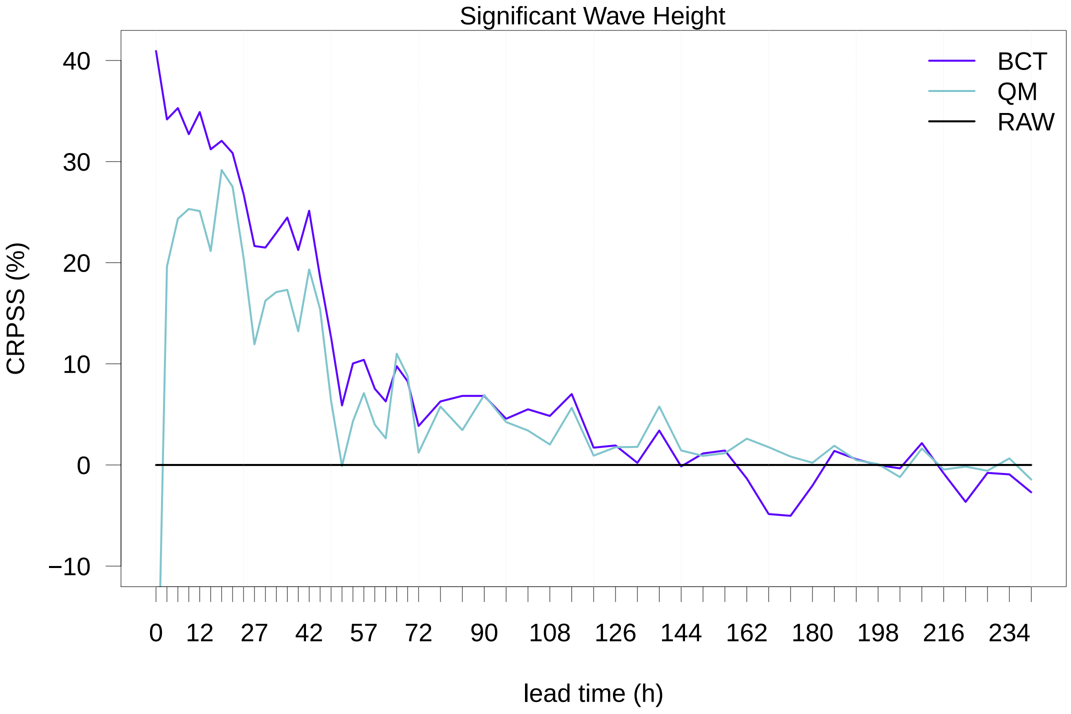
METreport

No. 1/2016
ISSN 2387-4201
Oceanography

Statistical calibration of wave ensembles

A local Box-Cox t-distribution approach

John Bjørnar Bremnes, Magnar Reistad and Birgitte Rugaard Furevik



METreport

Title Statistical calibration of wave ensembles. A local Box-Cox t-distribution approach	**Date** February 10, 2016
Divisions Numerical Weather Prediction and Oceanography and Marine Meteorology	**Report no.** 1/2016
Author(s) John Bjørnar Bremnes, Magnar Reistad and Birgitte Rugaard Furevik	**Classification** ● Free ○ Restricted
Client(s) The Research Council of Norway	**Client's reference** 225231/070

Abstract

Ensembles of wave forecasts are often biased and underdispersive when validated against site measurements. The main objective of this work is to develop a statistical method for calibration of ensemble forecasts in order to provide reliable input to simulation tools for marine operations. As ensemble statistics indicate that the shape of the forecast distribution should vary considerably in time and with the weather situation the use of the flexible four-parameter Box-Cox t-distribution (BCT) is proposed. The BCT regression method is applied to data at the FINO3 platform and compared to both the raw ensemble forecasts and a simple quantile mapping technique applied to each ensemble member. In terms of general validation scores for probabilistic forecasts the BCT method gave reliable forecasts that are up to 40% better (CRPS) than the raw ensemble for significant wave height and up to 60% better for wave period. The improvement due to the calibrated ensemble data set is illustrated by comparing forecasted weather windows over the test period from the deterministic dataset using α-factor, in the raw ensemble and in the calibrated ensemble.

Keywords

wave forecasting, ensemble, statistical calibration

Disciplinary signature

Responsible signature

Abstract

Ensembles of wave forecasts are often biased and underdispersive when validated against site measurements. The main objective of this work is to develop a statistical method for calibration of ensemble forecasts in order to provide reliable input to simulation tools for marine operations. As ensemble statistics indicate that the shape of the forecast distribution should vary considerably in time and with the weather situation the use of the flexible four-parameter Box-Cox t-distribution (BCT) is proposed. The BCT regression method is applied to data at the FINO3 platform and compared to both the raw ensemble forecasts and a simple quantile mapping technique applied to each ensemble member. In terms of general validation scores for probabilistic forecasts the BCT method gave reliable forecasts that are up to 40% better (CRPS) than the raw ensemble for significant wave height and up to 60% better for wave period. The improvement due to the calibrated ensemble data set is illustrated by comparing forecasted weather windows over the test period from the deterministic dataset using α-factor, the raw ensemble and the calibrated ensemble.

Contents

1 Introduction 5

2 Data 7

3 Method 7

3.1 Statistical method 7

3.2 Forecast validation measures 9

3.3 Forecasting weather windows 11

4 Results 11

5 Discussion and Conclusions 15

1 Introduction

Utilising the information about uncertainty in the weather forecasts provided by ensemble predictions is very important in many applications, in particular in offshore wind energy projects where cost of installation and operation constitutes a large part of the overall cost of the project. Ensemble prediction is a group of model runs using a forecast model, run from nearly the same initial condition, only perturbed slightly for each member in the ensemble. The control run is unperturbed and compares to a deterministic forecast. Usually, accurate deterministic models with data assimilation are applied for the first 2-3 days ahead, while coarser (often global) ensemble systems are optimised for longer forecasts. Since the atmosphere is a chaotic system, small differences in the initial conditions will lead to very different realisations by a numerical forecast model over time. The European Center for Medium-Range Weather Forecasts (ECMWF) runs a global atmosphere model ensemble with 51 members at $32km$ spatial resolution and a wave model with $55km$ spatial resolution. This ensemble prediction system (ENS) provides forecasts 15 days ahead. An example of an ENS wave forecast for a location in the North Sea is shown in Figure 1. In the example there is large spread, i.e. uncertainty, related to the increase in wave height on days one and three. In between, on days two and four, the forecast is more reliable (less spread). From day 7 and onwards, there is very little information to gain from this forecast as the spread is even and the median rather constant. Experience shows that the

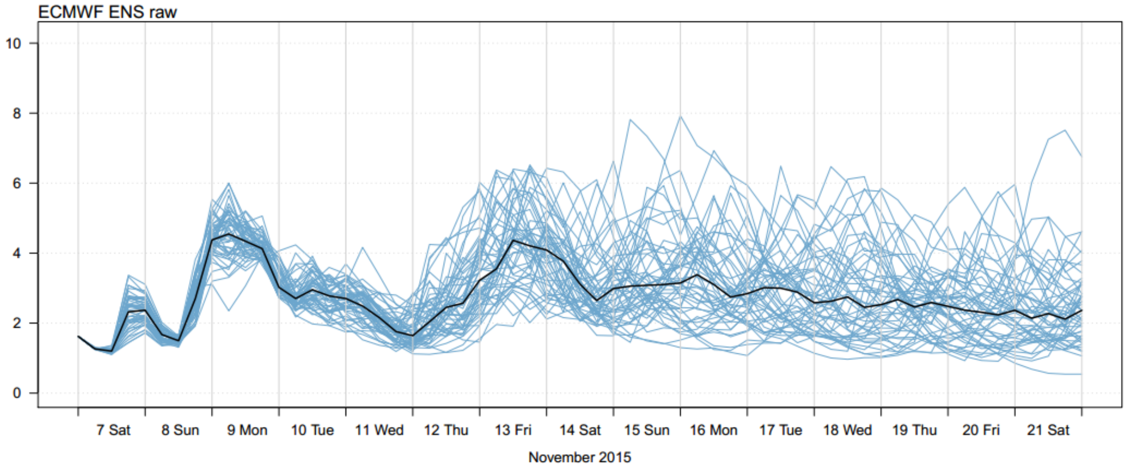


Figure 1: *Example of ENS wave forecast. Blue curves are the single ensemble members. The black curve is the ensemble mean.*

forecasts from the ENS are biased too low in the North Sea and the spread of the members may be too weak (the ensemble is underdispersed). In this work the ENS-forecasts for significant wave height (Hs) at FINO-3 are calibrated against wave observations from the Acoustic Wave and Current Profiler (AWAC) at FINO-3 during May 2013 - April 2014 (Figure 2). The resulting forecasts are then validated during a test period 30 April - 31 July 2014 at FINO-3 and the predicted weather windows are compared to the α-factor method (2).

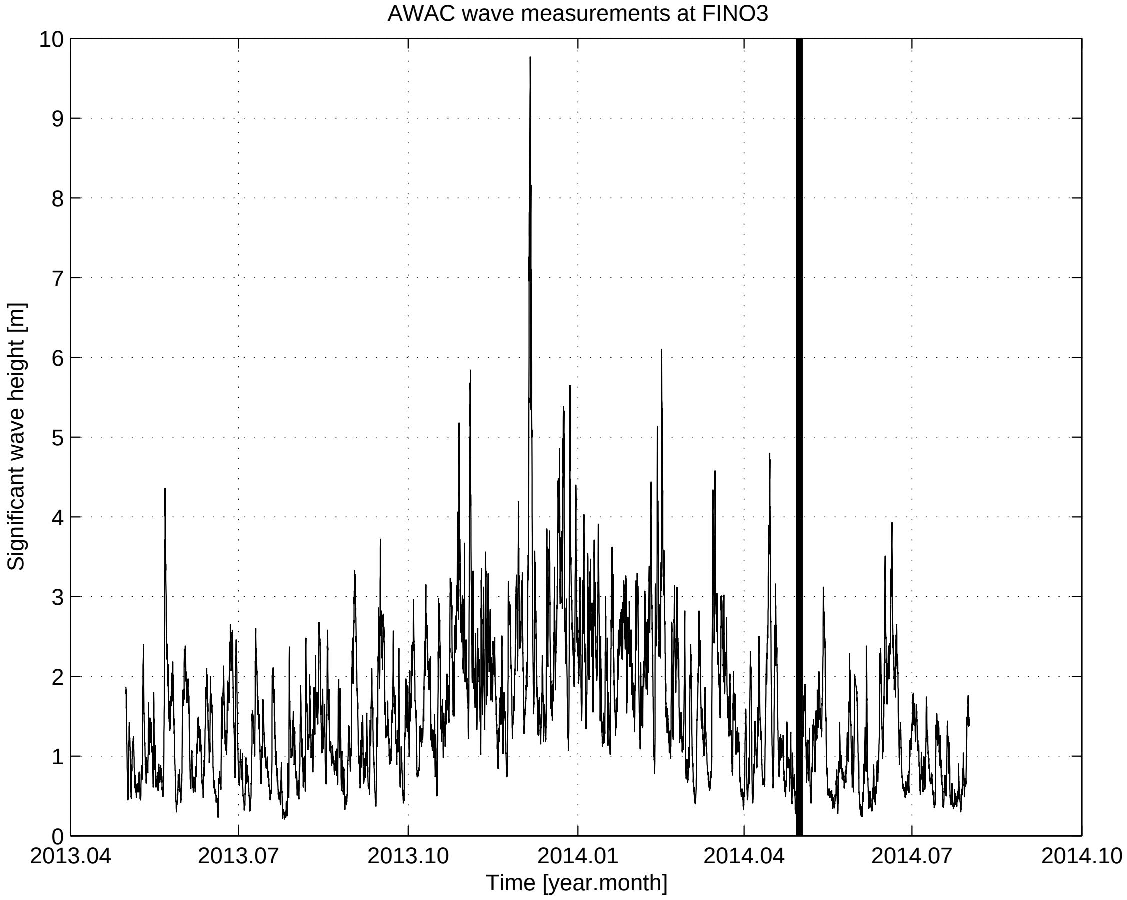


Figure 2: *Measurements of significant wave height at FINO3 1 May 2013 - 31 July 2014. Data before 1 May 2014 (vertical bar) are used for calibration of the ensemble. The data after 1 May 2014 is the test period used for validation and demonstration.*

2 Data

Data from the AWAC close to the FINO3 platform (55.195° N, 7.158° E) are organized for the period 30 April 2013 to 31 July 2014. Measurements of significant wave height (Hs) and upcrossing wave period (Tz) are interpolated to a temporal resolution of one hour. The ensemble forecast data (51 members) are extracted from the ECMWF archive on a latitude-longitude grid with a spatial resolution of 0.25° and then bilinearly interpolated to the position of the AWAC. The variables significant wave height (GRIB code 229 in table 140) and mean wave period based on second moment (GRIB code 221) are taken from the 00 UTC run with lead times +0, +3, +6, ..., +72, +78, +84, ..., +240h.

3 Method

3.1 Statistical method

In a statistical forecasting framework the objective is to model how the probability distribution of the wave variable of interest depends on information available at the time the forecast is to be issued. The predictive information in this study comes from an ensemble of forecasts from a wave model which has been compressed into two variables – the ensemble mean and the ensemble standard deviation. Thus, the aim is to determine the conditional distribution of the wave variable given the ensemble mean and standard deviation. In many modeling situations a specific parametric probability distribution is chosen a priori, as is done here. The problem is then reduced to model how its parameters depend on the predictive information. The choice of probability distribution is not obvious, but by studying how various ensemble statistics vary in time one can get an idea despite sampling uncertainty. In Figure 3 the time variation of the ensemble mean, standard deviation, skewness, and kurtosis of significant wave height is shown. Clearly, each of them vary considerably in time. To allow for such variations a flexible parametric probability distribution is therefore needed.

In this study, the Box-Cox t-distribution (BCT) is chosen. The BCT distribution is characterized by four parameters μ, σ, ν, and τ which can be interpreted to be related to the median, the coefficient of variation, the skewness and the kurtosis, respectively. Brief investigations revealed that a reasonable model where all parameters are related to ensemble statistics is not straightforward. Instead a local statistical modeling approach is adopted. For every forecast to be made the training data cases are each assigned a weight reflecting how similar it is to the current forecasting situation. Then a simple statistical

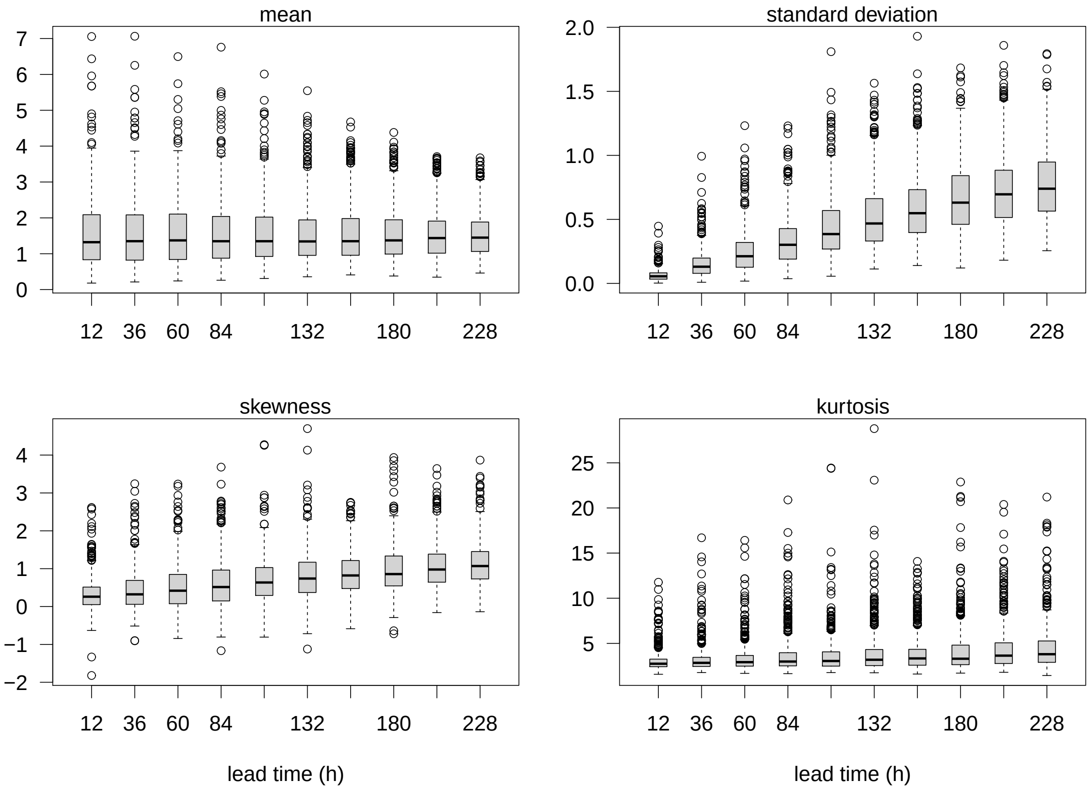


Figure 3: *Distributions of the ensemble statistics mean, standard deviation, skewness, and kurtosis for significant wave height as a function of lead time. The distributions are calculated over the whole data period 30 April 2013 - 31 July 2014. The box shows the median, and the 25 and 75 percentiles, and the whiskers are the 5 and 95 percentiles. The circles are the extremes.*

model with the BCT distribution is fitted to the weighted data. The estimated distribution at the new predictor is used as the probabilistic forecast and quantiles and probabilities of events can be derived from this.

In the first step, each case in the training data is given a weight. Assume the predictors (vectors) in the training set are denoted by $x_1, x_2, ..., x_n$ where n is the number of cases, and let x be the predictor value of the forecast to be made. The weights are computed using the so-called tri-cube weight function which is defined as

$$w(x_i, x) = w_0\left(\frac{\|\theta \odot (x_i - x)\|_2}{D(\lambda n)}\right) \quad (1)$$

for any $i \in \{1, 2, ..., n\}$, where $\odot$ denotes elementwise vector multiplication and $\|\cdot\|_2$ is the Euclidean norm. The vector θ with elements in the interval $(0,1)$ defines the relative impact of each predictor variable in the weight process and its element is assumed to add up to one. $D(\lambda n)$ is the Euclidean distance to the λn nearest predictor value where λ controls the proportion of training cases with positive weight. The function w_0 is defined as

$$w_0(u) = \begin{cases} (1-u^{3/2})^3 & \text{if } u \in [0,1] \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Ideally the parameters λ and θ should be tuned for each lead time, but based on previous experience λ is set to 0.2, that is only 20% of the training cases have impact on a single forecast. The relative impact of the predictor variables ensemble mean and standard deviation is set to 0.9 and 0.1, respectively. Prior to the weight process the variables are standardized by dividing by their standard deviations (over the training time period).

In the BCT model only the location parameter μ is assumed to be dependent on the raw ensemble, more precisely $\mu = \mu_0 + \mu_1 MEDIAN$, where $MEDIAN$ denotes the ensemble median. The remaining parameters are assumed to be constant, but unknown. Thus, in total five parameters had to be estimated for each forecast. The model parameters are fitted by maximum likelihood estimation using the Rigby and Stasinopoulos algorithm in the GAMLSS R-package (1).

Calibration using the BCT distribution will result in well calibrated forecasts if the modeling is done properly. It is, however, a more complex approach than for example simple bias adjustment of each ensemble member using quantile mapping (QM). In order to see the difference in skill a QM method is also applied. The quantile mapping is defined by separately sorting the ensemble control forecasts and the measurements for the training period, and then using linear interpolation to get a calibration function that can be applied to each ensemble member forecast. The outcome is an ensemble of equal size as the original/raw ensemble. It should be noted, though, that QM does not necessarily imply well calibrated ensembles.

3.2 Forecast validation measures

For probabilistic and ensemble forecasts the most common summarizing statistics for forecast quality is the continuous ranked probability score (CRPS) which is based on the difference between the forecasted cumulative distribution function and the observation.

Here, the CRPS for a single ensemble forecast is computed by

$$\frac{1}{m}\sum_{i=1}^{m}|f_i - y| \;-\; \frac{1}{2m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}|f_i - f_j| \tag{3}$$

where $f_1, f_2, ..., f_m$ denotes the m members of a single ensemble forecast and y the corresponding observation. The lower the CRPS is the better. For simplicity, the same formula is also used for the calibrated BCT forecasts which are converted to ensembles by calculating the $1/(m+1), 2/(m+1), ..., m/(m+1)$ quantiles from the BCT distributions. The CRPS for a set of forecasts is just the average CRPS over the forecasts. Further, for comparing several forecast systems it is often easier to interpret CRPS relative to some reference. The continuous ranked probability skill score (CRPSS) is defined as 1 - CRPS/CRPS$_{ref}$ where CRPS$_{ref}$ is the CRPS of a reference forecast system; in our results the raw ensemble forecast system are used as the reference. Thus, a positive CRPSS indicates improvement over the reference forecast system.

CRPS summarizes the quality of the full probability distribution or ensemble forecasts, but the quality of forecasts for specific events, e.g. significant wave height less than or higher than a threshold, is also of interest. A suitable metric for probabilistic forecasts of such events is the Brier score (BS) which is defined as

$$\frac{1}{T}\sum_{t=1}^{T}(p_t - o_t)^2 \tag{4}$$

where p_t is the probability for the event at time t, o_t is binary variable indicating whether the event occured, and T the number of forecasts. For the BCT calibrated forecasts the probability of the event is calculated directly from the BCT distribution, while for the raw and the QM calibrated ensemble it is the fraction of members for the event. The Brier skill score (BSS) can be computed the same way as CRPSS.

Summarizing measures like CRPS and BS does not give the full picture of forecast quality. For example, good probabilistic forecasts should be reliable, but at the same time also sharp (low uncertainty). It is therefore instructive to assess both properties. In order to measure reliability, the $m = 51$ ensemble members and the observation are concatenated and the rank of the observation is computed. In a reliable forecast system the ranks of the observations over time should be uniformly distributed. The degree of reliability can be quantified using the reliability index defined as

$$\frac{1}{m+1}\sum_{i=1}^{m+1}(c_i \;-\; \frac{1}{m+1})^2 \tag{5}$$

where c_i is the fraction of cases where the observation rank is i. In this study the validation data set only consisted of 92 cases which due to the relative large size of the ensemble would imply high sampling uncertainty in the score. Prior to computing the score the observation ranks are therefore grouped into 13 bins instead of the original 52 bins. The second property, sharpness, is evaluated by computing the average widths of the central 90% and 50% forecast intervals.

The Brier score is a measure for probabilistic forecasts of binary events. In the context of weather window forecasting probabilistic forecasts are often reduced to binary ones in decision making process. To assess such forecasts the hit and false alarm rate are applied. The hit rate is defined as the fraction of events which are correctly forecast, while the false alarm rate is the fraction of non-events (no weather window) which are forecast as weather windows.

3.3 Forecasting weather windows

Forecasting for an offshore operation requires a weather window of a certain length, which satisfies a criterion with high probability. This is usually handled by defining an operational criteria based on the design criteria times a factor called the α-factor (2). The forecasted value from a deterministic forecast should then be below the operational criteria during the time of operation. A high value for the α-factor means high confidence in the forecast. The α-factor takes a range of values dependent on the weather information available. The cases considered here is the base case (Level C) and the situation when a meteorologist is present at the site (Level A); table 4-1 and table 4-3 in (2). The α-factor approach is applied on deterministic forecasts. Deterministic forecasts give the most likely evolution of the weather as estimated by the forecast model, but does not include information of the predictability of the weather system, although this can vary quite dramatically. Such information is however included in ensemble prediction systems and this information may be utilised if the ensemble has realistic spread and no bias.

4 Results

In order to test the calibration methods the data are divided into a training and a test data set. Data before May 2014 is applied for model estimation/training, while predictions are made for the months May, June, and July 2014. All validation results to follow are for the latter period.

In Figure 4 summarizing measures for the quality of the significant wave height (Hs) and wave period (Tz) forecasts are presented. In terms of the CRPSS the improvements are clearly significant for both variables; up to about 40% and 60% for the shortest lead times, respectively. For significant wave height the improvement is most pronounced for the first three days, mainly because of the strong underestimation of forecasts uncertainty in the raw ensemble forecast. The gain in mean error and mean absolute error (ME/MAE) for Hs are more modest (at least relatively) and the ME possibly have less impact on the CRPSS. For the wave period, on the other hand, the positive CRPSS can likely be attributed to the large negative ME for the raw ensemble. The performance of the models is also assessed for 1.0 and 1.5 meter significant wave height threshold events. The BSS showed more or less the same pattern as for CRPSS, though with stronger periodic variations. As expected, the fully calibrated ensemble (BCT) gave better scores than the ensemble data set which is just bias corrected (QM) most likely due to the better theoretical foundation of the BCT.

The reliability and sharpness of the forecasts are shown in Figure 5. Clearly, the raw ensemble is inferior with respect to reliability up to day five for Hs and for all lead times for Tz. The latter can be linked to the strong negative ME for all lead times. Thus, probabilities derived directly from the raw ensemble could not be trusted. The poor reliability of QM at lead time $+0h$ is due to no variation/uncertainty at this lead time which is inherited from the raw ensemble. Apart from that, the reliability of the calibrated methods seemed adequate.

The performance of the forecasts are also assessed in a weather window decision context and compared to the α-factor approach. The α-factors are taken from (2, Table 4-1) (base case) and applied to the raw control forecast (ensemble member no. 0). Two definitions of weather windows are considered – significant wave height below 1.0 and $1.5m$ (α-factor of 0.65 and 0.705, respectively). Since this type of weather window forecasting is a deterministic approach the ensemble and probabilistic forecasts had to be converted to 'yes' or 'no' forecasts. Three probability thresholds are used: 1, 0.1, and 0.01%. That is, if the probability of exceeding the wave threshold are less than the probability threshold the forecast is converted to 'yes' (weather window) and 'no' otherwise. Note that for the raw and QM calibrated ensembles (of size 51) the conversion is in practice independent on the probability threshold. For these the ensembles are set to 'yes' when all ensemble members are below the wave threshold.

The outcome of the inter-comparison is summarized in Figure 6 showing the hit and false alarm rates for the various forecasting approaches and weather window definitions. For the first two days the α-factor approach gives the best forecasts. The raw ENS have higher hit rates, but does not meet the certainty requirement as the false alarm rates are too high. Both calibration methods provided quite conservative forecasts of weather windows. For forecasts beyond two days the false alarm rates of the α-factor approach are too high. This is expected since the α-factors are not designed for these forecast horizons.

Some of these results are also illustrated by counting the number of forecasted weather windows in the different data sets over the test period. The test period provides a maximum number of 104 possible weather windows if we look for a weather windows starting at the beginning of each forecast. (Due to the length of the forecasts, the files actually covers 16 April - 28 July 2014.) From the ensemble, the deterministic forecast can be represented by either the control run, the median or the mean of the members. We here consider the design Hs of 2.0 and $1.5m$. The α-factor related to a design wave height of $2.0m$ $(1.5m)$ and a weather window length of 72 (24) hours is given by approximately $\alpha = 0.78 - 0.15\frac{leadtime}{72}$ $(\alpha = 0.72 - 0.13\frac{leadtime}{72})$ for level C, which is recommended when only a standard offshore forecast is available. If a meteorologist is present, higher values $\alpha = 0.86 - 0.17\frac{leadtime}{72}$ $(\alpha = 0.80 - 0.15\frac{leadtime}{72})$ is applied to the deterministic forecast (2). Alternatively, if the ensemble forecast is able to represent the uncertainty in the weather forecast, the α-factor could be avoided. In that case one of the ensemble members could be used directly to forecast weather windows. Since there are 51 ensemble members, each member represents a probability of almost 2%. Tables 1 - 4 compares the number of weather windows ('yes') from using members (ENS50, ENS49 etc.) of the calibrated ensemble (BCT), the raw ENS and deterministic forecast (applying the α-factor on the control run, the median and the mean of the ensembles) under the condition of observed weather window ('yes'/'no' in far left column). The ensemble member number 50 (ENS50) is here defined as the member predicting the highest wave height at any given time. Likewise for the other members (ENS49-ENS45).

The effect of the calibration is evident. The BCT forecasts have fewer weather windows, but no false ones neither for the α-factor method nor the first two ensemble members in any of the scenarios. The raw ensemble gives two false weather windows using the high α-factor in table 2 and one can clearly not rely on the highest values of the members as there are three false weather windows predicted using ENS50 in table 1 and four in table 3.

There is not much difference in using the control run, the mean or the median of

Table 1: *Number of weather windows of 72 hours and design criteria Hs$<$ 2m. The α-factor* 0.78 *according to Level C, standard wave forecast is applied to the control run (Ctrl), the median (Med.) and the mean of the ensemble. Top part: BCT calibrated forecast; Bottom part: Raw ensemble.*

Obs	Forecast	Ctrl	Med.	Mean	ENS50	ENS49	ENS48	ENS47	ENS46	ENS45
Yes	Yes	41	42	42	32	48	51	57	58	59
Yes	No	27	26	26	36	20	17	11	10	9
No	Yes	0	0	0	0	0	2	2	2	3
No	No	36	36	36	36	36	34	34	34	33
Yes	Yes	44	45	46	53	60	63	63	64	66
Yes	No	24	23	22	15	8	5	5	4	2
No	Yes	0	0	0	3	6	6	8	8	8
No	No	36	36	36	33	30	30	28	28	28

BCT in the α-factor method. The method is rather conservative for standard weather forecast with around 41-42 predicted 72-hour 2.0*m* weather windows of a total of 68 and around 39-40 predicted 24-hours 1.5*m* weather windows of a total of 67 observed. With a meteorologist on site, the α-factor method predicts 46-48 and 44-46 of the weather windows, respectively.

The ENS50 of the calibrated ensemble (BCT) is too conservative in predicting the 72-hours 2.0*m* weather window (table 1), where only 32 weather windows are forecasted. Using ENS49 would result in 48 correctly forecasted weather windows in Table 1 and 51 24-hours 1.5*m* weather windows in table 3. This is less conservative than with the α-factor method, while still avoiding false predictions. A challenge is to estimate the certainty of this result, for which a longer data set is needed.

Table 2: *As for table 1 but with α-factor 0.86 according to Level A, meteorologist on site.*

Obs	Forecast	Ctrl	Med.	Mean
Yes	Yes	46	48	48
Yes	No	22	20	20
No	Yes	0	0	0
No	No	36	36	36
Yes	Yes	55	58	58
Yes	No	13	10	10
No	Yes	2	2	2
No	No	34	34	34

Table 3: *Number of weather windows of 24 hours and design criteria Hs*< 1.5*m. The* α*-factor 0.72 according to Level C, standard wave forecast is applied to the control run (Ctrl), the median (Med.) and the mean of the ensemble. Top part: BCT calibrated forecast; Bottom part: Raw ensemble.*

Obs	Forecast	Ctrl	Med.	Mean	ENS50	ENS49	ENS48	ENS47	ENS46	ENS45
Yes	Yes	39	40	40	49	51	53	54	54	54
Yes	No	28	27	27	18	16	14	13	13	13
No	Yes	0	0	0	0	0	0	0	0	0
No	No	39	39	39	39	39	39	39	39	39
Yes	Yes	44	44	44	60	63	64	64	64	64
Yes	No	23	23	23	7	4	3	3	3	3
No	Yes	0	0	0	4	4	4	4	6	7
No	No	39	39	39	35	35	35	35	33	32

Table 4: *As for table 3 but with* α*-factor 0.80 according to Level A, meteorologist on site.*

Obs	Forecast	Ctrl	Med.	Mean
Yes	Yes	44	45	46
Yes	No	23	22	21
No	Yes	0	0	0
No	No	39	39	39
Yes	Yes	53	53	53
Yes	No	14	14	14
No	Yes	0	0	0
No	No	39	39	39

5 Discussion and Conclusions

The main objective of the study was to provide well calibrated ensemble forecasts for input to a simulation tool for marine operations. In that context it has been demonstrated that the quality of ensemble forecasts can be considerably improved by means of statistical calibration; and maybe most important of all that properly calibrated probabilistic forecasts are reliable which is a prerequisite for any decision making based on probabilistic forecasts. There are, however, scope for improvements. For example, the dynamical information about forecast uncertainty inherent in raw ensemble forecasts has not been fully utilized. Partly this is because forecast quality measures like CRPS only weakly penalizes uncertainty/spread. The CRPS is in practice strongly related to the bias of the location parameter (mean/median) of the probability distribution. Further work in this direction is recommended.

A more direct application of forecasts is in deterministic forecasting of weather windows with high degree of certainty – typically in the order of 99.99%. The calibration methods in this study was not designed with this in mind and the results in terms of weather window hit rates at 0.01 and 0.1% (Figure 6) were inferior due to the heavy upper tails of the BCT distribution. The BCT distribution is very flexible and if any outliers due to measurement errors were present in the data then these will have significant negative impact in the statistical calibration. By higher focus on the extreme quantiles it should be possible to forecast weather windows with hit rates at least as high as for the α-factor approach. The main reason is simply that ensembles include more information about wave conditions than a single deterministic forecast do. Further, it would be interesting to compare statistically calibrated forecast systems based on ensemble versus dedicated (regional, fine scale) determinstic forecasts to quantify the potential gain.

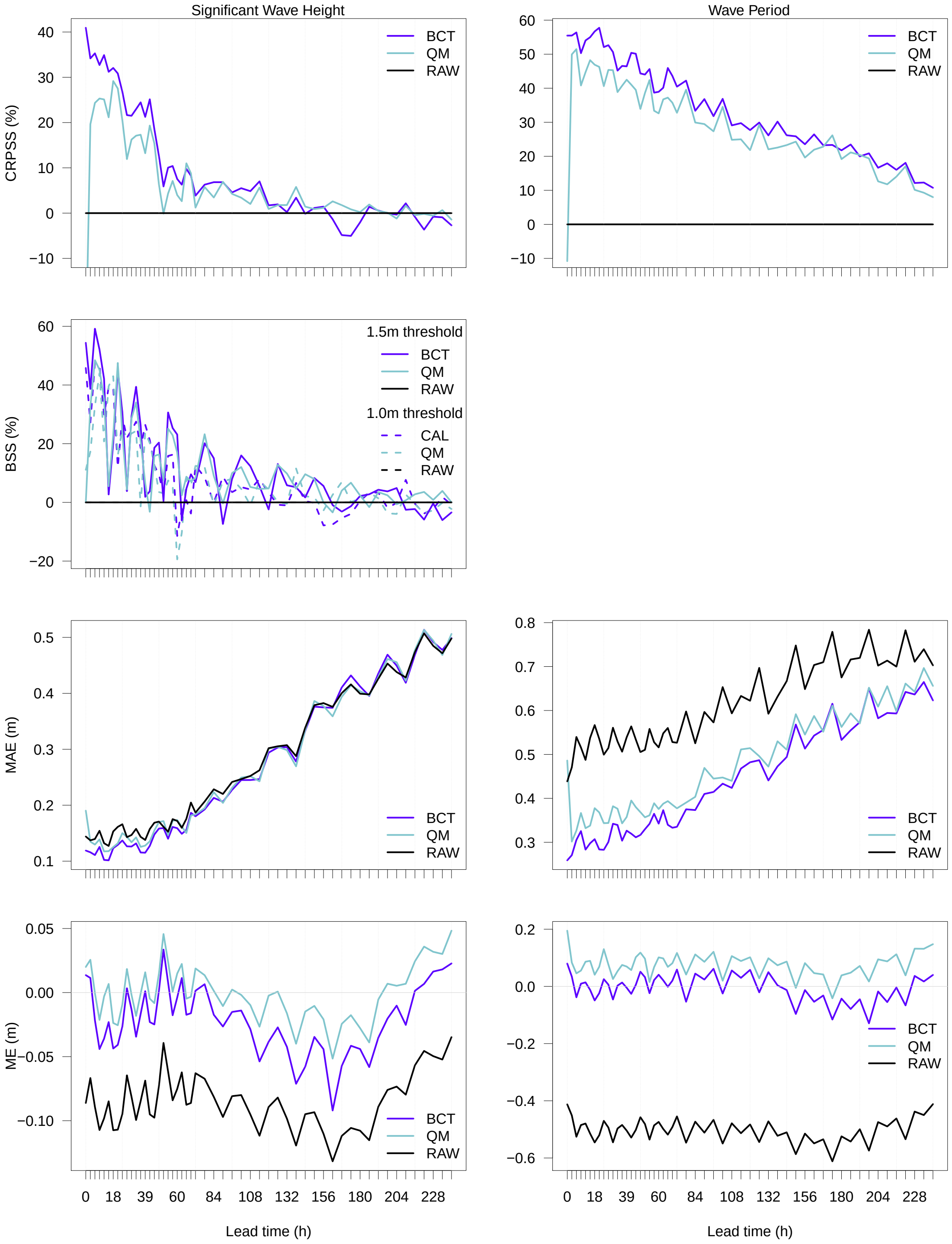


Figure 4: *Continuous ranked probability skill score (CRPSS), Brier skill score (BSS), mean absolute error (MAE) of the 50 percentile and mean error (ME) of ensemble mean as a function of lead time. Significant wave height in the left column and wave period in the right.*

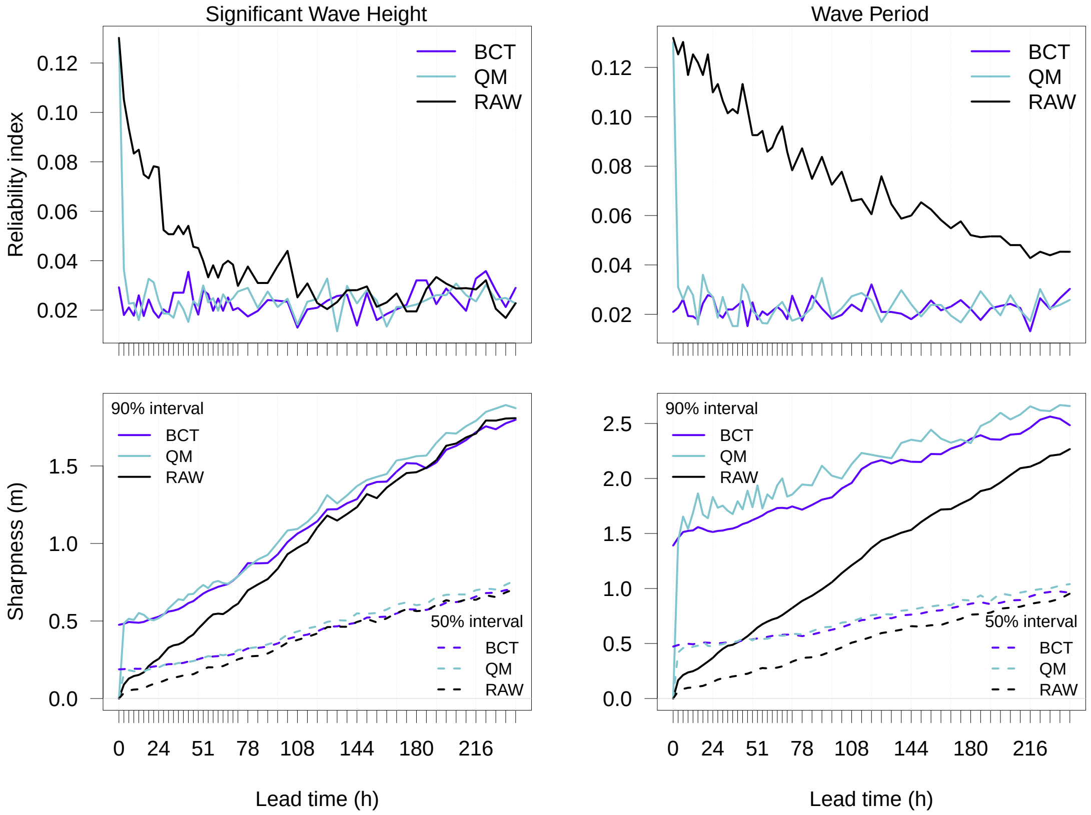


Figure 5: *Reliability index and sharpness for forecasts of significant wave height and wave period as a function of lead time. Sharpness is given in terms of the average lengths of* 50% *and* 90% *forecast intervals.*

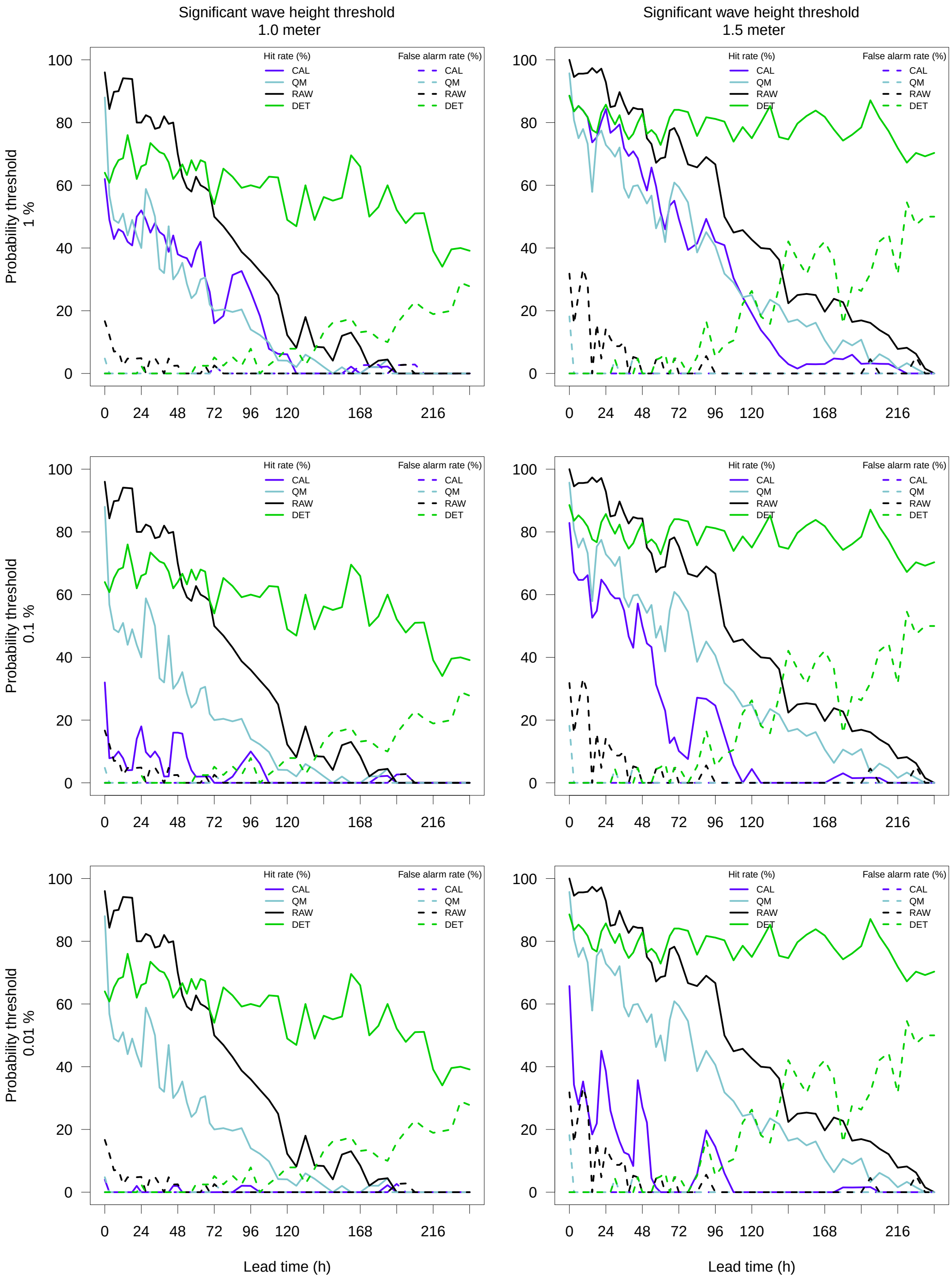


Figure 6: *Hit and false alarm rates for significant wave height thresholds* 1.0 *and* 1.5*m and probability thresholds* 0.01, 0.1, *and* 1%.

Acknowledgements

The work is part of the project “Decision support for offshore wind turbine installation” (Work Package 3.2) funded by the Norwegian Research Council and Statoil ASA. The wave measurements from FINO3 are obtained from Bundesministerium für Umwelt, Federal Ministry for the Environment, Nature Conservation and Nuclear Safety, Germany. Ensemble forecasts are obtained from European Center for Medium-Range Weather Forecasts, Reading, UK.

References

[1] Rigby, R. A. and Stasinopoulos D. M. (2005). Generalized additive models for location, scale and shape,(with discussion), Appl. Statist., 54, 3, pp. 507-554.

[2] DNV (2011). Marine operations, General. Offshore standard DNV-OS-H101.