Potential Benefits of Using Probabilistic Forecasts for Waves and Marine Winds Based on the ECMWF Ensemble Prediction System

ØYVIND SAETRA AND JEAN-RAYMOND BIDLOT

European Centre for Medium-Range Weather Forecasts, Shinfield Park, Berkshire, United Kingdom

(Manuscript received 20 March 2003, in final form 22 January 2004)

ABSTRACT

The potential benefits of using the ECMWF Ensemble Prediction System (EPS) for waves and marine surface winds are demonstrated using buoy and platform data as well as altimeter data.

For forecasting purposes, the spread of the different forecasts in the ensemble may indeed be regarded as a measure of the uncertainties in the deterministic predictions. In order to demonstrate this point, a new method is presented in which the ensemble spread is divided into different classes. An upper bound for the model errors is established by calculating the corresponding percentiles of the errors for each separate class. Using this upper bound for the model errors, a strong correlation between the ensemble spread and the deterministic forecast skill is shown.

The reliability of the probability forecasts as derived from the EPS for wind and waves is found to be good. However, the reliability diagrams indicate a small tendency for overconfidence in the wave probability forecasts for waves above 6 and 8 m. This is most pronounced in the Southern Hemisphere, whereas the reliability for the Northern Hemisphere is relatively good.

The impact of using of the wave EPS in decision making is studied by a cost–loss model for the relative economic value. For comparison, poor-man's ensembles (PMEs) are also created by adding normally distributed noise to the control forecasts. This study reveals that the real EPS performs better than both the PME and the control forecasts in terms of relative economic value. When more complex forecasting parameters are considered, such as the joint probability of wave height and period, benefits of using the EPS become even more pronounced.

1. Introduction

In June 1998, the Ensemble Prediction System (EPS) at the European Centre for Medium-Range Weather Forecasts (ECMWF) was coupled to the ocean wave model. From then on, daily ensemble wave forecasts have been available. Although the positive impact on both the atmospheric and the wave forecasts was the main reason for the introduction of the coupling (Janssen et al. 2002), probabilistic forecasts of ocean waves are also potentially very valuable products by themselves. There are a number of activities at sea where ocean waves are important and where the risk of high waves must be considered. Examples may be the towing and maintenance work on oil rigs, ship routing, or the construction of underwater pipelines (Haver and Vestbøstad 2001). During such high-risk maritime operations, ensemble forecasts of ocean waves could be a helpful tool in the decision process. To be useful, however, the forecasts need to be verified against observations.

A number of meteorological centers provide ensemble forecasts of atmospheric parameters operationally, but presently ECMWF is the only center that also produces wave ensembles. Because the wave forecasts are calculated from a nonlinear numerical model, forced by surface winds, they cannot be regarded as another output parameter from the atmosphere model. Also, the wave model includes swell, waves produced by remote winds, which introduces a memory into the wave system that would not generally be present for an atmospheric output parameter. One of the main objectives of this study is therefore to establish that ensemble wave forecasting, based on wind forcing from the atmospheric ensembles, is a feasible method in the sense that the probability forecasts are useful estimates of the observed distributions. Second, we want to verify how well the probability forecasts perform when compared with observations. In particular, we want to know how reliable the probabilities forecasted by the system are. The spread of the different forecasts in the ensemble is often used as an indicator of flow-dependent predictability of weather parameters, assuming a relationship between the ensemble spread and the skill of a deterministic model. Although the idea may be intuitively appealing, this relationship needs to be demonstrated quantitatively. To test the reliability of probability forecasts, well-

Corresponding author address: Øyvind Saetra, The Norwegian Meteorological Institute, P. O. Box 43 Blindern, 0313 Oslo, Norway. E-mail: Oyvind.Saetra@met.no

established tools such as reliability diagrams may be used. To our knowledge, no such well-established tool exists for the spread–skill relation. A number of authors have used a correlation coefficient between spread and skill, a method we later argue is quite unsuitable for this purpose. The correlation coefficient is a measure on how well a set of data pairs fit a straight line. Scatterplots of forecast errors versus ensemble spread are not anywhere near this. Nevertheless, the forecast errors tend to be distributed over a larger range when ensemble spread is large. To account for this shortcoming, a new method aimed at quantifying the relationship between ensemble spread and the expected skill of a deterministic forecast is presented. Here, the ensemble spread is divided into a number of bins. For each bin, a statistical upper bound to the forecast errors is calculated as a percentile of the forecast error.

In a recent study, Vogelezang and Kok (1999) tested the EPS waves by using two buoys in the North Sea for verification purposes for the period from October 1998 to February 1999. In our study, the forecasts are compared to buoy and platform observations for significant wave height and peak period obtained via the Global Telecommunication System (GTS). Since waves are strongly dependent on winds, wind speed is also included in the analysis. Since the majority of the buoys and platforms are located close to the continents, relatively few observations are obtained over the open oceans. To account for this shortcoming, the forecasts of significant wave height are also assessed against satellite altimeter observations, and the results are compared with those obtained from the buoy and platform observations. The study covers the period from 1 September 1999 to 31 March 2002, thus including three full Northern Hemisphere winters.

The structure of this paper is as follows: Section 2 gives a brief description of the ensemble prediction system. Section 3 describes the buoy, platform, and altimeter observations used in this study, and section 4 gives an example of the EPS forecast for an extreme event in the Norwegian Sea. In section 5, a new method of determining the relationship between spread and skill is presented. In section 6, the reliability of the probability forecasts is considered. Section 7 investigates the economic value of the EPS. Section 8 compares the EPS with the altimeter observations, and finally, the conclusions are drawn in section 9.

2. Ensemble Prediction System

In 1998 the coupling between wind and waves was introduced operationally. The resolution of the atmospheric component of the EPS was TL159L31 (spectral triangular truncation with 31 levels in the vertical), which is about 120-km horizontal resolution in gridpoint space on the Gaussian grid used at ECMWF. On 21 November 2000, the new high-resolution EPS was introduced with TL255, corresponding to approximately 80-km resolution in the horizontal. Note that this change in the EPS was carried out within the period spanned by this investigation. The changes introduced to the system during this period are reported by Buizza et al. (2003). The most important of these is the increase of resolution. For a detailed description of the atmospheric component of the ECMWF EPS, the reader is also referred to Buizza et al. (2000).

The wave model component in the EPS is the ECMWF version of the ocean wave prediction model [WAM; the Wave Modelling group (Wamdi)] cycle 4. The WAM model was developed during the 1980s by an international group of scientists (Komen et al. 1994) and marked the introduction of a new generation of ocean wave models. Since its implementation as the operational wave model at ECMWF in November 1991, the model has undergone numerous changes and improvements (Janssen 2000; Bidlot et al. 2002). The present version of the EPS wave model runs on a 110-km grid resolution with shallow-water physics and 12 directional and 25 frequency bins. This resolution was introduced into operations on 21 November 2000, at the same time as the implementation of the new high-resolution EPS for the atmosphere. The previous version, implemented in 1998, used a grid resolution of 1.5° with deep-water physics only. Thus, both the atmospheric and wave component of the EPS had a significant increase in model resolution during the period spanned by this investigation. To see whether this increase in resolution had any significant impact on the statistics for the EPS, time series of monthly mean values for bias, rms error, and frequency of outliers were plotted. We also divided the data into two batches, before and after the system change, and plotted reliability diagrams and rank histograms on both periods. It was found that for the open oceans, the change of resolution in the coupled EPS had only a small positive impact on the wave model part of the system as described in Saetra and Bidlot (2002). Hardly and differences were detected for the rank histograms and the reliability diagrams between the periods before and after the system change.

For the atmospheric component, the first 50 ensemble members are all initiated from the ECMWF analysis on which perturbations have been introduced (Buizza et al. 2000) and are therefore labeled perturbed forecasts. A last member, denoted the control run, used the unperturbed analysis as the initial field interpolated to the EPS resolution. For the waves, all ensemble members use the unperturbed analysis as the initial condition. The divergence between the wave ensemble members is therefore due only to different wind forcing when the coupled atmospheric ensemble members are subject to different evolutions (Farina 2002). In this comparison, we have used all EPS forecasts initiated from 1200 UTC. Currently, ECMWF also produces EPS forecasts from 0000 UTC, however, these forecasts were not available for the period covered in this study.

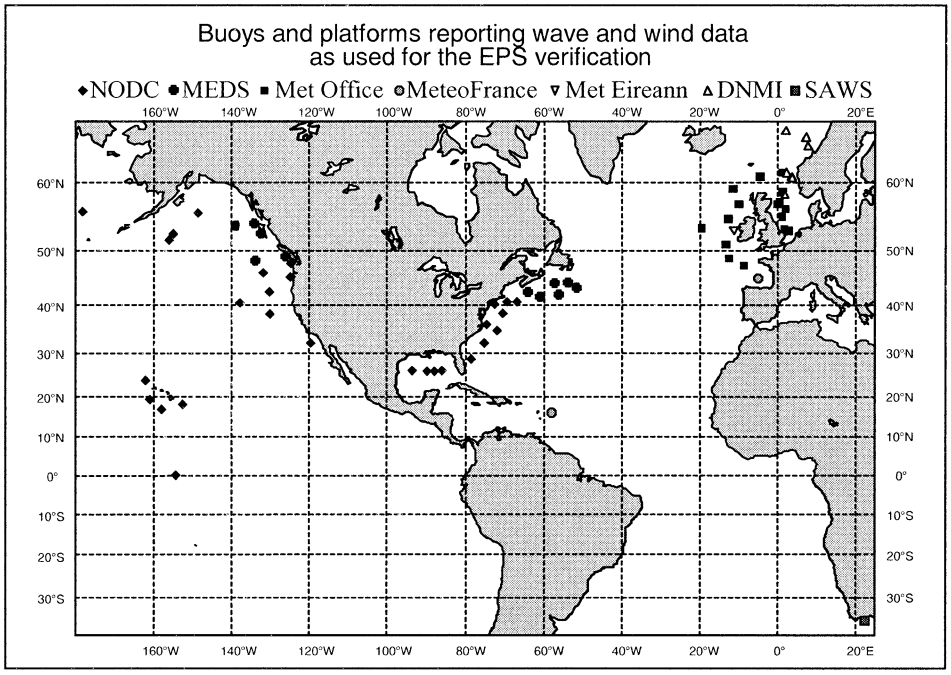

FIG. 1. Buoy and platform locations used in this study. The data providers are indicated above the map.

3. Observations

a. In situ

Sea state and ocean surface meteorological observations are routinely collected by several national organizations via networks of moored buoys or weather ships and fixed platforms deployed in their nearshore and offshore areas of interest. The geographical coverage of the data is still very limited, and at the present EPS wave model resolution, only a small number of all these stations are within the wave model grid. Nevertheless, about 66 stations that report both wind and wave data can be selected. They are well within the grid of the wave model, in relatively deep water (depth of 100 m or more), and since the EPS wave model was originally set up as a deep-water model, they have a high rate of hourly observation availability and reliability. The data coverage can be seen in Fig. 1, which shows the positions of all the observation sites. Except for one platform located off the South African coast and one buoy on the equator near Christmas Island in the Pacific, all measurements are taken in the Northern Hemisphere.

The hourly wave and wind data are transferred continually via the GTS to national meteorological centers and are usually archived with all other synoptic ship observations. In the remainder of the paper, the word buoy is used to refer to the selected moored buoys, weather ships, or platforms since most of the reliable observations come from moored buoys. Note, however, that the observation principle for waves is quite different for buoys than for platforms. Buoys usually rely on time series analysis of the buoy motion to derive wave spectra whereas radar imaging of the sea surface is employed by platforms to derive the wave spectra. Collocations between these observations and the corresponding model values interpolated to the buoy locations can easily be obtained. A direct comparison between model values and buoy and platform observations is, however, undesirable as some measurements may still be erroneous. Furthermore, model and observed quantities represent different temporal and spatial scales.

From the buoy records, time series are reconstructed and used to perform a basic quality check on the data (Bidlot et al. 2002). Spatial and temporal scales are made comparable by averaging the hourly observations in time windows of 4 h centered on 1200 UTC. It should be emphasized that buoy observations and the model represent different scales. Buoys exhibit high-frequency variability on a time scale of 1 h, which is absent in the model because the model value does represent a mean value over a grid box of size 100–150 km. In other words, since waves propagate across the area where the instrumental sensors are located, one should not consider a single observation at any given time to be equivalent to the actual statistical wave height computed at each model grid box. Averaging of the observed wave height is therefore preferable where the averaging period should match the scales still represented by the model. With a mean group velocity of 10 m s^{-1}, an averaging time of 4 h thus seems appropriate to represent a 100–150-km box-averaged observation. Not averaging the data will result in increased scatter between the model and observations, which can be linked to the high-frequency variability, not present in the model (Janssen et

al. 1997). Comparing averaged buoy data with 0000 and 1200 UTC analyses essentially yields the same statistics if done over several months. It is therefore expected that no significant differences would have been found if the 0000 UTC ensemble forecast had been included in this study. Based on the authors' experience, buoys, which are sufficiently offshore that the impact of the diurnal cycle is small, were selected for this comparison. GTS data are unfortunately provided with some truncation. Wave heights are rounded to the closest 0.1 m, wave periods to the closest second, and wind speed to the closest meter per second. Averaging will diminish the effect of these truncations. The resulting errors for wave data are well within what can be expected from buoy measurements (Monaldo 1988). It is, however, unfortunate that wind speed observations are encoded with such a large truncation error (up to 0.5 m s^{-1}) since most of them still need to be adjusted to the standard height of 10 m.

Buoy anemometers are not usually at an average height of 10 m. However, the wind observations used here are compared to model counterparts assumed to represent the wind 10 m above mean sea level. Therefore the height of the anemometers was obtained from the data providers (usually in the order of 4–5 m), and the wind speed statistics were produced by adjusting the buoy winds to 10 m. The wind speed is corrected assuming that on average the wind profile in the planetary boundary layer is neutral, as described in Bidlot et al. (2000). In case of strong winds, the neutral wind profile is usually appropriate. It might be less appropriate for low wind speeds when the boundary layer is stably stratified and decoupled. However, these low wind situations are not of interest here because we focus on the higher wind speed cases. Winds from platforms are usually adjusted to 10 m by the data providers. A reduction factor is used even though the height of the anemometer could be in the several tens of meters. Winds from platforms are therefore less reliable than buoy observations.

Roughly 46 000 wave height data values were used. Besides wave height, buoys also report wave period measurements. There is, however, no consensus on what type of period should be reported. Canadian and U.S. buoys report the period corresponding to the peak in the one-dimensional wave spectrum, the peak period (*Tp*), whereas the other data providers use a mean period, usually the zero mean crossing period (*Tz*), which can roughly be equated to the reciprocal of the square root of the normalized second moment of the frequency spectrum. The peak period has always been a standard output of the operational model; however, *Tz* has only been archived in the operational EPS since 27 October 2001. We therefore only show results based on *Tp*.

b. Wave data from satellite measurements

Wind and wave data are also available from the radar altimeter onboard the *European Remote Sensing Satellite-2* (*ERS-2*). ECMWF receives the fast-delivery ERS data in near–real time and archives them. The wave height observations from *ERS-2* are of relatively good quality and have been used by the wave model data assimilation since May 1996. Nevertheless, Janssen (2000) showed that the altimeter wave height might actually be slightly too low, especially in situations where waves are steep. A scheme for the correction of altimeter wave heights due to the non-Gaussian shape of the sea surface elevation and slope distribution was introduced in the ECMWF operational wave model in July 1999 (Janssen 2000). Deviations from a Gaussian distribution are measured by the skewness factor and the elevation–slope correlation, which depends in a complicated way on the wave spectrum. The wave model spectra are used to estimate these two quantities to derive a correction to the *ERS-2* wave height data. Before correcting the altimeter wave heights, the data are preprocessed by running them through a quality control procedure that is very similar to the one used for the buoy data except that the processing considers 30 consecutive data points following the satellite track. A few quality indicators provided with the data are also used to discard suspicious data points. The valid individual altimeter wave height data, which are available in a ± 3 h time window centered around the main synoptic times, are collocated to the closest model grid point. The average value is computed for all grid boxes with at least two individual observations. The mean position is assumed to coincide with the model grid point, and the time at the center of the time window is taken as the verifying time. These mean values are then corrected using wave model spectra available just before assimilation of the altimeter data. Both uncorrected and corrected datasets are then archived on the same grid as all the operational analysis wave model fields (roughly with a 55-km resolution) with the time stamp of the analysis.

Because these corrected wave data are averaged on the wave model grid of the analysis, it is trivial to find the gridded values that are closest to the EPS grid points and to retrieve the corresponding EPS wave heights for all output forecast steps considered in this study (0, 24, 48. . . h). Note that the altimeter data were averaged over the operational analysis grid of roughly 55 km and not on the EPS grid. Nevertheless, 310 000 gridded observations were obtained. Fast-delivery wave heights for low wave heights (below $\sim$1.5 m) are known to be overestimated (Challenor and Cotton 1997); however, no corrective fit is used here to remove this inconsistency because the comparison between altimeter wave heights and the EPS is intended to focus on high waves.

The altimeter wind speeds are not yet processed onto the wave model grid and are therefore not used in the verification. Furthermore, the ECMWF verification of the *ERS-2* fast-delivery wind speed product has identified a few periods during which the quality of the retrieved winds had degraded. It was therefore decided not to use this data.

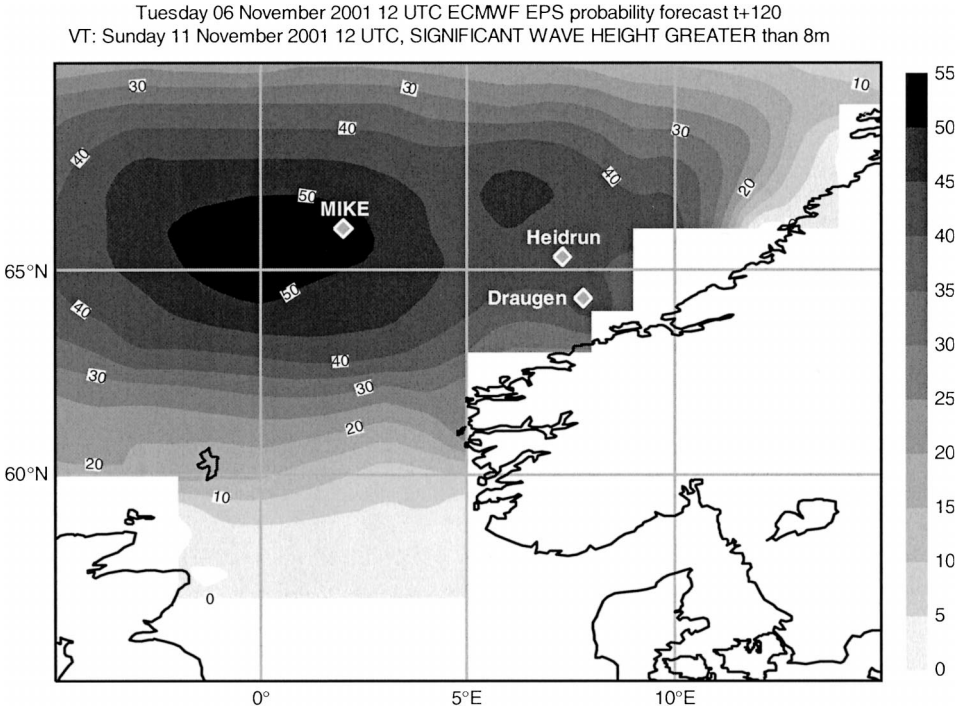

FIG. 2. Day 5 probability forecast that significant wave heights would exceed 8 m, 1200 UTC 11 Nov 2001. The three stations mentioned in the text are marked by their names.

c. Relative error estimation

By collecting buoy and altimeter observations, model analysis, and model hindcast, Janssen et al. (2003) were able to estimate the relative random error of each dataset (with respect to the mean). They found that the buoy wave height relative errors were of the order of 9%–12% and 7%–11% for the *ERS-2* fast-delivery wave height. Somewhat similar values were found for the wind speed errors.

4. Example of the wave EPS

On the night of 10 to 11 November 2001, extreme wave conditions were experienced in the Norwegian Sea (Haver and Vestbøstad 2001). At two oil rigs, Heidrun (65.30°N, 7.30°E) and Draugen (64.30°N, 7.80°E), significant wave heights in excess of 15 m were observed; the highest individual wave was of the order of 25 m. Draugen has been in operation since 1994 and Heidrun since 1996. In the vicinity of where Heidrun is now, a buoy had also been deployed between 1980 and 1988. The waves observed in November 2001 were the largest ever recorded at these two locations. Also, at the weather ship *Polarfront* (Station Mike: 66°N, 2°E), which is positioned farther out at sea, a maximum significant wave height of 15.5 m was measured during this storm. The *Polarfront* has measured waves regularly since the late seventies and had only recorded a wave height of this magnitude two times earlier. For reference, see Reistad et al. (2003) and Haver and Vestbøstad (2001).

On a daily basis, ECMWF issues global forecasts of the probability of significant wave height above 2, 4, 6, and 8 m based on the EPS. Looking at the probability forecast 5 days ahead of the 10–11 November event for this area, it is obvious that something dramatic was about to take place. Figure 2 shows the day 5 probabilities of waves exceeding 8 m. The positions of the weather ship and the two oil platforms are marked with their respective station names in this plot. In the area where the weather ship was positioned, more than a 50% probability of waves above 8 m was predicted. For both Heidrun and Draugen, the forecast probabilities of waves above 8 m were between 40% and 45%. Taking into account the fact that the ECMWF wave model tends to underestimate extremes (Bidlot et al. 2002), it is obvious that the ensemble forecast provided an early warning 5 days ahead of this extreme situation.

To have a closer look at the actual ensemble forecasts in this case, the plume diagram showing the forecasts from 1200 UTC 6 November 2001, for Heidrun is given in Fig. 3. The plot gives the swell wave height, wind speed, and the significant wave height. The respective deterministic high-resolution forecast is plotted with a thick solid line, and the control forecast with a dotted line. The individual ensemble members are indicated with thin gray lines. Although none of the ensemble members predicted waves above 15 m, five of the members were above 12 m, and one member was slightly above 14 m. It is important to note that this plot is based on 12-hourly output, noon and midnight. From the wave

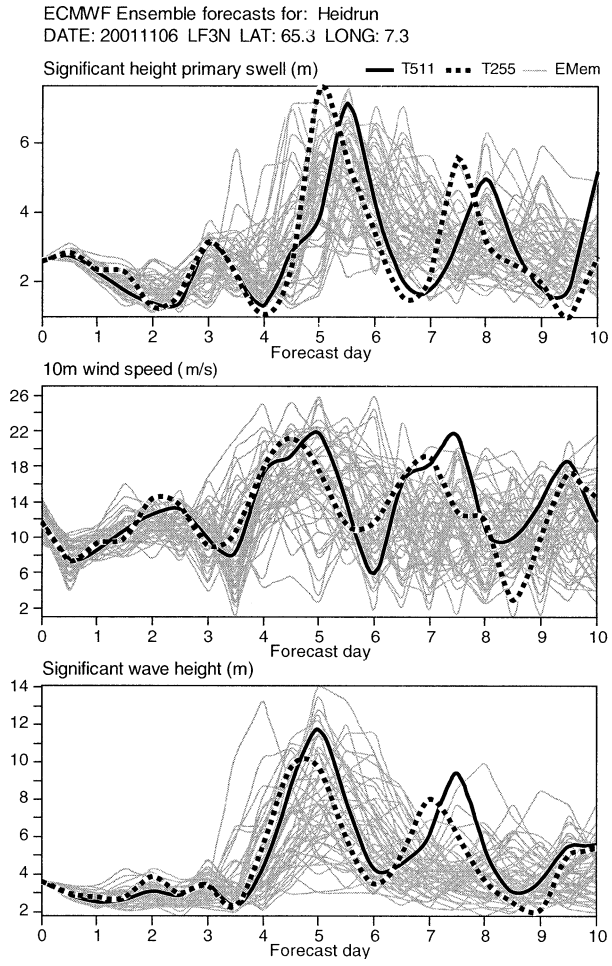

FIG. 3. Plumes showing the deterministic high-resolution and ensemble forecasts for platform Heidrun (65.30°N, 7.30°E) for top the significant wave height of the swell part of the wave spectrum, (middle) the wind speed, and (bottom) the significant wave height. The deterministic high-resolution forecasts are given by the thick solid lines, and the control forecasts by the thick dashed lines. The thin solid lines are the ensemble members. The forecasts were issued at 1200 UTC 6 Nov 2001.

recording taken at the Heidrun, the largest waves were measured between 0500 and 0800 UTC. At both 0000 and 1200 UTC the measured wave heights were about 12 m. Note that those output steps were the only ones available at the time. Currently, wave EPS forecasts are output every 6 h.

5. Relation between ensemble spread and skill

We now compare the ensemble forecasts with the buoy observations. A similar comparison with altimeter wave heights is presented in section 8. The spread of the EPS for wind and waves has been tested by plotting rank histograms (Hamill 2001) for wave height, peak period, and wind speed. The results of this comparison are given by Saetra and Bidlot (2002) and Saetra et al. (2002). When effects of observation errors are taken into account, the system has a slightly too low spread. For significant wave height, the number of outliers—that is, the frequency of the observations being either the lowest or highest value—is a little less than 10% for the day 5 forecasts, compared to the theoretical value of 3.8% for a 51-member ensemble. For the peak period, the number of outliers is even less. Here, however, the outliers tend to be mainly in the lower rank. The reason for this is probably caused by the positive bias, meaning the forecasts are too high, for the peak period. For the control forecast, the bias is about 0.6 s at day 5 (Saetra and Bildot 2002). Similarly for the wind speed, the number of outliers is only about 6% at day 5.

One easy way to utilize the ensemble forecasts is to use the ensemble spread to determine the uncertainty of the corresponding deterministic forecasts. Forecasters tend to be more confident in deterministic forecasts when the ensemble spread is small than when the spread is large. The assumption is based on the idea that the ensemble spread can be a measure of the flow-dependent predictability. For example, stable weather regimes are less sensitive to errors in the initial state and the ensemble members stay closer for a longer period than during unstable weather regimes. From time series of all ensemble members, the width of the plume may then be used to estimate when predictability is lost. For this to be a meaningful approach, however, the relation between spread and skill has to be demonstrated. Some authors have computed the correlation coefficient between the ensemble spread and the forecast errors (Buizza et al. 2000). We argue that since the correlation coefficient is a measure of how well a set of data points fit a straight line, it is unsuitable for determining the relationship between spread and skill. Figure 4 shows a density plot of ensemble spread versus forecast errors for the wave height at day 5. Here, the forecast errors are the absolute difference between the observations and the control forecasts. The ensemble spread is defined as the difference between the upper and lower quartiles of the ensemble. The dashed line gives the best linear fit with a correlation coefficient of 0.433. As is clear from this plot, the data do not fit to a straight line, which is also reflected by the low correlation coefficient. Still, for the cases with large spread, the errors are distributed over a larger range. Thus, it seems from this plot than an upper statistical bound to the forecast errors exists, and this upper bound is an increasing function of the ensemble spread.

In order to test this, we take as an example the 90 percentile of the absolute errors as a measure of a statistical error bound. For a given spread, we are then seeking the value that separates the 10% largest errors from the rest of the data. Since the spread is also a stochastic parameter, it should be treated equally by using percentiles. In this paper, therefore, we refer to spread as the interquartile range. To relate the percentiles of the forecast errors to the ensemble spread, it is necessary to divide the spread into different classes or

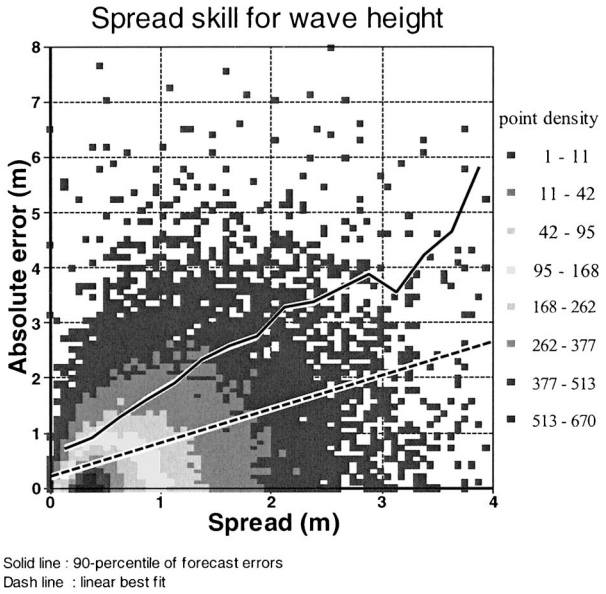

FIG. 4. Density plot of ensemble spread vs forecast errors for the wave height at day 5. All buoy data were used. The dashed line is the best linear fit, and the solid line is the 90 percentile of the forecast errors (see text for details).

bins, and then rank the observed errors within each class to find the value that constitutes the boundary between the 10% largest errors and the rest.

In Fig. 5, the 90 percentile of the absolute error for significant wave height and wind speed is given as a function of the ensemble spread for the day 5 forecast range. This plot also shows the result when the data were divided into different seasons. For both wave height and wind speed, the 90 percentile shows a dependency on the ensemble spread. In fact, the correlation coefficients for these two cases are 0.966 and 0.989 for wave and wind, respectively. The correlation is most apparent for spread below 3.5 m. As seen from the histogram inserted in the plot, the number of data pairs are very low for spread values above this. The data have also been divided in different areas (not shown here), and hence different variability. Nonetheless, the spread–skill relationship calculated here did not show any regional or seasonal dependencies.

We may look at it the other way around such that for a fixed error, the 90 percentile of the ensemble spread is calculated. This is shown in Fig. 6 for wave height and wind speed. As before, the data have also been divided into four seasons, as shown by the symbols in the plots. The relationship is still apparent, but with slightly lower values for the correlation coefficients, which are 0.938 and 0.944 for wave and wind, respectively.

Of course, the choice of percentile for the observed errors in this case is more or less arbitrary; any other percentile is expected to give qualitatively similar results, that is, except maybe for a percentile close to 100, for which the percentile may be very sensitive to a few outliers. In Fig. 7, the relationship between the spread and the 75, 80, 85, and 90 percentile for significant wave height and wind speed are given. Again, the forecast range in the example is day 5. Note that the gradient is steeper for the waves than for the wind speed. For the wave height, the 75 percentile fits roughly with the diagonal line. A very approximate rule of thumb may then be that the error in the wave forecasts is expected, with 75% probability, to be less than or equal to the interquartile range of the wave ensemble.

6. Reliability of the probability forecasts

To test how well an ensemble forecasting system is predicting the probability of certain events, reliability diagrams are useful tools (Wilks 1995). In Fig. 8, the reliability diagrams for the day 5 forecast probabilities of waves (*Hs*) above 2, 4, 6, and 8 m are plotted. For a given event, the forecast probabilities are split into discrete bins ranging from zero to one. For each probability class, the fraction of times the event is observed (with respect to the total number of ensemble forecasts in that class), defined as the observed frequency, is plotted against the corresponding forecast probability. For a perfectly reliable forecasting system, these points lie on the diagonal line. The probabilities have been calculated in the simplest way possible, by dividing the number of ensembles that forecast an event with the total number of ensemble members. It is important to keep in mind that this method has limitations, in particular for probabilities in the tail of the distribution. Generally, the results indicate good reliability, particularly for the 4-m threshold. For the 2- and 6-m thresholds, the reliability is also quite good, but there is a small tendency for the points to lie below the diagonal line, which indicates that high probabilities are forecasted slightly too often. For the 8-m threshold, the reliability curve shows the behavior typical for situations with insufficient sample size. Reliability diagrams for wind speed are given in Fig. 9. Here, the threshold wind speeds (*Ws*) are 10, 14, 17, and 20 m s^{-1}. As for the waves, the model apparently has a tendency to overforecast, in particular for the highest wind speed thresholds.

In Fig. 10, we have calculated the reliabilities of the joint probability of wave height and period for the day 5 forecasts. Four different combinations of wave height and peak period (IF) thresholds have been chosen. A summary of the threshold values for all four cases is given in Table 1. Two relatively low threshold levels, 2 and 4 m for wave height, have been used. For 2-m wave height, the intervals are for periods between 3.5 and 6.5 s and between 5.5 and 8.5 s. These are referred to as case 1 and case 2, respectively. The reliability diagram for case 1 is typical for rare events, but with relatively good reliability. The second case is much more common, as the reliability curve also reveals. For the 4-m wave height threshold, the periods are between

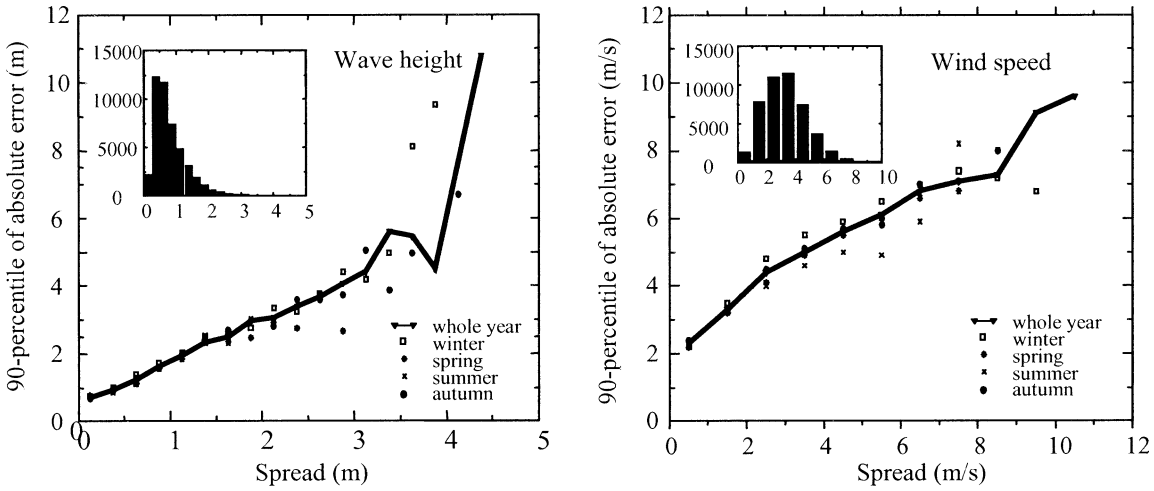

FIG. 5. Day 5 forecast spread–skill for (left) wave height and (right) wind speed, using the 90 percentile of forecast errors. The histograms in the upper-left corner of the plots show the frequency distribution of the spread bins. All buoy data were used. The solid line is the result when all available data are taken into account. The squares, stars, cross, and circles denote the results for winter, spring, summer, and autumn, respectively. Winter is defined as Dec, Jan, and Feb; spring as Mar, Apr, and May; summer as Jun, Jul, and Aug; and autumn as Sep, Oct, and Nov.

7.5 and 10.5 s (case 3), and between 9.0 and 13.0 s (case 4). For this threshold level, case 3 shows the behavior typical for rare events with relatively good reliability. Case 4 shows quite good reliability, but this is a much more common situation.

In Fig. 11, the Brier skill scores (see appendix A) for all forecast ranges have been plotted for the wave height, wind speed, and joint probability of wave height and period. The threshold values are the same as in the reliability figures. Here, the sample climate has been used as reference, and the limit when predictability drops below the climatological value is marked with a dashed horizontal line. The skill scores for the wave height are rather good for all threshold values. For waves above

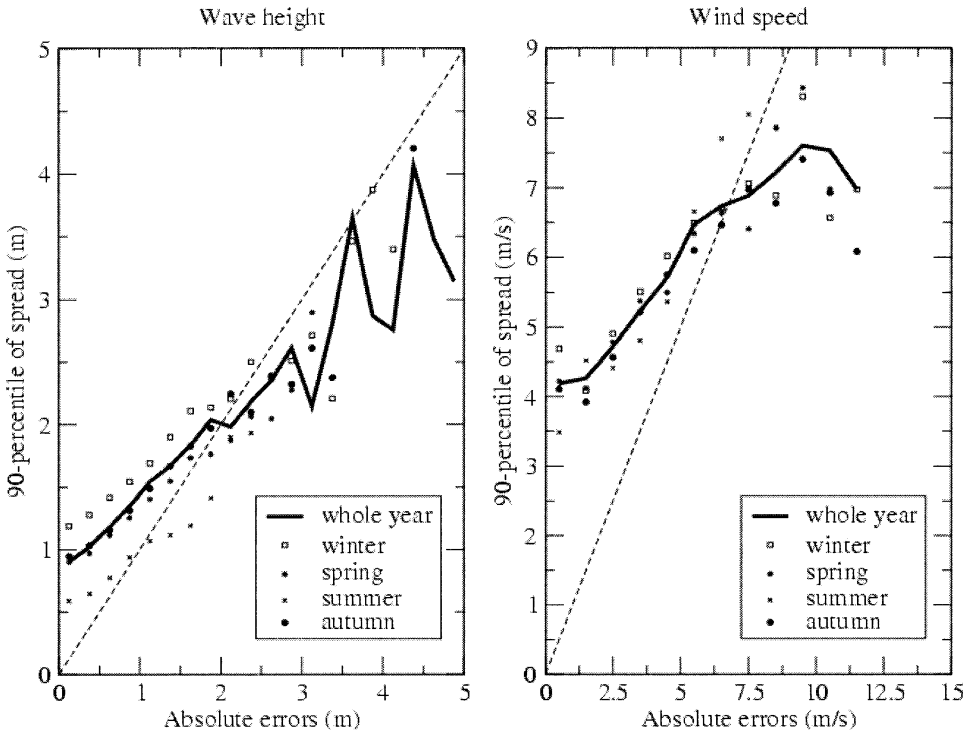

FIG. 6. Day 5 forecast for 90 percentile of spread vs absolute errors for (left) wave height and (right) wind speed. The perfect-reliability diagonal is the dashed line. All buoy data were used.

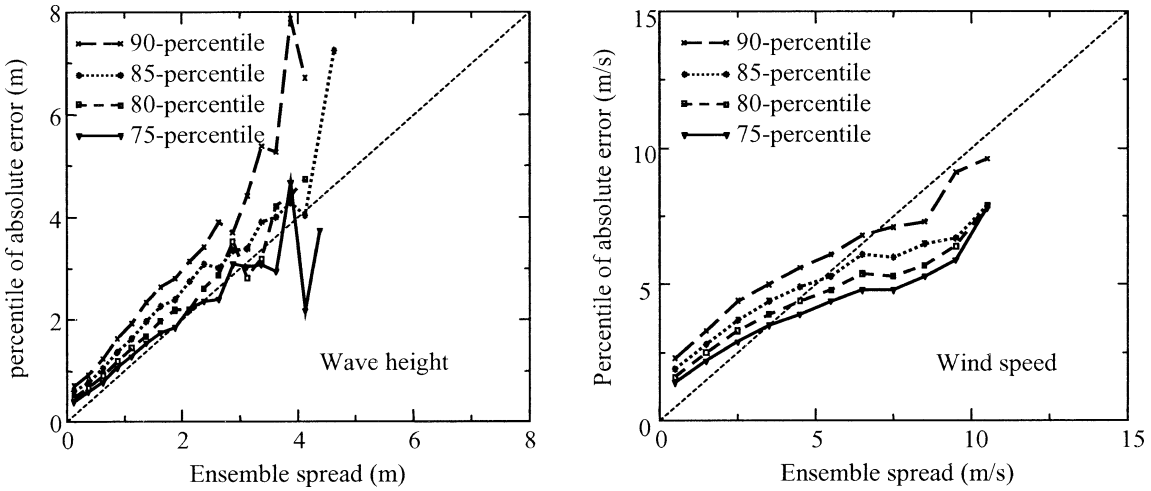

FIG. 7. Day 5 forecast spread–skill for (left) wave height and (right) wind speed, using different percentiles as upper bounds to errors. All buoy data were used.

8 m, predictability is lost around day 6, whereas for the lower wave height values the Brier skill scores are above zero even at day 10. For wind speed the scores are less impressive. For wind speeds above 20 m s^{-1}, predictability is already lost between days 2 and 3. For the two lowest wind speed thresholds, the skill score stays above zero for all forecast steps. The reason for this striking difference between the scores for wind speed and wave height is most likely explained by the presence of swell in the wave system. Waves may propagate over long distances and thus contain a memory that is not present for the winds. For the joint probabilities, the skill scores are lower than for both wind and waves. Note that for case 1, which is a rather unusual combination of wave height and period, the scores are below zero for all forecast steps. Despite the fact that the day 5 reliability curve is close to the diagonal line for this case, the forecast is beaten by climatology at all forecast ranges

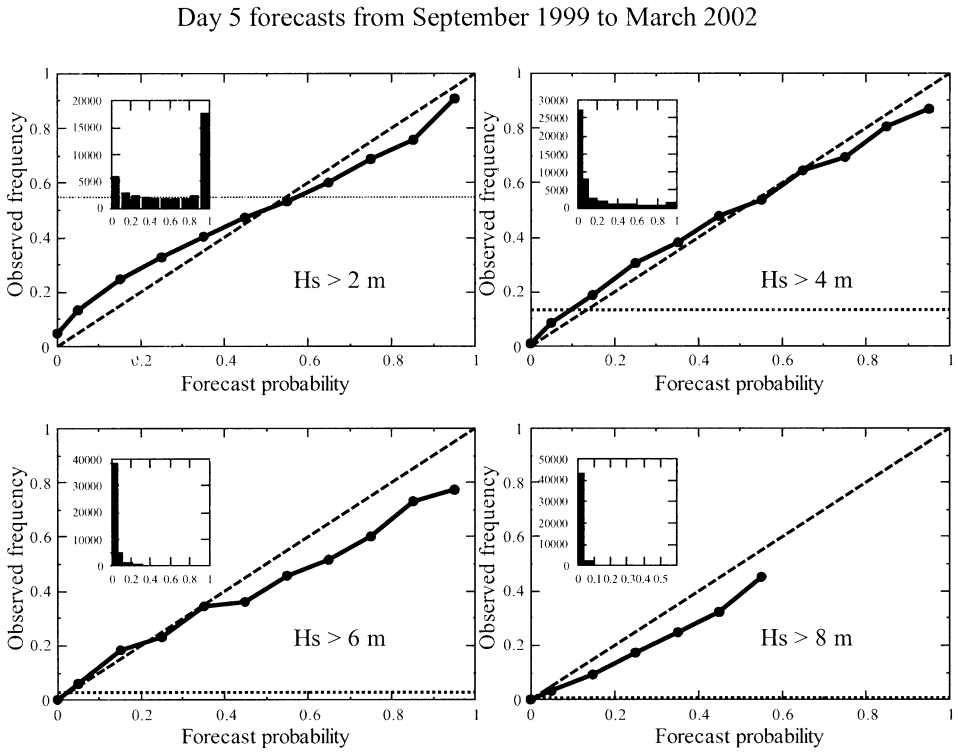

FIG. 8. Day 5 reliability diagram for wave height. All buoy data were used.

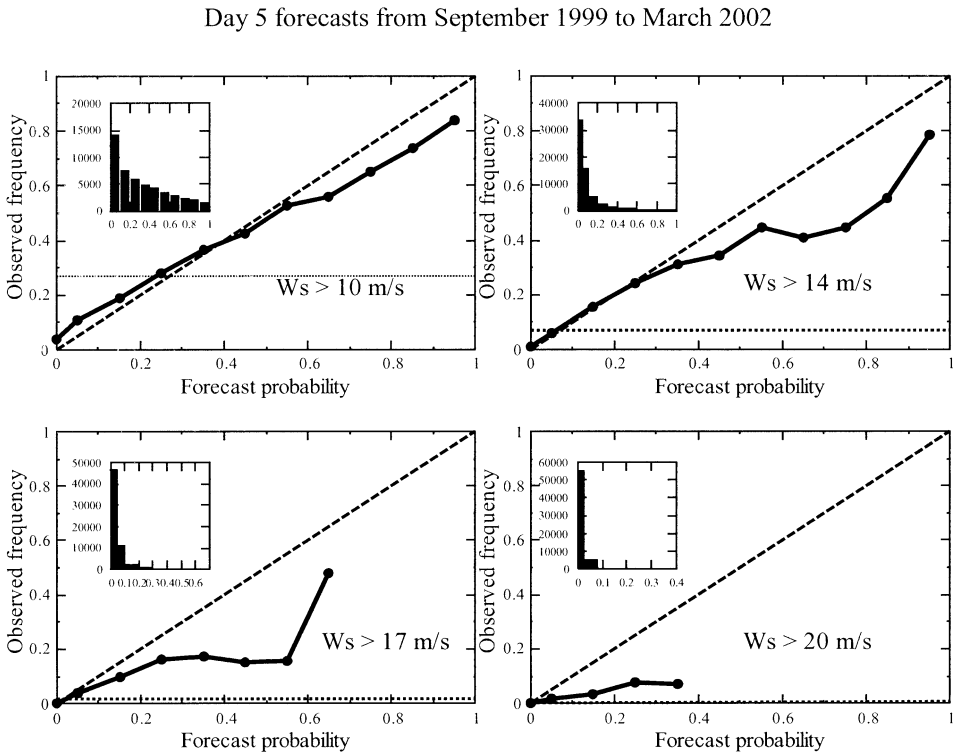

FIG. 9. Day 5 reliability diagram for wind speed. All buoy data were used.

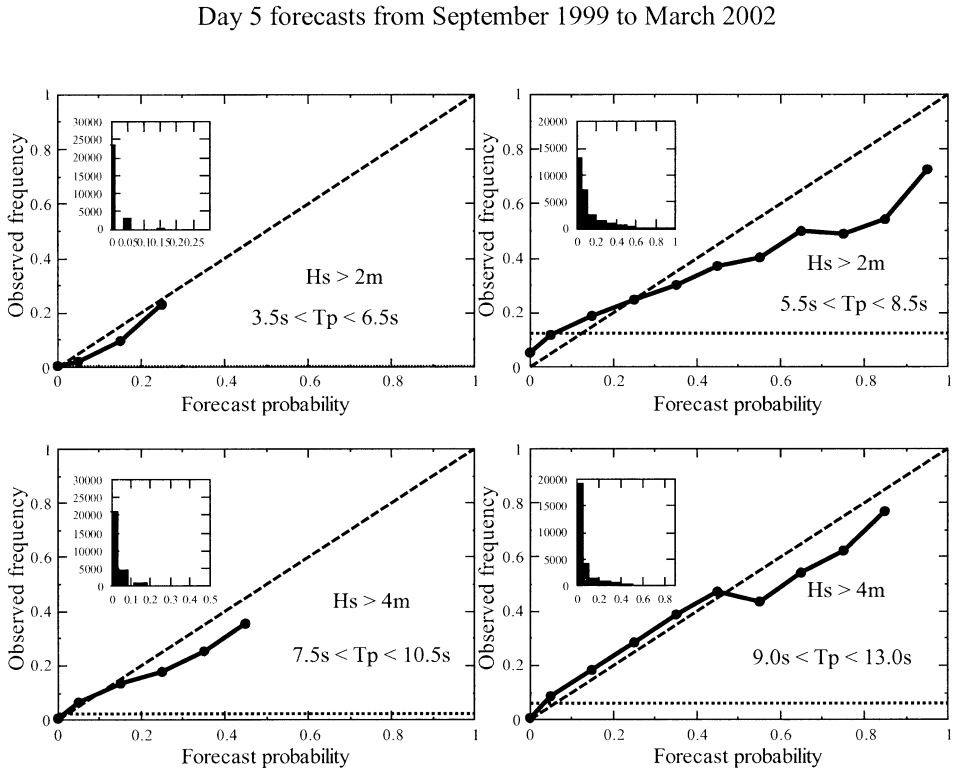

FIG. 10. Day 5 reliability for the joint probability of significant wave height and peak period. Only the U.S. and Canadian buoys were used.

TABLE 1. Cases for the joint probabilities of wave height (*Hs*) and peak period (*Tp*).

	Case 1	Case 2	Case 3	Case 4
Hs larger than	2 m	2 m	4 m	4 m
Tp larger than	3.5 s	5.5 s	7.5 s	9.0 s
Tp smaller than	6.5 s	8.5 s	10.5 s	13.0 s

according to the Brier skill score. This is discussed more in the next section, where the relative economic value for the same cases is considered.

7. Economic value of the EPS

It is important to assess the economic value of the ensemble forecasts. In many operations involving weather-related risks, the decision of whether to carry out the operation or not must be taken at some point, while the potentially dangerous part of the operation may lie several days into the future. For instance, when an oil rig is to be towed, a perilous part of the operation is the installation of the platform at the operation site, in some cases many days after the start of the operation. In such cases, the ensemble forecast should provide valuable information.

As an example we consider a hypothetical company that frequently operates a vessel out to an oil rig in the North Sea. If the wave height exceeds a certain threshold level, the company will experience on average a loss, *L*, due to damages to the vessel or goods. If this situation is forecasted, the company can take some protective measures to avoid part of the loss. This, however, involves the costs, *C*. If L_o is the part of the average loss that is saved by taking protective action, the cost–loss ratio is defined as $\alpha = C/L_o$. Now, three different strategies can be taken. The first one is to base the decision on the monthly mean wave climate for the area. For months when the climatological wave height is larger than the maximum wave height in which the ship can safely operate, protective measures are taken daily during the whole month, otherwise no action is taken. The second strategy is to base the decision on a deterministic wave forecast from a forecasting center. In this case, action to protect is taken only when the model predicts higher waves than the threshold value. Option three is to base the decision on a certain probability, *P*, that the wave height threshold will be exceeded. The probability can then be calculated from the ensemble forecasting system. If the EPS probabilities are reliable, the company should on average save money if action to protect is taken whenever $C < PL_o$. Therefore, the probability threshold to decide whether to take action or not is when the probability of dangerous sea state exceeds the cost–loss ratio. For this hypothetical example we have chosen a company that frequently anchors at platform Auk A in the North Sea (56.40°N, 2.0°E). For the cost of taking action, we have chosen $C = \$500$ and for the cost–loss ratio $\alpha = 0.1$. Since the real weather is not always as predicted, the company will have to endure the cost *C* whenever the event for which action was taken did not occur. Similarly, it will have to suffer a loss if the event did happen but it did not protect for it. The buoy observations from this location have then been used to calculate this additional expenditure the company would experience depending on which of the strategies mentioned above is used. For this example, we have selected

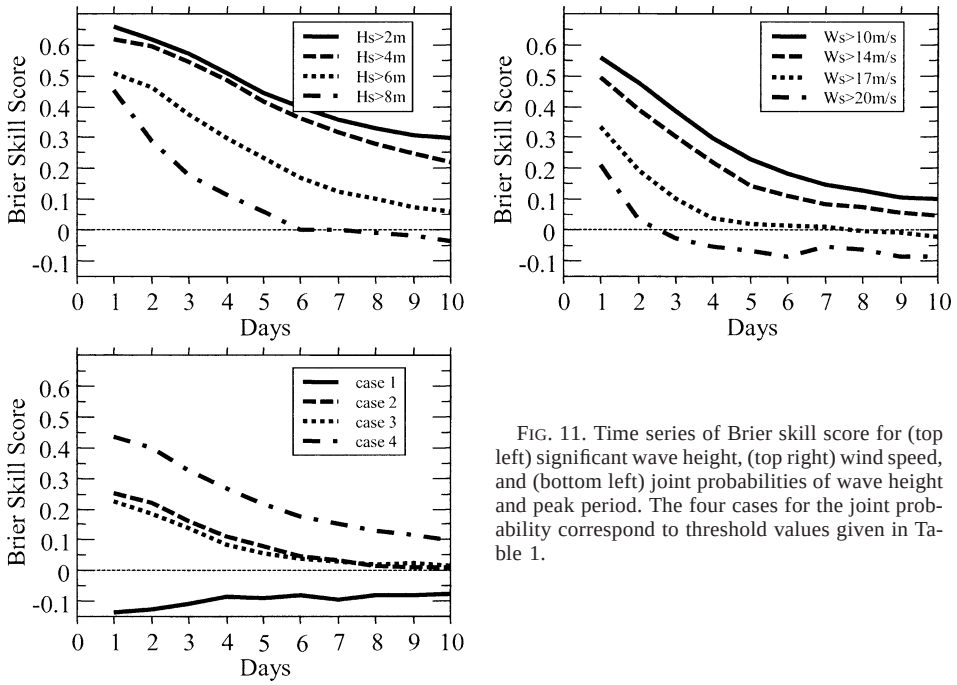

FIG. 11. Time series of Brier skill score for (top left) significant wave height, (top right) wind speed, and (bottom left) joint probabilities of wave height and peak period. The four cases for the joint probability correspond to threshold values given in Table 1.

TABLE 2. Monthly mean significant wave height for platform Auk A in the North Sea.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Hs (m)	2.05	2.45	2.04	1.74	1.11	1.47	1.27	1.10	1.78	2.06	2.38	2.58

three different threshold levels for the wave height: 2, 3, and 4 m. The monthly mean wave heights for Auk A, based on the buoy observations in this study, are given in Table 2. For wave heights larger than 2 m, the monthly mean values are larger during 6 months of the year. For the strategy based on wave climate, the cost of taking action is experienced for each journey during these 6 months. For the wave height thresholds of 3 and 4 m, the monthly mean values are always smaller. For these cases, the strategy based on climate will experience a loss L every time wave height observations exceed the threshold level. The decisions for the two other strategies are based on day 3 forecasts. For the deterministic forecasts, we use the control forecasts. The additional expenditure for the three different strategies is given in Table 3. For comparison, we have also given the extra expenditure associated with a perfect knowledge of the future weather (i.e., the observations). These data are based on 734 journeys to the platform. In all cases, the optimal strategy is to use the ensemble forecasts, and the second best is to use the deterministic forecasts. The benefit of using forecasts is most pronounced for the 3-m wave height threshold, where the expenses based on climatology are more than 3 times those based on the ensemble forecasts.

The result obtained above is of course strongly dependent on the choice of parameters and wave height threshold. A more general approach has been suggested by Richardson (2000). This method is also well suited for comparing the relative value of ensemble forecasts with that of traditional deterministic forecasts. Based on the observations and the corresponding forecasts, the hit rate and false alarm rate can be calculated. The relative economic value is defined such that zero is equal to the economic value of using climatology and one is equal to the economic value of using a perfect forecasting system. For details, see Richardson (2000).

In Fig. 12, the relative economic value of the ensemble wave forecast and a deterministic forecast, represented here by the control forecast, is given as a function of the cost–loss ratio. In addition, the EPS has been compared to the poor-man's ensemble (PME), which was constructed by adding normally distributed noise to the control forecasts. The standard deviation used for this is 0.96 m, which is the root-mean-square error for the day 5 forecasts for waves. These latest figures correspond to the reliability diagrams in Fig. 8 and show the results for the day 5 forecast for waves above 2, 4, 6, and 8 m. For the curves corresponding to the EPS, the appropriate probability level has been found by calculating the expression for a discrete set of probabilities ranging from 0 to 1 for each cost–loss value and choosing the one that maximizes the economic value. The benefit of using the EPS is again apparent in most cases.

The relative economic value of the forecasts for the joint probability of wave height and period is given in Fig. 13. The threshold levels here correspond to the four cases in Table 1. The standard deviation used to create the PME for the peak period is 2.71 s, corresponding to the root-mean-square error of the control forecast at day 5. Encouragingly, for all cases shown, the relative economic value of the EPS is larger than the value of both the control forecast and the PME. As mentioned in the previous section, cases 1 and 3, on the left-hand side of the plot, represent rare combinations of wave height and period. For these cases, relative economic values above climatology are obtained only for very low cost–loss ratios (note the logarithmic cost–loss axis). The results indicate that the relative difference between the PME and the EPS is larger for rare, or complex, situations. It is important to remember that this is strictly dependent on the correct choice of probability level for deciding whether or not to take action. For case 1, this

TABLE 3. Additional expenditures (U.S. dollars) when using ensemble forecasts, deterministic forecasts, or climatology to make the decision of whether to take protective action or not. For comparison, the expenditures based on perfect knowledge of the future weather are also given.

	Ensembles	Deterministic	Climatology	Perfect forecast
$Hs > 2$ m	324 500	544 000	562 000	148 000
$Hs > 3$ m	209 500	332 000	710 000	71 000
$Hs > 4$ m	107 000	186 000	285 000	28 500

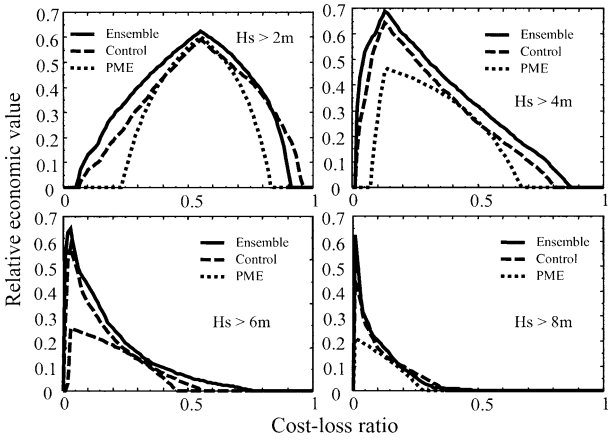

FIG. 12. Relative economic value for the day 5 wave height forecasts as a function of the cost–loss ratio. The threshold levels are 2, 4, 6, and 8 m and correspond to the values used for the reliabilities in Fig. 8.

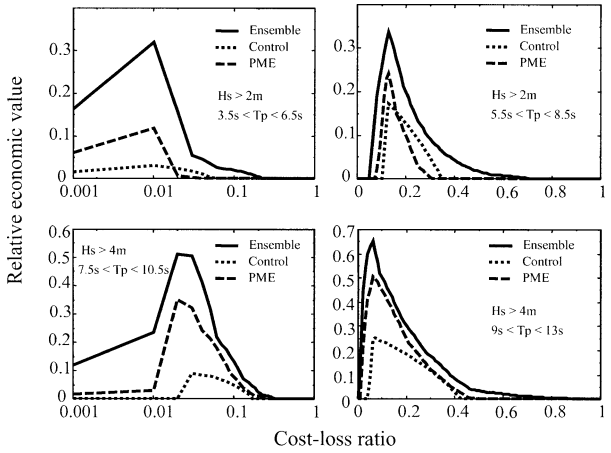

FIG. 13. Relative economic value for the day 5 forecasts of the joint probability of wave height and peak period (see Table 1 for threshold levels). Note that the two plots on the left-hand side represent rare combinations of wave height and period and exhibit economic value above climatology only for cost–loss ratios below 0.1. To see the differences more clearly, these two graphs are plotted with logarithmic scale along the cost–loss axis.

result is more encouraging than the results obtained in the previous section, where all scores showed skill below climatology. For the relative economic value the EPS performs better than climatology at least for very low cost–loss ratios. However, care must be taken when interpreting these results since the number of cases when the wave height and period combination for case 1 were observed was very low. Of more than 27 000 observations, this combination appeared only 230 times.

Further insight into the performance of different forecasting systems can be obtained by comparing their respective hit rates and false alarm rates. The relative operating characteristics (ROC) curves can be obtained by plotting the hit rate against the false alarm rate for probabilities from 0 to 1 associated to different thresholds. For a totally random forecasting system, the hit rate will be equal to the false alarm rate and will result in points along the diagonal line. A perfect forecasting system, with hit rate of 1 and a false alarm rate of 0, would give one point in the upper-left corner of the graph. Here, instead of showing plots of the ROC curves, we concentrate on the areas under these curves. These ROC areas can be used as an index of accuracy, with 1 for a perfect system and 0.5 for totally random forecasts. The results for wave height and joint probability of waves and periods are given in Tables 4 and 5. For wave height, the thresholds are the same as for Fig. 12. For the joint probabilities of waves and periods,

TABLE 4. Area under the ROC curves for significant wave height and four different thresholds.

	$Hs > 2$ m	$Hs > 4$ m	$Hs > 6$ m	$Hs > 8$ m
Ensemble	0.880	0.912	0.914	0.858

TABLE 5. Area under the ROC curves for joint event of peak period and significant wave. Threshold values are given in Table 1.

	Case 1	Case 2	Case 3	Case 4
Ensemble	0.682	0.725	0.805	0.876

the four cases correspond to the thresholds used in Fig. 13.

8. Comparison with altimeter data

To compensate for the shortage of buoy and platform data for the open oceans in general, and the Southern Hemisphere in particular, the wave EPS has in addition been compared to satellite altimeter data covering the whole globe. The data cover the same 3-yr period as the buoy and platform data. The data shown in the plots focus on the results for the day 5 forecasts. All the other forecast ranges, from day 1 to day 10, have been examined and the main impression is that the results below are, in general, representative of the other forecast steps. Starting with the spread–skill relationship, given in Fig. 14, the absolute errors show a clear relationship with the ensemble spread. As before, the spread is taken to be the interquartile range of the ensemble forecasts. In addition to the global results, the figure also shows the spread–skill relation for different regions: Northern Hemisphere, Southern Hemisphere, Tropics, North Atlantic, and Pacific. Here, the North Atlantic and North Pacific are defined as the areas of these oceans that are poleward of 20° N.

Figures 15 and 16 show the reliability graphs for the Northern and Southern Hemispheres, respectively. The corresponding results for the Tropics are given in

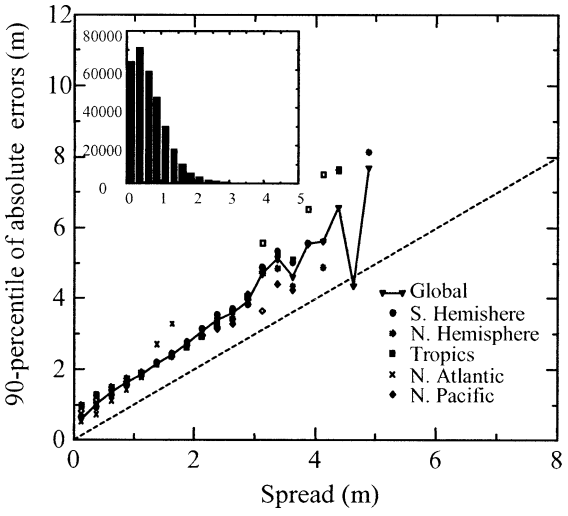

FIG. 14. Day 5 forecast spread–skill based on the verification against altimeter wave height using the 90 percentile of the forecast errors. The frequency in the various bins for ensemble spread are depicted in the bar diagram. The Tropics are defined as the area between 20°N and 20°S.

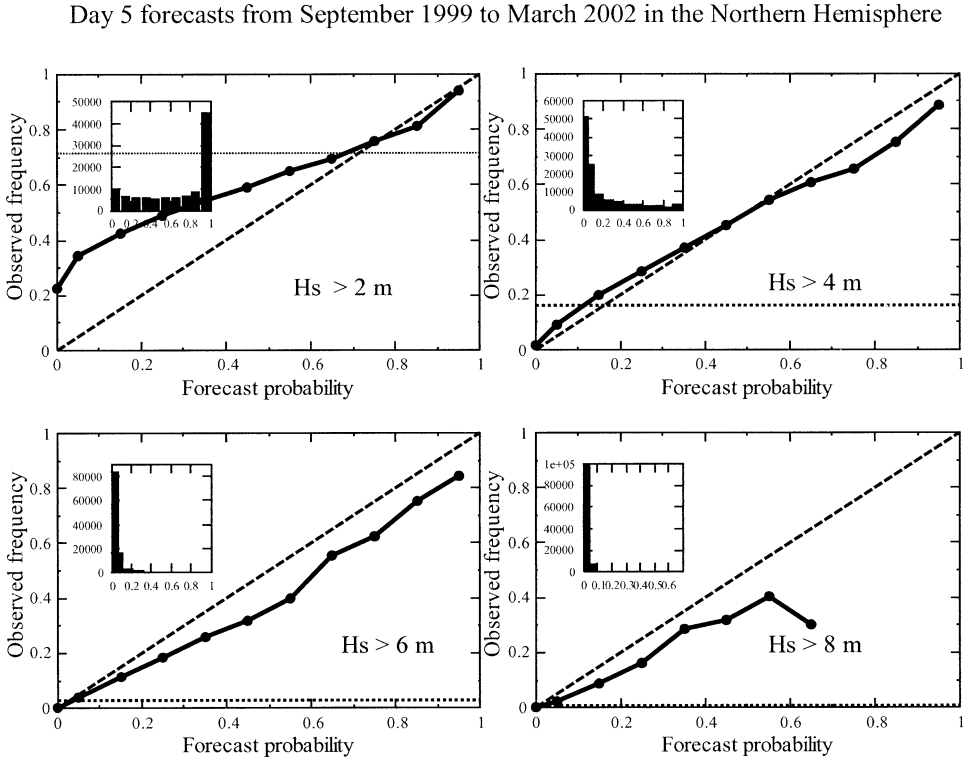

FIG. 15. Reliability diagrams for significant wave height at day 5 for the verification against altimeter data in the Northern Hemisphere. The threshold values are shown above each plot and are the same as those used for the buoy data (Fig. 8).

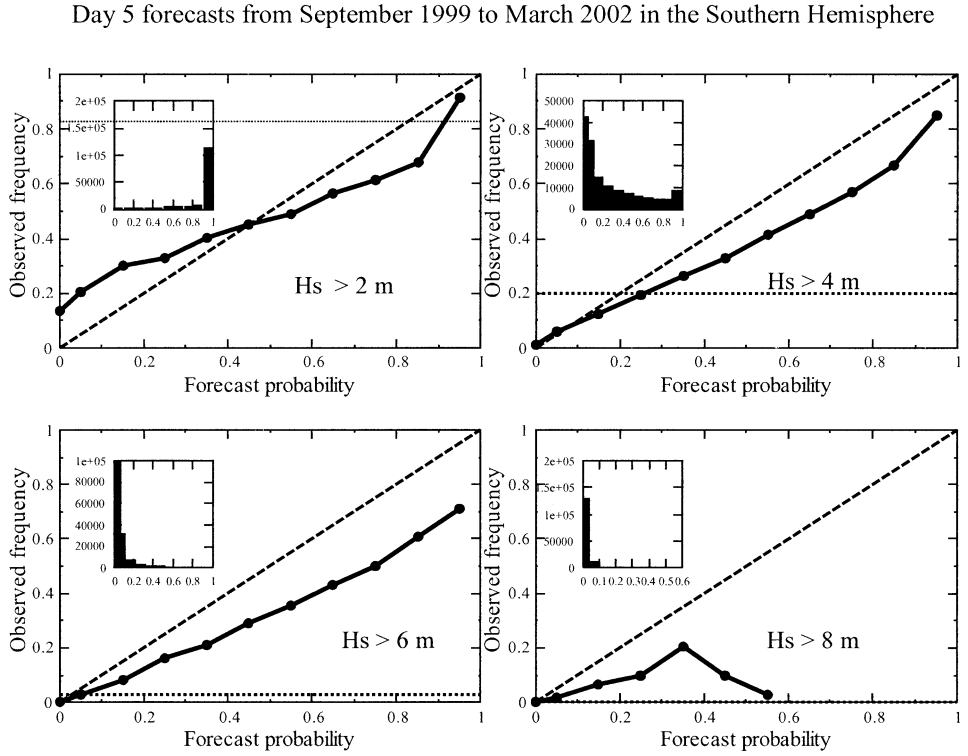

FIG. 16. Reliability diagrams for significant wave height at day 5 for the verification against altimeter data in the Southern Hemisphere.

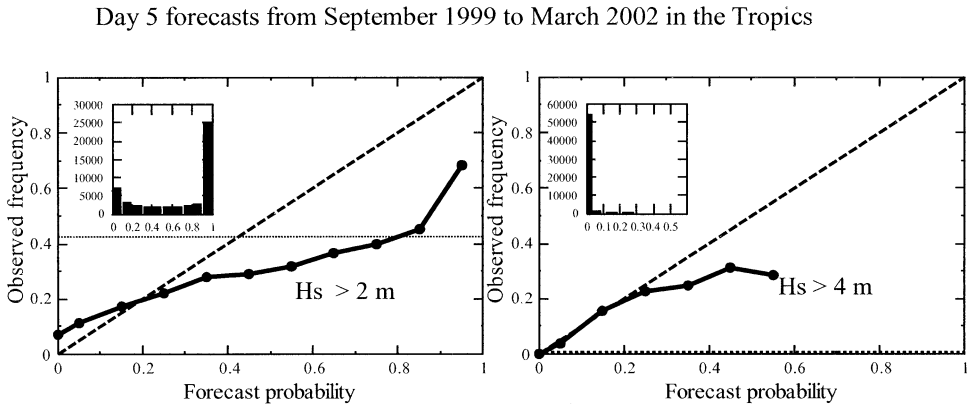

FIG. 17. Reliability diagrams for significant wave height at day 5 for the verification against altimeter data in the Tropics. For this area, there are almost no wave height observations above 6 m.

Fig. 17. For the Northern Hemisphere, the results are comparable to those obtained with buoy observations. Note however that for the 2-m threshold, the tendency to underforecast low probabilities is further enhanced. We believe this is caused by the fact that the *ERS-2* altimeter has problems measuring low wave heights (Janssen 2000). For the Southern Hemisphere, the EPS seems to have a small tendency to overforecast the probability of wave height of more than 4 m over most probability levels. In the Tropics, there are almost no observations of waves exceeding 6 m, and even for wave heights above 4 m, there are very few observations, making it difficult to draw any conclusions. For the lowest threshold level, the forecast system is overforecasting probabilities above approximately level 0.4.

The analyses of the EPS performance in terms of the relative economic value due to altimeter data are shown for waves above 2, 4, 6, and 8 m in Fig. 18. The solid line represents the relative values of the EPS, and the dotted line represents the deterministic forecasts, which is the control forecast in this case. At first glance, the results look very similar to those obtained from the buoy and platform data (see Fig. 12). However, a closer examination reveals some important differences. The maximum relative economic values are slightly reduced by roughly 0.1. For the lowest threshold value, the economic value does not exceed the climatological value for cost–loss rations below 0.2 when the altimeter data are used. Using the buoy and platform observations, this was achieved already at cost–loss ratios above 0.05. On the other hand, for cost–loss ratios above 0.5 and wave heights above 6 and 8 m, the economic values have improved compared to the results acquired from the GTS data.

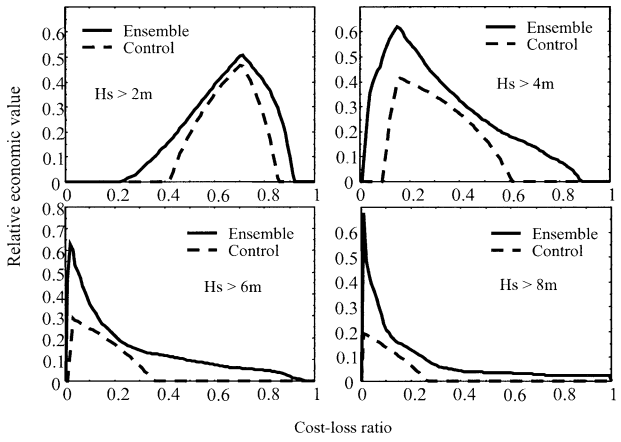

FIG. 18. Relative economic value for the day 5 wave height forecasts as a function of the cost–loss ratio. Results are based on the verification against altimeter data. The threshold levels are 2, 4, 6, and 8 m and correspond to the results in Fig. 12.

9. Conclusions

The ECMWF Ensemble Prediction System for waves and winds over marine areas has been compared with observed data for the period between September 1999 and March 2002. Two different datasets have been used. First, the model has been compared with quality controlled data from platforms and buoys, available via the Global Telecommunication System (GTS). This amounts to approximately 46 000 observations for wave heights, and about 60 000 observations for the wind speeds. Because of the limited geographical coverage of the GTS data sites, which are largely confined to the Northern Hemisphere and mostly located over the continental shelves, the EPS wave heights have in addition been evaluated using global satellite altimeter data. This dataset, covering the same time span as the GTS data, consists of 310 000 grid-box mean observations. The probabilistic forecasts have been tested for spread, spread–skill relation, reliability, and relative economic value.

In order to demonstrate that the spread can be regarded as a measure of the uncertainties in the deterministic forecasts, the ensemble spread was sorted into different bins, and the percentiles of the absolute errors were calculated for each bin. In this way a statistical

upper bound to the forecast error was found. For waves, the slope of the curve is more or less parallel to the diagonal line, indicating that the forecast error of the deterministic model could be expected to be bounded by the interquartile range of the ensemble spread. For the wind speed, the slope is less steep. Nonetheless, an apparent correlation between spread and skill is also demonstrated for this parameter.

The reliability of the probability forecasts has been tested by analyzing reliability diagrams for four different threshold levels for both wave height and wind speed. The reliability seems to be very good, although the buoy and platform observations indicate a tendency toward overconfidence in forecasting wave heights above 6 and 8 m. The reason for this is not clear to us at the moment. Generally, the ECMWF wave model is known to underestimate high waves (Bidlot et al. 2002), but when looking at individual time series for cases with very high waves, we can see that in most cases a number of the ensemble members have predicted wave heights that are well above the observed values. The reason seems to be that these members have been forced by sufficiently strong wind speeds, resulting in the forecasting of too large probabilities for the larger waves classes. For the two lowest threshold levels, the curves of observed frequency versus forecasted probability are very close to the diagonal line. To a certain degree, the altimeter observations confirm the above result, at least for the Northern Hemisphere. However, there is a more pronounced tendency for overconfidence in the probability forecast when tested against altimeter data.

In this study, the reliability of four different combinations of wave height and wave period has also been calculated. Two of these cases are considered to be rare, although indeed possible combinations of height and period. The reliability as it turns out is rather good. Even for the single most atypical combination, the points in the reliability diagram are located relatively close to the diagonal. However, this result must be interpreted with caution since the number of data pairs for this case is very low indeed. Also, when the Brier skill score is considered, this case is beaten by climatology at all forecast ranges.

To test the value of the EPS forecasting system for decision making, the method suggested by Richardson (2000) for calculating the relative economic value as a function of the cost–loss ratio has been applied. The value of the forecast is measured relative to that of climatology and perfect knowledge of the future weather. The method also enables comparison with other forecasting methods. The value of the forecasts has been compared with those of traditional deterministic forecasts. A poor-man's ensemble was created by simply adding normally distributed noise to the deterministic forecast, using information on the error statistics to determine the spread. The spread from such a forecast will be constant for a given forecast range and consequently cannot be used to decide the expected confidence in the deterministic forecast. It still performs relatively well as we have demonstrated in this investigation, even though it is in almost all situations outperformed by the real ensemble. For more complex forecasting parameters, the benefit of using the real ensemble becomes even more apparent. This encouraging result should hopefully serve as an inspiration for the development of more interesting products based on the EPS. The potential of the wave ensembles as a marine forecasting tool could then be exploited to its full extent.

Acknowledgments. The comments from the three anonymous reviewers helped to greatly improve the analysis and quality of this paper. This research is partly funded by the European Commission through the SEAROUTES project, Contract GRD-CT_2000-00309. We thank Peter Janssen, Hans Hersbach, David Richardson, Roberto Buizza, and Tim Palmer for support and valuable discussions.

APPENDIX

Definition of Brier Score and Brier Skill Score

Given N forecast and observation pairs, the Brier score (BS) is defined as

$$\mathrm{BS} = \frac{1}{N}\sum_{k=1}^{N}(y_k - o_k)^2,$$

where y_k is the forecasted probability, $o_k = 1$ when the event occurs, and $o_k = 0$ when the event does not occur. The Brier skill score (BSS) relative to climatology is calculated as

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{clim}}},$$

where $\mathrm{BS}_{\mathrm{clim}}$ is the Brier score of the climatological data. See Wilks (1995) for more details.

REFERENCES

Bidlot, J. R., D. J. Holmes, P. A. Wittmann, R. L. Lalbeharry, and H. S. Chen, 2002: Intercomparison of the performance of operational ocean wave forecasting systems with buoy data. *Wea. Forecasting,* **17,** 287–310.

Buizza, R., J. Barkmeijer, T. M. Palmer, and D. S. Richardson, 2000: Current status and future developments of the ECMWF Ensemble Prediction System. *Meteor. Appl.,* **7,** 163–175.

——, D. S. Richardson, and T. N. Palmer, 2003: Benefits of increased resolution in the ECMWF ensemble system and comparison with poor-man's ensembles. *Quart. J. Roy. Meteor. Soc.,* **129,** 1269–1288.

Challenor, P. G., and P. D. Cotton, 1997: The SOC contribution to the ESA working group calibration and validation of ERS-2 FD measurements of significant wave height and wind speed. *Proc. CEOS Wind and Wave Validation Workshop,* Noorwijk, Netherlands, ESTEC, ESA WPP-147, 95–100.

Farina, L., 2002: On ensemble prediction of ocean waves. *Tellus,* **54A,** 148–158.

Hamill, T. M., 2001: Interpretation of rank histogram for verifying ensemble forecasts. *Mon. Wea. Rev.,* **129,** 550–660.

Haver, S., and T. M. Vestbøstad, 2001: Uværshelga utenfor Midt-Norge 10–11 november 2001. STATOIL Doc. PTT-KU-MA-024, 94 pp.

Janssen, P., 2000: Wave modeling and altimeter wave height data. *Satellite, Oceanography and Society*, D. Halpern, Ed., Elsevier Science, 35–56.

——, B. Hansen, and J.-R. Bidlot, 1997: Verification of the ECMWF forecasting system against buoy and altimeter data. *Wea. Forecasting,* **12,** 763–784.

——, J. D. Doyle, J. Bidlot, B. Hansen, L. Isaksen, and P. Viterbo, 2002: Impact and feedback of ocean waves on the atmosphere. *Atmosphere–Ocean Interactions,* N. Perrie, Ed., *Advances in Fluid Mechanics*, Vol. I, WIT Press, 155–197.

——, S. Abdalla, and H. Hersbach, 2003: Error estimation of buoy, satellite, and model wave height data. ECMWF Research Dept. Tech. Memo. 402, 17 pp.

Komen, G., J. L. Cavaleri, M. Donelan, K. Hasselmann, S. Hasselmann, and P. A. E. M. Janssen, Eds., 1994: *Dynamics and Modelling of Ocean Waves.* Cambridge University Press, 533 pp.

Monaldo, F., 1988: Expected difference between buoy and radar altimeter estimates of wind speed and significant wave height and their implications on buoy–altimeter comparisons. *J. Geophys. Res.,* **93C,** 2285–2302.

Reistad, M., A. K. Magnusson, and D. Kvamme, 2003: Extreme waves at Haltenbnken: Wind and waves in extreme weather scenarios. Research Rep. 156, The Norwegian Meteorological Institute, 49 pp.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF Ensemble Prediction System. *Quart. J. Roy. Meteor. Soc.,* **126,** 649–667.

Saetra, Ø., and J. R. Bidlot, 2002: Assessment of the ECMWF Ensemble Prediction System for waves and marine winds. ECMWF Research Dept. Tech. Memo. 388, 29 pp.

——, ——, H. Hersbach, and D. S. Richardson, 2002: Effects of observation errors on the ensemble statistics. ECMWF Research Dept. Tech. Memo. 397, 12 pp.

Vogelezang, D. H. P., and C. J. Kok, 1999: Golfhoogteverwachtingen voor de Zuidelijke Noordzee. KNMI Tech. Rep. 223, 24 pp.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction.* Academic Press, 467 pp.