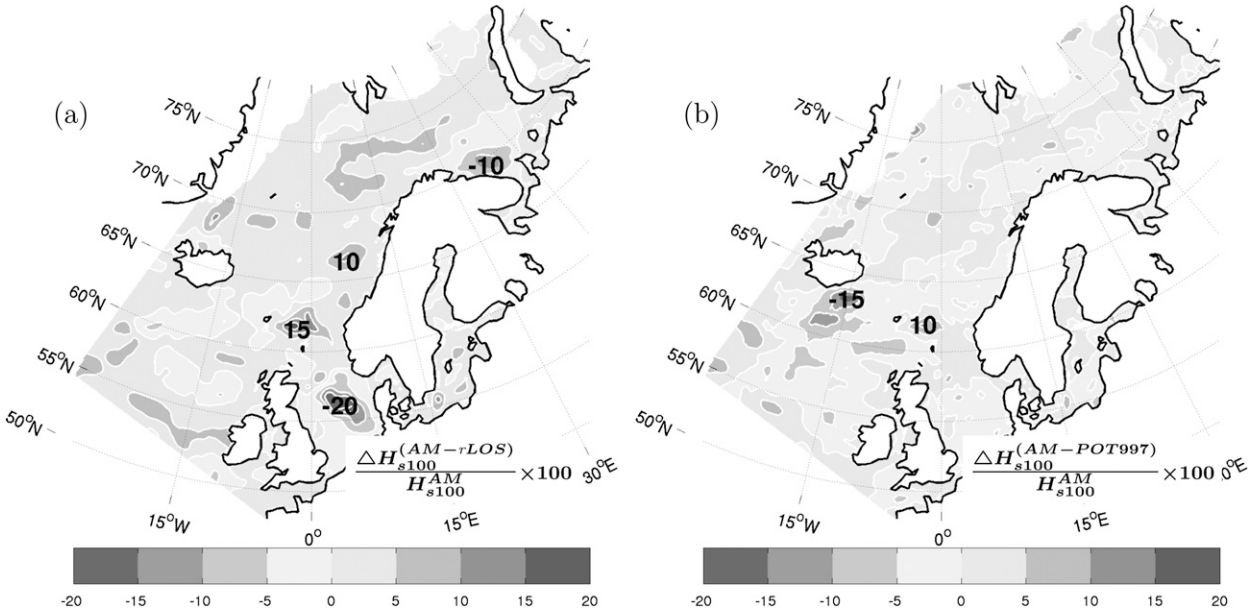

FIG. 6. (left) Discrepancy between the H_{s100} estimates of AM and *r*LOS, relative to AM. (right) Discrepancy between the H_{s100} estimates of AM and POT997, relative to AM, given in percent.

the Norwegian Sea, and 4 and 6 south of Iceland. To investigate the quantitative difference to the AM model, we have plotted the discrepancy between the H_{s100} results of the AM model and the *r*LOS model, relative to the AM model (Fig. 6a). Overall, the H_{s100} estimates of the AM model exceed those of the *r*LOS model and differ by less than ±5% for the most part. The largest deviation is found in the central North Sea, where the *r*LOS model exceeds the AM model by about 20%.

Compared to the AM model, we now see a much narrower 95% confidence interval, where the global maximum spans an interval of 18–24 m versus 18–30 m for the AM model. The relative uncertainty of the *r*LOS model is most pronounced in the central parts of the North Sea and just northeast of the Kola Peninsula, where the width of the confidence interval constitutes about 30% of the best estimate. These areas have a positive shape parameter ($\xi > 0$); otherwise, approximately 95% of all grid points have a negative shape parameter.

Figure 5 represents the results of the POT model. Based on the Anderson–Darling goodness-of-fit method, we have set the threshold at the 99.7th percentile of the initial data. Below the 95th percentile, the GP distribution is rejected at a 5% significance level at all grid points. At the 99th percentile there is scattered rejection, with more or less total rejection in the North Sea and the Baltic Sea. The GP distribution is almost fully accepted at the 99.9th percentile, with the exception of the Baltic Sea. However, to limit both the variance and the bias between the model and the data, we have set the threshold at the 99.7th percentile of the initial data (POT997). This leaves a total of 100–150 entries per grid point.

In general there are only small deviations between the H_{s100} estimates of the POT997 and the AM (i.e., within ±5% of the AM estimate). However, two areas stand out, an area east of the Faroe Islands, where POT997 is approximately 10% smaller than AM, and another area southeast of Iceland, where POT997 exceeds AM by ~15% (Fig. 6b). The POT997 model produces tighter confidence intervals, and as with the AM model, the highest uncertainty is related to the areas with the highest return value estimates.

About 30% of all grid points have a positive shape parameter; otherwise, H_s is bounded above. The total area of grid points having a negative shape parameter for the three different models is summarized in Table 3.

Figure 7 illustrates the model diagnostics at six locations, presented from north to south. These are primarily chosen because they represent areas of high deviation between at least two of the models (Fig. 6). We have also included the position 67.98°N, 02.21°E (Fig. 7c), as it is located within the local minimum of the H_{s100} estimates

TABLE 3. Percentage of grid points having a negative shape parameter ($\xi < 0$, bounded above) according to the three different models.

AM	*r*LOS	POT997
77.2	94.8	70.5

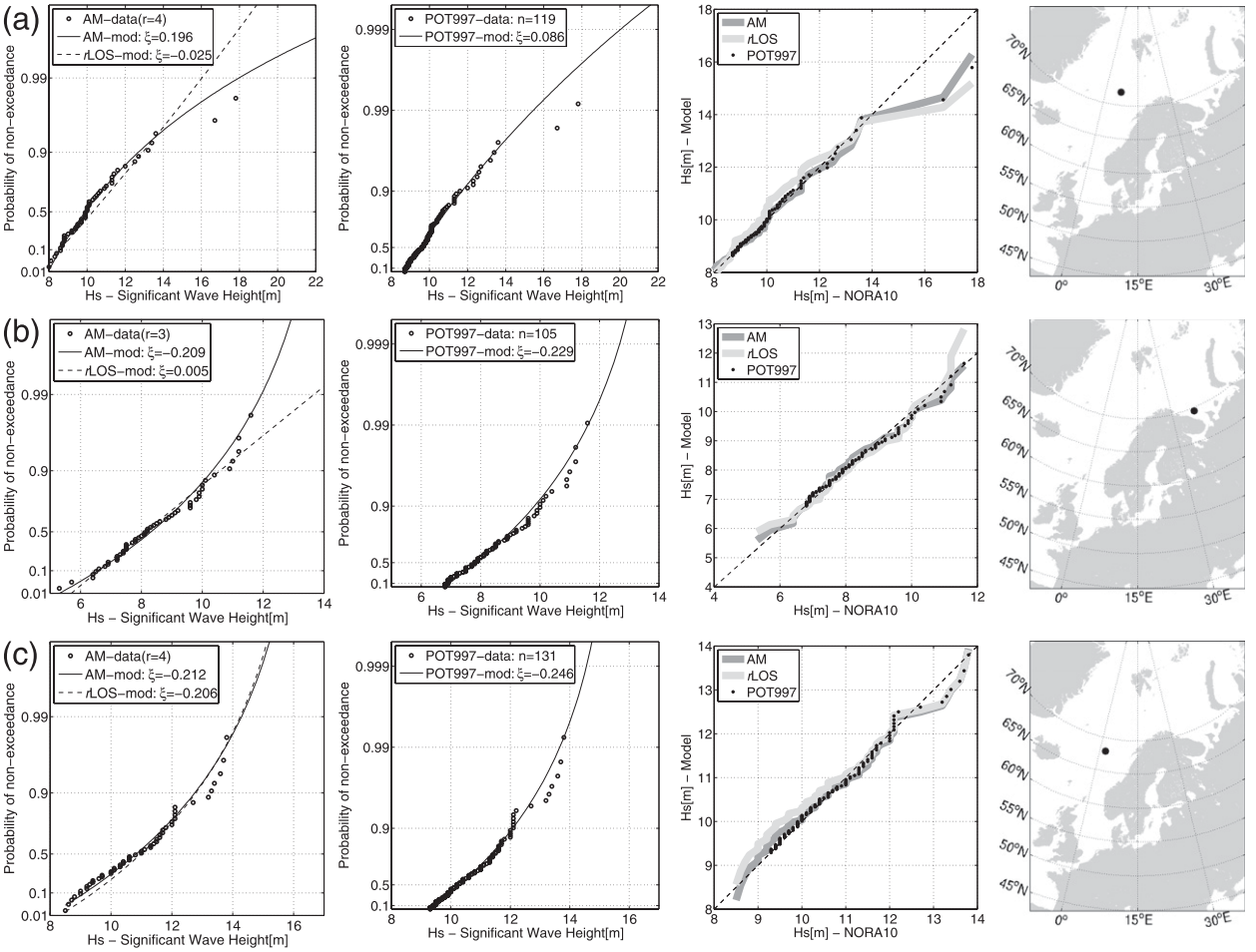

FIG. 7. Model diagnostics at (a) 71.92°N, 7.24°E; (b) 69.54°N, 39.43°E; (c) 67.98°N, 2.21°E; (d) 63.07°N, 15.47°W; (e) 61.88°N, 2.56°W; and (f) 56.52°N, 3.25°E. (left to right)) Return value plots of the AM–*r*LOS model together with the AM data, where ξ represents the shape parameter and r is the number of the largest-order statistic, return value plots of the POT997 model together with the POT997 data where n represents the total number of entries and ξ is the shape parameter, quantile plots of the three models, and the geographical location.

found in the Norwegian Sea (Figs. 3a, 4a, and 5a). In the first column we present the return value plots of the AM and *r*LOS models plotted together with the AM data. In the next column the POT997 model is presented together with the POT997 data. In both cases each entry is assigned the probability rank/(n + 1), where n represents the total number of data points. In the third column the three models are compared in quantile plots, while the last column shows the geographical location of the comparison.

The sensitivity of the three different models may be illustrated by a bootstrap experiment. In the following we have constructed 1000 resamples based on random draws with replacement from the original data subsets at two locations—71.92°N, 7.24°E and 67.98°N, 2.21°E—presented in Figs. 7a,c, respectively. Each resample can be illustrated by its own return value plot, providing an alternative way of defining the 95% confidence interval. As bootstrap samples have a tendency to generate shorter tails than the true sample distribution, we follow the example of Coles and Simiu (2003) and apply a bias correction to the parameter estimates of each resample. This ensures that the bootstrap mean coincides with the best estimate (maximum likelihood) of each of the original data subsets. In Fig. 8 the individual return value plots are presented together with the best estimates and the confidence intervals based on the bootstrap procedure and the profile likelihood. Notice that the bootstrap produces a slightly more symmetrical confidence interval around the best estimate, while the

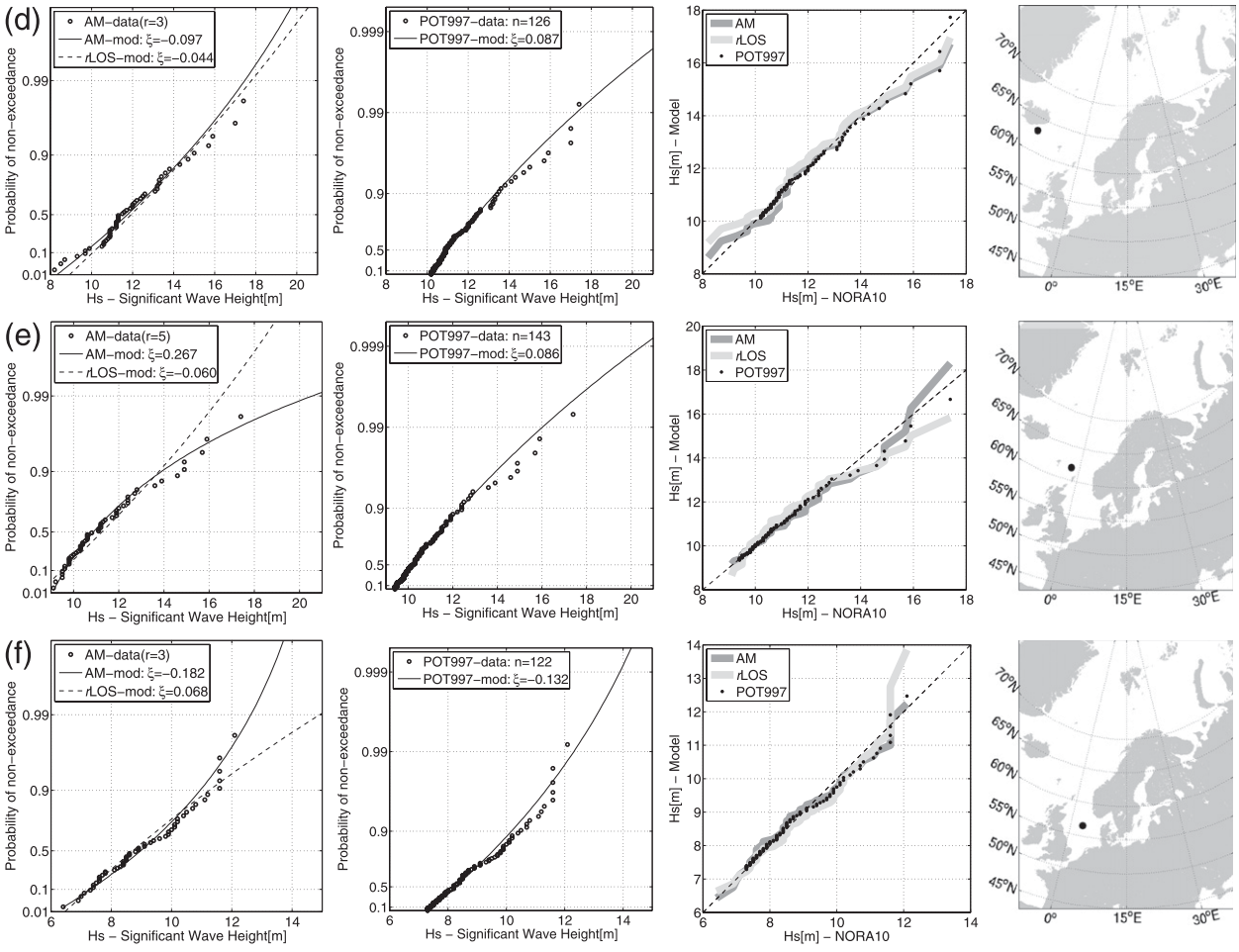

FIG. 7. (*Continued*)

profile likelihood is skewed toward higher H_s. The relative widths of the two confidence intervals seem to have a dependence on the shape parameter. For $\xi \leq 0$, the profile likelihood produces wider confidence intervals than its counterpart, while the opposite is true for $\xi \geq 0$.

5. Discussion

Of the three approaches applied in this study, the AM model is the most wasteful. Here, 52 yr of data are reduced from 151 944 to 52 entries per grid point, ~0.03% of the initial dataset. This data subset is easy to handle and easy to obtain, but decreases our confidence in the return value estimates. In several areas the relative width of the 95% confidence interval constitutes more than 60% of the best estimate (Fig. 3b). Nor does the AM model account for any year-to-year variation. A single year can have several wave events higher than the next; still, only one entry is retained. In that way important information may get censored. Taking into consideration that the AM model is highly influenced by individual storms, particularly the strongest ones, it is fair to conclude that the approach should be applied with some care. Nevertheless, we find that the AM model provides a relatively good fit in areas where the discrepancies among the other models are most pronounced, (Figs. 7a,b,e,f).

With the *r*LOS model, we retain three or more entries per year over most of the model domain, letting fewer severe storms go unnoticed. This clearly tightens the 95% confidence interval compared to the AM model, almost reducing it in half (i.e., peaking just above 30% of the best estimate). For the most part the *r*LOS model provides somewhat lower H_{s100} estimates than the AM model, and primarily within ±5% of the AM model. However, in a few areas the discrepancy exceeds ±10%, as illustrated in Fig. 6. In none of these cases does *r*LOS show better conformity to the data compared to the AM model (Figs. 7a,b,e,f). It is therefore somewhat of a paradox that more data provide a tighter confidence

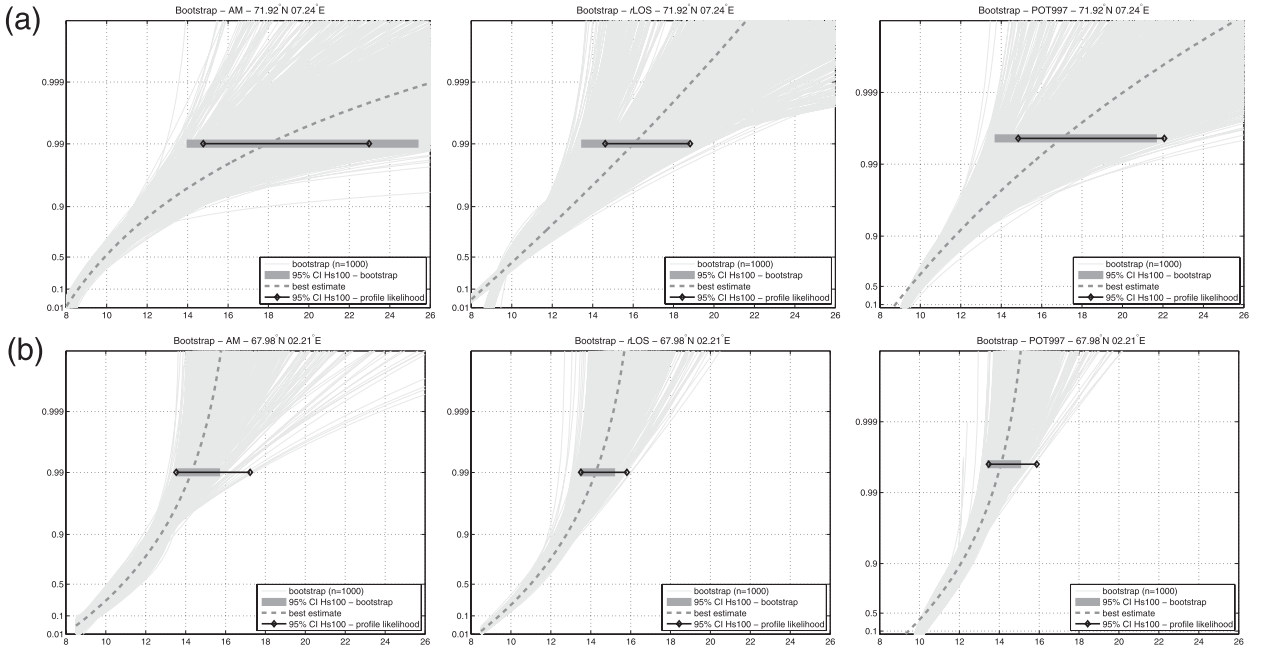

FIG. 8. Return value plots of H_s based on 1000 resamples of the data subsets obtained with (left to right) AM, *r*LOS, and POT997 at (a) 71.92°N, 07.24°E and (b) 67.98°N, 02.21°E, marked in light gray. The 95% confidence intervals based on the bootstrap and the profile likelihood are marked by a gray and a black bar, respectively. The best estimate of the corresponding model is represented by the dashed line.

interval, while the best estimate of the model shows an increased deviation from the actual data. This only shows that larger data subsets do not necessarily provide better results, a consequence of the well-known bias–variance trade-off. In general, increased data subsets will give more weight to the lower entries, and put restrains on the shape parameter of the distribution. For the *r*LOS model, the shape parameter shows significantly less variation compared to the AM model and is often closer to a Gumbel-type model(i.e., $\xi = 0$). In some cases this may prove more unfortunate than others, for example, when the data belong to more than one population.

The highest uncertainty related to the H_{s100} estimates of the *r*LOS model is found in the central parts of the North Sea (Fig. 4b). This also corresponds to the area of highest discrepancy relative to the AM model, where the latter is about 20% lower than the estimates of the *r*LOS model (Fig. 6a). This area is located on the continental shelf in water depths $h < 100$ m (Fig. 1). For wave lengths $> 4h$ (transitional depth), the bottom friction will have an increasing effect on the wave field (WMO 1998). It is therefore plausible that the H_s distribution may belong to two different populations above and below some threshold of wavelength. With the majority of the data originating from ''deep water'' conditions, where the bottom friction is negligible, a model extrapolation is likely to produce too conservative (i.e., too high) return value estimates at some point. In such cases it is very important that the highest entries are sufficiently weighted, which is more likely with the AM model (Fig. 7f). Some of the same features are found north of the Kola Peninsula, in the Barents Sea (Fig. 7b).

The POT model provides an alternative approach to the AM and *r*LOS models. With no consensus on how to set the threshold, we have explored several options in this study. In addition to the Anderson–Darling goodness-of-fit test, we have also tried to automate the threshold selection by using mean residual life plots (i.e., plotting the threshold against the mean excesses):

$$\left\langle \left\{ u, \frac{1}{n_u} \sum_{i=1}^{n_u} [x_{(i)} - u] \right\} : u < x_{(i)} \right\rangle. \qquad (10)$$

According to Coles (2001), the curve of the mean residual life plot should become linear above a threshold u if the Pareto distribution is a valid approximation. Such model diagnostics are normally made visually, but for a larger model domain this is too time consuming and needs to be automated. We have tried to do so by first fitting a polynomial function to the mean residual life plot. This expression is then doubly differentiated to find a local inflection point, which in most cases should indicate where the function is straightening out. This

method seems to locate a threshold where the mean residual life plot becomes linear, but the threshold is too low according to the Anderson–Darling goodness-of-fit test. So, initially we wanted to use a nonsubjective approach to set the threshold. In the end, we have found it necessary to check the model diagnostics at a selection of grid points, with our primary focus on those areas where the H_{s100} estimates of the three models differ the most. With the threshold set at the 99th percentile, the H_{s100} estimates are very much comparable to the *r*LOS model; at the 99.7th percentile the estimates are leaning more closely to the AM model. Based on the model diagnostics in Fig. 7, we prefer the threshold set at the 99.7th percentile, even though this has a negative effect on the confidence intervals, as the total area having a negative shape parameter decreases from 87.7% to 70.5%.

With the threshold set at the 99.7th percentile, the total number of entries per grid point lies somewhere in between the AM and *r*LOS models. This is also reflected in the relative width of the confidence interval, peaking around 45% of the best estimate (Fig. 5b). On the whole, the H_{s100} estimates are comparable to the AM model, though the POT997 is the only model providing a global maximum above 22 m. The relative difference between the two models exceeds 10% at two locations (Fig. 6). However, the model diagnostics at these positions (Figs. 7d,e) are inconclusive and do not indicate any model being superior to the other.

All of the models are somewhat sensitive to individual storms. An extreme example is found at 71.92°N, 7.24°E (Fig. 7a), where the H_{s100} estimate of the AM model is reduced from 18.0 to 16.2 m when censoring the highest entry. With the two highest entries removed, the H_{s100} estimate is only 14.6 m. This result is further emphasized by the bootstrap experiment at the same location (Fig. 8a), where all models indicate a large spread in the return value plots. For comparison, we have also chosen a position where the three models agree well on both the H_{s100} estimate and the shape parameter at 67.98°N, 02.21°E. This area is also associated with higher confidence in the estimates, defined by the relative width of the 95% confidence intervals. Even so, the distribution of the random draws varies significantly (Fig. 8b).

The three different approaches applied in this study are based on the assumption that each of the time series is IID. Independence is attained by exclusively extracting entries separated by a minimum of 48 h, while the criterion of stationarity is a different matter. The periodic cycle of the seasons is somewhat accounted for by only extracting the highest entries, providing a dataset clearly dominated by winter data. Trends, on the other hand, are unaccounted for. A number of studies have treated climatic trends of H_s within the northeast Atlantic (Wang and Swail 2001, 2002; Weisse and Günther 2007) and indirectly by studying changes in storminess (Alexandersson et al. 2000; Solomon et al. 2007; Wang et al. 2009a,b). Common to these studies is a worsening of the wave climate from the mid-1960s to the beginning of the 1990s. It should also be mentioned that the ERA-40 dataset, which provides the boundary conditions for the NORA10, is inhomogeneous itself due to a growing amount of observation assimilations over the reanalysis period (Uppala et al. 2005), though no severe inhomogeneities related to cyclone activity have been detected in the boreal extratropics of the ERA-40 data (Wang et al. 2006). Still, to what degree this influences the return value estimates of this study is left open. Future work may follow the example of Menéndez et al. (2009, 2008); Méndez et al. (2008).

This study is comparable to Caires and Sterl (2005) as it combines data from a hindcast–reanalysis and the POT model to obtain H_{s100} estimates. With the finer resolution of the NORA10, the corresponding return value estimates are believed to represent the 1-h mean sea state, while the H_{s100} estimates of Caires and Sterl (2005) are calibrated to buoy observations averaged over 3 h ($\pm$1). Even so, we find our H_{s100} estimates to be lower than the corresponding estimates of Caires and Sterl (2005). This is probably influenced by several factors. First and foremost, the datasets are different, as is clearly demonstrated by the superior performance of NORA10 relative to ERA-40 in Reistad et al. (2011). It was therefore found unnecessary to calibrate the initial dataset, a feature more likely to add bias in areas less well represented by observations. Second, our fitting procedure is not limited to the exponential distribution. As stated by Caires and Sterl (2005), the exponential distribution does not apply well in the storm tracks of the high latitudes where the shape parameter more often is negative (Table 3). The use of a purely unbounded distribution may explain why the estimates of Caires and Sterl (2005) are excessive in the northeast Atlantic. Third, we have set the threshold higher, at the 99.7th percentile versus the 93th percentile, putting more weight on the higher entries and improving the fit. Fourth, the two datasets cover different periods (1958–2002 versus 1958–2009). Fifth, the NORA10 does account for shallow water effects, unlike the ERA-40. Note, however, that the different period and the shallow water mode are expected to have minor impacts on the final results over most of the model domain. On the other hand, we believe the H_{s100} estimates of this study to be less precise in the westernmost part of the model domain (i.e., south-southwest of Iceland), as this area is more influenced by the boundary conditions of the model (Fig. 1). With the performance of the outer WAM

model and the ERA-40 highly comparable (Reistad et al. 2011), there are good reasons to believe that the H_{s100} estimates are too low in this particular area. This may explain why the global maximum is shifted further east-northeast compared to the similar estimates of Alves and Young (2003), Caires and Sterl (2005), and Wimmer et al. (2006).

The general wave climate, represented by the mean and the 95th–99th percentiles of H_s, is decreasing going northeast into the Norwegian Sea and the Barents Sea. It is therefore somewhat unexpected that we find a local minimum in the H_{s100} estimates in the central parts of the Norwegian Sea (Figs. 3a, 4a, and 5a). Bearing in mind the model sensitivity discussed above, this result would have appeared more nuanced exclusive of the two extraordinary wave events above 16.5 m illustrated in Fig. 7a. However, the bootstrap experiment presented in Fig. 8 offers a fairly strong indication that the two distributions, located at 71.92°N, 07.24°E and 67.98°N, 02.21°E, possess shape parameters of opposite signs, indicating that the northernmost position will see the highest wave conditions in time. This feature is probably a combined effect of preferred low pressure tracks and some shadowing effects from Iceland and the Faroe Islands.

6. Conclusions

In this study we have presented 100-yr return value estimates of significant wave height, H_{s100}, covering the northeast Atlantic. The estimates are based on a new hindcast developed by the Norwegian Meteorological Institute spanning the period 1958–2009. With three different subsets of the initial data, the annual maximum, the r largest-order statistic, and the peaks over threshold, we have utilized three commonly applied extreme value models, all based on the assumption of stationary wave conditions. Our choice of r is determined by the likelihood ratio test and varies over the model domain, while the threshold selection is based on the Anderson–Darling test and set at the 99.7th percentile of the initial data.

The levels of model performance have been investigated by return value plots and quantile plots, and primarily focused toward areas with the highest discrepancies. However, such model diagnostics are subjective and leave it up to the individual researcher to determine the best model. In general, the H_{s100} estimates differ by less than $\pm 5\%$, with local discrepancies peaking around 20%. In these areas we have found the annual maximum and the peaks-over-threshold methods to outperform the r largest order statistic, as the former two conform better to the highest entries.

The main advantage of utilizing larger data subsets is tighter confidence intervals. With the annual maximum, the width of the 95% confidence interval constitutes as much as 60% of the H_{s100} estimate, while the r largest-order statistics peak just above 30%. However, this seemingly increased confidence in the estimate is not necessarily indicative of an improved fit between the data and the model, which is a paradox.

In the end, no model has been found to be superior in all cases. The model that utilizes the most data should be preferred provided that the conformity between the model and the data are intact, as this increases our confidence in the model estimate. However, bigger data subsets combined with the maximum likelihood approach will put more weight on the lower part of the distribution and are therefore more likely to be biased at the high end. This feature has been particularly evident for the r largest-order statistic in this study. With the considerable time span of the hindcast (52 yr), the annual maximum has proved to perform well with little bias and acceptable variance, but low confidence. Overall, the peaks-over-threshold model has shown the best results provided the threshold is set high, showing good fit and reasonable confidence intervals.

Acknowledgments. This study is part of a Ph.D. program financed by the Norwegian Centre for Offshore Wind Energy (NORCOWE). We would also like to express our gratitude to the EU project ExtremeSeas and the Wave Ensemble Prediction Offshore (WEPO) project for partial funding and the reviewers for constructive feedback. Finally, thanks to the Norwegian Deepwater Programme for financing the NORA10 hindcast.

REFERENCES

Alexandersson, H., H. Tuomenvirta, T. Schmith, and K. Iden, 2000: Trends of storms in NW Europe derived from an updated pressure data set. *Climate Res.,* **14,** 71–73.

Alves, J. H. G. M., and I. R. Young, 2003: On estimating extreme wave heights using Geosat, Topex/Poseideon and ERS-1 altimeter data. *Appl. Ocean Res.,* **25,** 167–186.

Anderson, C. W., D. J. T. Carter, and P. D. Cotton, 2001: Wave climate variability and impact on offshore design extremes. Shell International Tech. Rep., 88 pp. [Available online at http://info.ogp.org.uk/metocean/JIPweek/att/WCEReport_2sided.pdf.]

Caires, S., and A. Sterl, 2005: 100-year return value estimates for ocean wind speed and significant wave height from the ERA-40 data. *J. Climate,* **18,** 1032–1048.

Carter, D. J. T., 1993: Estimating extreme wave heights in the NE Atlantic from GEOSAT data. Health and Safety Executive Offshore Tech. Rep. OTH 93 396, 28 pp.

Choulakian, V., and M. Stephens, 2001: Goodness-of-fit test for the generalized Pareto distribution. *Technometrics,* **43,** 478–485.

Coles, S., 2001: *An Introduction to Statistical Modelling of Extreme Values*. Springer-Verlag, 208 pp.

——, and E. Simiu, 2003: Estimating uncertainty in the extreme value analysis of data generated by a hurricane simulation.

J. Eng. Mech., **129,** 1288–1294, doi:10.1061/(ASCE)0733-9399(2003)129:11(1288).

Cooper, C. K., and G. Z. Forristall, 1997: The use of satellite altimeter data to estimate extreme wave climate. *J. Atmos. Oceanic Technol.,* **14,** 254–266.

Komen, G. J., M. Cavaleri, M. Donelan, K. Hasselmann, S. Hasselmann, and P. A. E. M. Janssen, 1994: *Dynamics and Modelling of Ocean Waves*. Cambridge University Press, 532 pp.

Lopatoukhin, L. J., V. A. Rozhkov, V. E. Ryabinin, V. R. Swail, A. V. Boukhanovsky, and A. B. Degtyarev, 2000: Estimation of extreme wind wave heights. WMO/TD-No. 1041, JCOMM Tech. Rep. 9, Joint WMO–IOC Technical Commission for Oceanography and Marine Meteorology, 73 pp.

Méndez, F. J., M. Menéndez, A. Luceño, R. Medina, and N. E. Graham, 2008: Seasonality and duration in extreme value distributions of significant wave height. *Ocean Eng.,* **35,** 131–138.

Menéndez, M., F. J. Méndez, I. J. Losada, and N. E. Graham, 2008: Variability of extreme wave heights in the northeast Pacific Ocean based on buoy measurements. *Geophys. Res. Lett.,* **35,** 1–6.

——, ——, C. Izaguirre, A. Luceño, and I. J. Losada, 2009: The influence of seasonality on estimating return values of significant wave height. *Coastal Eng.,* **56,** 211–219.

Panchang, V., L. Zhao, and L. Z. Demirbilek, 1999: Estimation of extreme wave heights using Geosat measurements. *Ocean Eng.,* **26,** 205–225.

Peters, D. J., C. J. Shaw, C. K. Grant, J. C. Heldeman, and D. Szabo, 1993: Modelling the North Sea through the European Storm Study (NESS). Offshore Technology Conf. Rep. OTC 7130, 479–493.

Reistad, M., O. Breivik, and H. Haakenstad, 2007: A high-resolution hindcast study for the North Sea, the Norwegian Sea and the Barents Sea. *Proc. 10th Int. Workshop on Wave Hindcast and Forcasting and Coastal Hazard Symp.,* North Shore, HI, WMO.

——, ——, ——, O. J. Aarnes, and B. R. Furevik, 2011: A high-resolution hindcast of wind and waves for the North Sea, the Norwegian Sea and the Barents Sea. *J. Geophys. Res.,* **116,** C05019, doi:10.1029/2010JC006402.

Semedo, A., K. Sušelj, A. Rutgersson, and A. Sterl, 2011: A global view on the wind sea and swell climate and variability from ERA-40. *J. Climate,* **24,** 1461–1479.

Soares, C. G., and M. G. Scotto, 2004: Application of the *r* largest-order statistics for long-term predictions of significant wave height. *Coastal Eng.,* **51,** 287–294.

Solomon, S., D. Qin, M. Manning, M. Marquis, K. Averyt, M. M. B. Tignor, H. L. Miller Jr., and Z. Chen, Eds., 2007: *Climate Change 2007: The Physical Science Basis*. Cambridge University Press, 996 pp.

Soukissian, T. H., and G. D. Kalantz, 2006: Extreme value analysis methods used for extreme wave prediction. *Proc. 16th Int. Offshore and Polar Engineering Conf.,* San Francisco, CA, Int. Society of Offshore and Polar Engineers. [Available online at http://www.isope.org/publications/proceedings/ISOPE/ISOPE%202006/papers/2006_TS_01.pdf.]

Undén, P., and Coauthors, 2002: HIRLAM-5 scientific documentation. HIRLAM-5 Project, SMHI Tech. Rep. S-601 76, 144 pp.

Uppala, S. M., and Coauthors, 2005: The ERA-40 Re-Analysis. *Quart. J. Roy. Meteor. Soc.,* **131,** 2961–3012.

Wang, X. L., and V. R. Swail, 2001: Changes of extreme wave heights in the Northern Hemisphere oceans and related atmospheric circulation regimes. *J. Climate,* **14,** 2204–2221.

——, and ——, 2002: Trends of Atlantic wave extremes as simulated in a 40-yr wave hindcast using kinematically reanalyzed wind fields. *J. Climate,* **15,** 1020–1034.

——, ——, and F. W. Zwiers, 2006: Climatology and changes of extratropical cyclone activity: Comparison of ERA-40 with NCEP–NCAR reanalysis for 1958–2001. *J. Climate,* **19,** 3145–3166.

——, ——, W. Z. Francis, X. Zhang, and Y. Feng, 2009a: Detection of external influence on trends of atmospheric storminess and northern oceans wave heights. *Climate Dyn.,* **32,** 189–203.

——, F. W. Zwiers, V. R. Swail, and Y. Feng, 2009b: Trends and variability of storminess in the northeast Atlantic region, 1874–2007. *Climate Dyn.,* **33,** 1179–1195.

Weisse, R., and H. Günther, 2007: Wave climate and long-term changes for the southern North Sea obtained from a high resolution hindcast 1958–2002. *Ocean Dyn.,* **57,** 161–172.

Williams, O. M., 2005: Wave mapping in UK waters. Health and Safety Executive Res. Rep. 392, 18 pp. [Available online at http://www.hse.gov.uk/research/rrhtm/rr392.htm.]

Wimmer, W., P. Challenor, and C. Retzler, 2006: Extreme wave heights in the North Atlantic from altimeter data. *Renew. Energy,* **31,** 241–248.

WMO, 1998: *Guide to Wave Analysis and Forecasting*. 2nd ed. World Meteorological Organization, 159 pp.