Assessments of Surface Winds and Waves from the NCEP Ensemble Forecast System

RICARDO MARTINS CAMPOS  
Department of Atmospheric and Oceanic Science, University of Maryland, College Park, College Park, Maryland

JOSE-HENRIQUE G. M. ALVES  
NOAA/NCEP/EMC/SRG/Center for Weather and Climate Prediction, College Park, Maryland

STEPHEN G. PENNY  
Department of Atmospheric and Oceanic Science, University of Maryland, College Park, College Park, Maryland

VLADIMIR KRASNOPOLSKY  
NOAA/NCEP/EMC/Center for Weather and Climate Prediction, College Park, Maryland

(Manuscript received 24 May 2018, in final form 21 August 2018)

ABSTRACT

The error characteristics of surface waves and winds produced by ensemble forecasts issued by the National Centers for Environmental Prediction are analyzed as a function of forecast range and severity. Eight error metrics are compared, separating the scatter component of the error from the systematic bias. Ensemble forecasts of extreme winds and extreme waves are compared to deterministic forecasts for long lead times, up to 10 days. A total of 29 metocean buoys is used to assess 1 year of forecasts (2016). The Global Wave Ensemble Forecast System (GWES) performs 10-day forecasts four times per day, with a spatial resolution of 0.5° and a temporal resolution of 3 h, using a 20-member ensemble plus a control member (deterministic) forecast. The largest errors in GWES, beyond forecast day 3, are found to be associated with winds above 14 m s⁻¹ and waves above 5 m. Extreme percentiles after the day-8 forecast reach 30% of underestimation for both 10-m-height wind (U10) and significant wave height (Hs). The comparison of probabilistic wave forecasts with deterministic runs shows an impressive improvement of predictability on the scatter component of the errors. The error for surface winds drops from 5 m s⁻¹ in the deterministic runs, associated with extreme events at longer forecast ranges, to values around 3 m s⁻¹ using the ensemble approach. As a result, GWES waves are better predicted, with a reduction in error from 2 m to less than 1.5 m for Hs. Nevertheless, under extreme conditions, critical systematic and scatter errors are identified beyond the day-6 and day-3 forecasts, respectively.

1. Introduction

Accurate wave forecasts are important for monitoring waves that threaten ships either at sea or at harbor, as well as offshore and coastal structures. When dealing with structures and wave energy converters, model skill becomes very important due to the quadratic relation of the wave energy in relation to the wave height, according to linear wave theory (Airy 1841). For a monochromatic wave, the total energy is integrated over the wavelength, which is directly proportional to the wave period. Using the spectral representation, skillful predictions of total energy rely on accurately estimating wave periods and the square of the significant wave heights. Furthermore, as discussed by Cavalieri et al. (2007), successful wave modeling strongly depends on the quality of input winds due to the quadratic dependence of the input source term to the wind speed.

The skill of wind and wave simulations inevitably decreases with forecast time due to the chaotic behavior of the atmosphere and ocean surface, which is discussed by Lorenz (1963) and Ghorbani et al. (2017). Since the 1990s, operational weather prediction has benefited from using ensemble forecasting approaches, where several numerical model integrations are performed simultaneously with perturbations applied either to the initial
conditions and/or to the model parameters (Kalnay 2003). Two advantages of using ensemble forecasts are 1) the average of the ensemble members tends to smooth out uncertain components, which leads to better skill than a single deterministic forecast, and 2) the spread of the ensemble members provides forecasters with an estimation of the uncertainty of the forecast (Kalnay 2003). The mean of the ensemble forecast is well known to be more accurate than any deterministic forecast after the first few days (Zhou et al. 2017). Another aspect of ensemble forecasts, especially interesting for forecasters in operational agencies, is that it can offer alternative scenarios that a single deterministic system cannot provide.

We perform an assessment of wind and wave forecasts produced by the National Centers for Environmental Prediction (NCEP), comparing the ensemble and deterministic forecasts. Special attention is given to higher percentiles compared to average conditions. The target variables are 10-m-height winds (U10), significant wave height (Hs), and peak period (Tp). We evaluate the ability of ensembles to predict extreme winds and waves well in advance by examining the joint distribution of the error as a function of forecast range and severity (percentiles). In section 2, we describe the NCEP ensemble prediction systems GEFS and GWES. Section 3 presents buoy data and evaluation methods, whereas section 4 provides the assessment results. Finally, section 5 contains the final discussion. The first results are presented based on the direct deviation from the forecast models to the buoy data. However, as the impact of observation errors should not be overlooked, the final results in section 5 present the metrics calculated considering the buoy measurement errors.

2. The Global Wave Ensemble Forecast System

The NCEP Global atmospheric Ensemble Forecast System (GEFS) was implemented in 1992. The Global Wave Ensemble System (GWES) was implemented in 2005 (Chen 2006), and validated by Cao et al. (2007) and Alves et al. (2013). In the initial GWES implementation, wave generation and decay were estimated by the source terms proposed by Tolman and Chalikov (1996), nonlinear wave–wave interactions were calculated using the discrete interactions approximations (DIAs) of Hasselmann and Hasselmann (1985), and propagation was computed using a third-order-accurate scheme (Leonard 1991). Alves et al. (2013) selected a 2-yr-long database (April 2010–March 2012) of along-track altimeter measurements of Hs made by Jason-1, Jason-2, and Envisat. Their results indicate that although the general bias of the ensemble system does not show significant improvement over the deterministic global wave forecasts, after the fifth forecast day, root-mean-square errors (RMSEs) from the GWES become smaller than the deterministic forecast. Furthermore, the GWES continuous ranked probability scores (CRPSs) systematically outperformed the corresponding deterministic model’s mean absolute error (MAE) at all forecast times.

The upgrades to the GWES closely follow the upgrades of the GEFS. The current implementation of GEFS had its last major upgrade in May 2015 when the ensemble initialization scheme, the breeding-based ensemble transformation with rescaling (ETR) in the operational GEFS, was replaced by the ensemble Kalman filter (EnKF). The horizontal resolution of GEFS has increased from Eulerian T254 (~52 km) for the first eight days of the forecast and T190 (~70 km) for the second eight days to semi-Lagrangian T574 (~34 km) and T384 (~52 km). The number of sigma pressure hybrid vertical layers increased from 42 to 64. The ensemble size remains the same (20 members and one control run) due to a limitation of computer resources. The initial perturbations are now drawn from the EnKF 6-h forecast ensemble instead of the breeding cycle. Zhou et al. (2017) describe and evaluate the most recent implementation of GEFS.

At present, the GWES runs a 10-day forecast four times per day, using a spatial resolution of 0.5° and intake of atmospheric forcing every 3 h. A total of 20 perturbed members plus a control member (deterministic run) compose the GWES, which is forced by GEFS winds on the WAVEWATCH III model (Tolman 2016), version 5.16. Therefore, the errors of the GWES wave fields (Hs and Tp) depend on the skill of the atmospheric model. The current implementation of GWES had its last major upgrade in December 2015 and differs from the configuration described in Alves et al. (2013) in relation to the input and dissipation source terms. Since version 4.18 was implemented, the WAVEWATCH III configuration of GWES replaced the Tolman and Chalikov (1996) source terms with those proposed by Ardhuin et al. (2010), which reflect more closely the current state-of-the-art in terms of wind input and wave dissipation physics parameterizations. It has the wind input source term adapted from Janssen (1991), with adjustments performed by Bidlot et al. (2005, 2007), as well as new forms for several dissipation processes relevant to wave evolution. Ardhuin et al. (2010) and Alves et al. (2014) show the differences between the last and the current package, and they calculate and discuss the improvements of the new source terms.

The period of GWES and observations selected for the present study corresponds to the year of 2016, without major GEFS or GWES upgrades (homogeneous
database), and it comprises a complete seasonal cycle. Although the GWES has its 10-day forecast with the same resolution, the GEFS maintains high resolution only for the first 8 days. For the 9th and 10th forecast days, GWES interpolates the surface winds with coarser resolution from GEFS. The use of low-resolution wind forcing provides a significant limit to the forecast skill for the GWES, which may affect the wave forecast under certain metocean conditions.

The assessments of Cao et al. (2007) and Alves et al. (2013) were based on bulk error metrics, joining calm and extreme events in the same discussion. A few studies have evaluated the skill of the models for different percentiles, for example, Stopa and Cheung (2014), Campos and Guedes Soares (2016a,b), Campos and Guedes Soares (2017), and Campos et al. (2018) for wave hindcasts, and Campos et al. (2017) for wave forecasts. There has not yet been an assessment of GWES analyzing the range of calm to extreme wind/wave conditions. Additionally, there has not yet been a published study addressing the joint distribution of wave ensemble errors as a function of forecast time and severity. These are addressed in the next sections.

3. Buoy data and evaluation method

Wave simulations are typically validated using altimeter and buoy data. However, at any given geographic site, altimeter data often suffer from coarse temporal sampling. As discussed by Alves and Young (2004), this is a particular concern when evaluating extreme events. Polar-orbiting satellites revisit a site once every 10–35 days with tracks typically separated by 100–200 km (Cooper and Forristall 1997). While the spatial coverage is sparse, as in situ instruments, buoys record regular hourly measurements that are able to capture the evolution of wave peaks generated by storms at sea. For this reason, only buoy data are considered presently when we evaluate GWES.

The quality controlled buoy data were obtained from the National Data Buoy Center (NDBC). Their standard meteorological format (stdmet) covers all variables of interest to the present study: Hs, Tp, and U10. Only metocean buoys with at least these three parameters were selected, as well as instants with all variables U10, Hs, and Tp qualified, in order to ensure the same data length for all variables. The surface wind speed from GWES is fixed at 10 m, while NDBC buoys have anemometers from 2.7 to 5 m. The wind profile power law was applied to convert the wind speed from the anemometer height to the 10-m level (Det Norske Veritas 2007):

\[ u = \left( \frac{z}{z_r} \right)^a, \]

where \( u \) is the wind speed (m s\(^{-1}\)) at height \( z \) (m), and \( u_r \) is the known wind speed at reference height \( z_r \). The constant \( a \) is the friction coefficient, a function of the topography at a specific site and usually assumed equal to 1/5 for open land. According to Hsu et al. (1994) the value of 0.10 is more appropriated for lakes and oceans, which was applied during the conversion.

To avoid any misinterpretation of the GWES error, only deep water buoys have been selected, which leads to 29 NDBC buoys with the following ID numbers: 41010, 41040, 41041, 41043, 41048, 41049, 42001, 42002, 42003, 42039, 42055, 42056, 42057, 42058, 42059, 42060, 42360, 46001, 46002, 46028, 46035, 46047, 46066, 46070, 46085, 46089, 51000, 51003, and 51004. Figure 1 shows the positions of the buoys where the GWES assessment was made. It restricts the representativeness of GWES error in the present study to the western portions of the North Atlantic Ocean and the eastern portions of the North Pacific Ocean.

The error metrics were selected based on the study of Mentaschi et al. (2013), who discuss the limitation of RMSE (Fig. 3 of their work) and the advantages of interpreting the systematic and scatter components (SC) of the error separately. In addition, they suggest the calculation of an additional metric (HH) developed by Hanna and Heinold (1985), which is a corrected indicator that overcomes the problem of low values of RMSE and SI not being always associated with better performances of the numerical models. Therefore, eight metrics were calculated to evaluate GWES, following the description of Mentaschi et al. (2013), where \( x \) is the observation (buoy data) and \( y \) is the forecast. In the following equations, the overbar indicates the arithmetic mean:

\[ \text{Bias} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{x}_i), \]  
\[ \text{NBias} = \frac{\sum_{i=1}^{n} (y_i - x_i)}{\sum_{i=1}^{n} x_i}, \]  
\[ \text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{x}_i)^2}, \]  
\[ \text{NRMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - x_i)^2}{\sum_{i=1}^{n} x_i^2}}. \]
The error is computed as a function of the percentiles by resampling the data moving a minimum percentile level from 0 to 99.9 and calculating the metrics, with several iterations that generate error-metric values and their curves. Each percentile (always from the buoy data) leads to a quantile that is considered to be a threshold to resample the whole dataset, and is plotted together with percentiles. The metrics are initially calculated without considering the observation errors and, in section 5, an estimation of measurement errors is used to recompute the metrics to obtain the final assessment results. Saetra et al. (2004) showed that in the short range, ensemble assessment results are quite sensitive to observation errors and that there is even a nonnegligible effect in the medium range.

4. Assessment results

Initial results are first presented in section 4a for one buoy in the North Atlantic Ocean, to provide an illustration of some interesting features of GWES errors. Next, the data from 29 buoys are combined to increase the data size and improve the statistical confidence for the analysis as a function of forecast time (section 4b) and the percentiles (section 4c).

a. Initial analysis at specific locations

Figure 2 presents the dynamic bias (i.e., GWES minus the observations) showing all the 10-day forecasts for each simulation. The x axis contains 1 year of GWES forecasts (2016), while the y axis shows the forecast time of each model step. Therefore, one could fix a position on the x axis and draw a vertical line to see the bias of that specific cycle throughout the forecast days. The x and y axes both represent time but not on the same scale; the x axis corresponds to 1 year while the y axis corresponds to 10 days (forecast time), which was stretched to facilitate the visualization. With a uniform aspect ratio, the data in Fig. 2 would appear as a 45° diagonal. Warm colors indicate overestimation of the numerical forecast while cold colors show underestimation, as related to the buoy. An unbiased simulation would show as white in the plot. Each diagonal colored stripe represents an event that was predicted by the GWES forecast 10 days prior and evolved forward. In most cases these forecasts approximate the nowcast.

The bias of Fig. 2 corresponds to NDBC buoy 41048 at 31.860°N and 69.590°W. While forecast errors increase with lead time, the error alternates between over- and underestimation, depending on the event. Lorenz (1963) explained the chaotic behavior of the atmosphere and the expected increase of forecast errors at longer forecast ranges, which is confirmed by increased negative and positive biases after forecast day 6 (y axis). The misrepresentation of events is reduced as forecast time
decreases, especially for U10 and Hs. For example, during Hurricane Matthew, which occurred in the beginning of October 2016, U10 and Hs were severely underpredicted by GWES until 4 days prior to the event. This is shown in Fig. 2, indicating the overestimation and underestimation of the model (illustrated by the red portion followed by blue). Many other events produce the same pattern, with the improvement in forecast skill approaching the nowcast.

This improvement occurs because the data assimilation system of the atmospheric forecast (GEFS), the ensemble Kalman filter (EnKF) described by Zhou et al. (2017), continuously pushes the simulated state toward the assimilated observations at each forecast cycle that is built four times per day at NCEP. The improved conditions using observations ensure an analysis with high skill, which is propagated through the forecast days. As the forecasted atmospheric conditions deviate from the analysis, the influence of assimilated measurements decays in time and, consequently, the errors increase. Figure 2 indicates that this process can be very complex and depends on the meteorological system, severity, and season.

Another feature is that the variables U10, Hs, and Tp are highly correlated, especially Hs with U10. The wind underestimation of some events generates underestimated wave heights, with the same dependence valid for overestimation. This behavior is expected, since NCEP does not have a wave data assimilation system as of the date of this publication and, therefore, the errors on the wave fields are highly dependent of the wind inputs. The peak period, however, presents a different pattern of behavior when compared to U10 and Hs. The Tp errors have a weaker dependence on forecast time, so that events are generally either consistently misrepresented or well predicted throughout the entire forecast range. This could suggest a different source of error for Tp, probably associated with the parameterization and calibration of the WAVEWATCH III model. For this reason, events with large errors in Tp tend not to be significantly improved at lower forecast ranges.

A total of 29 other plots similar to Fig. 2 and associated with each buoy (not shown) were examined. They confirmed the same dependence of errors to the forecast time and events but with small differences associated with distinct wave climates, as expected. More influence of extratropical storms was found by the Pacific buoys while the Atlantic buoys, especially those farther south, respond more to tropical cyclones.

b. GWES skill as a function of forecast time

From this section on, all data from the 29 NDBC buoys are bundled to form a single dataset. Figure 3...
presents the error metrics (y axis) as a function of forecast time (x axis). Once again, a negative bias indicates underestimation while a positive bias indicates overestimation, as described by Eq. (2). The bias of the ensemble mean (cyan) is always contained within the dispersion of the ensemble members (top row in Fig. 3). The spread of the members increases with time, as do the errors. The SCrmse and CC metrics, however, show the ensemble mean with much lower error, indicating the benefit of using an ensemble compared to a deterministic forecast. The HH and SCrmse metrics also indicate improved skill using the ensemble mean relative to the deterministic forecast. The positive impact increases with time, particularly after forecast day 4, similar to results reported by Alves et al. (2013). GWES leads to SCrmse of Hs related to day 10 using the ensemble mean at a level equivalent to the deterministic run of day 7, for example, a gain of 3 days in the forecast skill.

In the top row of Fig. 3, the nonhomogeneity of GEFS surface wind accuracy, with strong negative bias in the...
analysis, is improved with forecast time. A direct impact is observed on the wave bias, where Hs deteriorates with time. Therefore, the wave model shows reduced bias in the short range, which increases with forecast time as the GEFS winds become stronger, despite having a lower bias. The consistency here is with the bias trend, which indicates weaker surface winds at shorter ranges, and stronger winds at longer forecast times.

c. GWES skill as a function of percentiles

Figure 4 presents the error as a function of severity, for three different forecast ranges: a nowcast, a 5-day forecast, and a 10-day forecast. Different than Fig. 3, the x axis in Fig. 4 now shows the percentiles and associated increasing quantiles; the different forecast times are now plotted as curves with distinct colors. Figure 4 shows that errors tend to increase toward extreme conditions, especially the bias. The rate of deterioration with severity is larger for longer forecast ranges, mainly for U10 followed by Hs. The patterns of evolution of Hs biases for the nowcast and 5-day forecast are similar, with deterioration at longer leads at day 10. This feature is similar to the bias of Tp; from day 5 to shorter forecast ranges the accuracy of Hs and Tp with higher percentiles does not improve significantly.

The SCrmse values for U10 and Hs converge to similar errors for forecast days 5 and 10 at higher percentiles, which means that large errors of extreme events are nearly the same for the mid- and long forecast ranges, in terms of the scatter component of the error. This differs from the systematic error (i.e., bias) discussed before, where the nowcast and day 5 were grouped. Therefore, the deterioration of forecasts, in terms of the scatter component of the error, is strongly dependent on the percentile. To illustrate, a simple comparison would be to fix two x-axis values, for example the 20th and 80th percentiles, and compare the gap in between different forecast ranges. The reduction of forecast errors is much more significant for calm conditions than during severe weather. This is valid for the scatter components of the errors of U10 and Hs but not for Tp, which, once again, showed a different pattern with nearly constant scatter errors with percentile levels. Comparing the bias with SCrmse of Tp, it is possible to visualize that the errors
during the wave period are different from those for the wind speed and wave height, which presents a stronger systematic component. The metric HH, not presented here, confirms the behavior found in the SCrmse plots.

d. Combined error as a function of forecast time and percentile levels

The combination of Figs. 3 and 4 provides the error as a function of two variables: the forecast range (y axis in Figs. 5–8) and the severity (quantiles; x axes in Figs. 5–8). Therefore, the x axes in Figs. 5–8 are the same as in Fig. 4, with the forecast time included along the y axis and errors plotted as shaded colors, following the standard used in previous figures: warm colors indicate overestimation of the numerical forecasts related to the buoys while cold colors indicate underestimation. The error generally increases as a function of both forecast range and severity, with different aspects depending on the variable and metrics addressed. This section also continues the comparisons between deterministic and ensemble forecasts.

Figure 5 shows the bias [Eq. (3)], where U10 and Hs have very similar distributions. The errors of the analysis/nowcast as well as for other forecast ranges under calm conditions tend to zero. Moving to higher percentiles and forecast times, the bias becomes negative with large underestimation of GWES for both wind speeds and wave heights, which confirms the first line of plots in Fig. 4. The interval of normalized biases is similar for U10 and Hs, with values from 0 to −0.3; that is, there is a 30% underestimation for extreme percentiles after the day-8 forecast. There is a slightly greater dependence of Hs on the severity than the forecast time, when compared to the same pattern for U10. As discussed before, Tp presents different error distributions, with the bias decreasing with higher values of Tp. Thus, the wave model tends to overestimate small wave periods and underestimate large ones. This behavior is amplified at longer forecast ranges.

The top row in Fig. 5 shows the deterministic run while the bottom row has metrics calculated for the GWES ensemble mean. Zhou et al. (2017) explain how the perturbations are introduced in the atmospheric simulations of GEFS using the EnKF, which assumes a Gaussian distribution of the model errors. As Fig. 5 addresses the systematic error only, when the ensemble spread is added to the deterministic run, the geometric mean remains centered, since the Gaussian distribution has skewness equal to zero. Hence, the bias of U10 for the deterministic run is equal to the ensemble mean, as expected, since the ensemble approach is not meant to correct the bias but to reduce the scatter errors. However, the EnKF is applied to the atmospheric model only, which generates U10 that is introduced as input into the wave model, producing Hs and Tp. During this process, each ensemble member is run independently, and the final wave ensemble can present a small skewness, leading to a displaced arithmetic average. This can be visualized by comparing the top and bottom plots, of Hs and Tp, in Fig. 5, which show small differences.

Figure 6 compares the scatter component of the error [Eq. (6)] of the control member with the error of the ensemble mean. Unlike the systematic error [Eqs. (2) and (3)], the error of the ensemble mean is different than the mean of the error of the ensemble members; the latter is an inappropriate performance measure. The SCrmse in Fig. 6 presents an impressive reduction of the error of the ensemble mean compared to the control member (deterministic), showing the success of the method implemented by Zhou et al. (2017). The SCrmse values of 5 m s−1 in U10 associated with extreme events at longer forecast ranges were improved to values of around 3 m s−1. As a direct impact, the wave forecasts of GWES produce better skill, dropping from 2 m of SCrmse for Hs to less than 1.5 m. The same is valid for the wave periods, showing the benefit of using an atmospheric ensemble to produce more accurate extreme waves at longer forecast ranges. Figure 6 highlights where the largest errors, which are associated with winds and waves above 14 m s−1 and 5 m, are concentrated beyond forecast day 3. There is an interesting boundary delimiting well-predicted extreme events up to day 2. The results show that NCEP forecasts with good precision under extreme conditions are restricted to horizons of only 2 or 3 days.

The error metric HH proposed by Hanna and Heinold (1985), from Eq. (8), is plotted in Fig. 7 to confirm what has been discussed so far. Unlike Fig. 6, the limit where GWES prediction skill significantly deteriorates occurs at the longer forecast ranges, beyond day 5. Moreover, the HH error metric associated with Hs appears to be more dependent on the forecast time than on the quantiles, when compared to SCrmse in Fig. 6. Once again, the error distribution of U10 and Hs are similar. The HH of Tp is again different from the results for U10 and Hs, with large errors for very low and very high quantiles, and better results between 8 and 15 s. Comparing the errors of the deterministic run with the ensemble (top and bottom panels in Fig. 7), it is possible to conclude that the improvement of the ensemble methodology is also confirmed by the HH metric.

From Eq. (8) we see that HH considers the absolute differences between GWES and the observations, through the squared differences, divided by the product of them, which amplifies the absolute differences when the model diverges from the measurements. The result is a metric that better addresses the differences under extreme conditions, as can be observed in Fig. 7.
5. Final discussion

Characteristics of the NCEP’s ensemble products from the GEFS and the GWES were analyzed based on a number of metrics. In terms of the bias, we found a non-homogeneity of GEFS surface wind skill, containing strong negative bias in the U10 analysis that improves with forecast time (see the top-left panel in Fig. 3). This wind bias has a direct impact on the wave fields, as expected, with a similar shape but displaced to higher values of the positive bias. In this case, Hs is in better agreement with the data in the short forecast range (0–24 h) than in the longer ranges. Therefore, in terms of bias, GEFS winds improve with time (reducing the underestimation) while GWES waves deteriorate with time (increasing the overestimation).

A possible explanation is that the WAVEWATCH III parameterizations were originally tuned with respect to the analysis winds (forecast day 0). The wave model calibration displaces the overall bias but cannot change the bias trend of Hs, which is derived directly from the wind input. Consequently, for the wholesale improvement of wave model product skill scores, considering that the surface winds’ inhomogeneity is given, it may be necessary to develop an approach that corrects the error as a function of the forecast range. This fact in itself explains the observed wave-height biases and may be related to the systematic biases observed during peak periods, particularly in severe sea states, given the cumulative nature of the wave development (e.g., if winds are systematically low, their cumulative effect on the waves will result in sources of error regardless of the wave model tuning attempting to compensate for the low bias). Saetra et al. (2004) describes that ensemble systems are commonly verified with respect to verifying the analysis fields rather than with respect to the observations. For our specific case in Fig. 3, it would lead to a nonrealistic bias since the GEFS analysis of U10 underestimated the observations.

It is important to remember that these features involving the bias are common to deterministic and ensemble forecasts, simply because the ensemble methodology was not conceived to reduce the systematic errors but the scatter errors, which is why post-processing bias correction algorithms are important and widely used by operational centers like NCEP. The success of the NCEP ensemble systems in reducing the scatter errors can be verified by examining the red curves in Fig. 3, which are related to the ensemble members (black) in the bias plots, and below the black and cyan curves (deterministic run) in the SCrmse plots.

The discussion and analysis above was possible due to Mentaschi et al. (2013), who described error metrics involved with dividing the systematic component from the scatter component of the total error. However, the
assessments so far have been calculated based on Eqs. (2)–(9) without considering the observation errors of \( x \), which are related to the buoy measurements. Saetra et al. (2004) studied the effects of observation errors on ensemble assessment using a perfect model approach. Saetra et al. found that, when comparing the wave ensembles with buoy measurements, the total number of outliers for the day-3 forecasts is reduced from more than 25% to less than 10% when a reasonable estimate of the observation errors is taken into account. Bowler (2006) showed that given the representation of the measurement errors, it is possible to remove the effect of those errors from the assessment scores, which improves the apparent performance of a forecasting system. Bowler (2008) and Ciach and Krajewski (1999) describe that the \( \text{RMSE}_o \) obtained from Eq. (4) is actually composed of

\[
\text{RMSE}_o = \sqrt{\text{RMSE}_t^2 + \text{RMSE}_b^2}, \tag{10}
\]

where \( \text{RMSE}_t \) is the “true” RMSE of the forecasts if the true state of the ocean/atmosphere was known, and \( \text{RMSE}_b \) is the error of the buoy, as measured against the “truth.” The goal of our assessment is to properly calculate \( \text{RMSE}_t \); however, the exact \( \text{RMSE}_b \) is unknown. Bowler (2006) describes that the source of the observation errors changes dramatically depending on the type of observations. Moreover, it strongly depends on the type of accelerometer and buoy size/hull, the mooring system, and the environmental conditions. Liu et al. (2015) examined the errors from different types of buoys compared to accurate wave gauges in a tank. They found relative errors between 3.47% and 3.79% for \( H_s \) and between 1.87% and 3.05% for \( T_p \). Lawrence et al. (2012) investigated the accuracy of several wave buoys, including Wavescan, SeaWatch Mini II, Directional Waverider, and the TRIAXYS directional wave buoy. They pointed to accuracies better than 2% and errors of less than 5 cm. We should not neglect that part of the observation errors from the buoys is dynamic and depends on the drag forces on the buoy (Ashton and Johanning 2015), such as strong currents and breaking waves. Bender et al. (2010) points to increasing errors on \( H_s \) when the buoy is heeled over for long periods of time. Additional neglected sources might include errors induced by biofouling, tying up of small craft to the buoy (Thomas 2016), and, for the spectral analysis processing and zero-order moment, the sampling variability and number of degrees of freedom in the FFT (Donelan and Pierson 1983).

Therefore, it is impossible to perfectly track all the dynamic errors at each wave buoy in order to exactly calculate \( \text{RMSE}_b \) for each measurement. This is also not automatically computed and provided by NDBC (NDBC 2015).
On the other hand, using buoy data from the same agency, NDBC, ensures the same quality control and spectral estimation of the data processing, which increases the consistency of RMSE\textsubscript{b} among buoys. Our approximated estimation of RMSE\textsubscript{b} is based on the study of Abdalla et al. (2011), who developed a triple-collocation technique to calculate the errors of wind and wave products from altimeters and buoys. In addition to the bulk computation of such averaged observation errors, they properly calculated the buoy errors of U10 and Hs as a function of wind and wave severity, similar to Fig. 4. These two error functions for U10 and Hs, provided by Figs. 3 and 6 in Abdalla et al. (2011), indicate relative errors of 0.11–0.22 m s\(^{-1}\) for U10 and 0.06–0.09 m for Hs. They were applied to our data to estimate buoy errors and RMSE\textsubscript{b} as a function of percentiles, which was utilized to recompute the metrics.

Figure 8 presents the final bias and RMSE values calculated using the methodology of Bowler (2008) and Ciach and Krajewski (1999), removing the buoy errors estimated by Abdalla et al. (2011). Indeed, as stated by Bowler (2006), the apparent performance of GWES is slightly improved, where the highest biases of U10 dropped from 6 to 4 m s\(^{-1}\) under extreme conditions and over long forecast times, and RMSE decreased from 9 to 8 m s\(^{-1}\) at the top-right corner of Fig. 8. Consequently, the Hs under extreme conditions and over long forecast times dropped from 2 to 1.5 m while the RMSE was reduced by approximately 0.2 m under these conditions. Figure 8 also compares the control run with the ensemble mean (EM), for the RMSE only, since the systematic error of the deterministic run and EM are the same, as discussed before. We again see similar patterns in U10 and Hs, with larger RMSEs under extreme conditions beyond forecast day 6. The ensemble approach is confirmed to reduce the RMSE, but the improvement is not as evident as in Fig. 6 for the isolated scatter component of the GWES error.

Figure 8, together with Figs. 4–7, shows that forecast skill is significantly reduced under extreme conditions for all metrics analyzed. The level of deterioration in forecast skill depends on the forecast range and error metric considered. Under calm and moderate conditions, even long forecasts are quite precise and accurate, apart from Tp, which presented some problems (see bias and HH) at low periods. Under extreme conditions, critical systematic errors are found beyond forecast day 6 while scatter errors are found beyond day 3. By combining both error components of the RMSE, we can confirm that skilled GWES forecasts under extreme conditions are expected up to forecast day 3.

We gave more attention to the systematic and scatter components of the errors, using the metrics bias and SCrmse, because they can be finally combined into the...
RMSE in Fig. 8; however, the SI could have been used instead of SCrmse with the same effect of looking at the scatter errors [see Eqs. (6) and (7)]. Moreover, we found that the discussion of Mentaschi et al. (2013) suggesting the use of a metric developed by Hanna and Heinold (1985) to be extremely relevant. The metric HH was found to be interesting in that it better addresses the differences under extreme conditions, one of the main goals of the present paper, and helped to confirm some characteristics pointed out by other metrics; complementing the approach based on the separation of systematic and scatter errors.

6. Conclusions

The skill of NCEP’s operational GEFS and GWES, which generate wind and wave ensemble forecasts, were examined as a function of severity (percentiles) and forecast time. A progressive deterioration in forecast skill was identified for extreme events. Under calm and moderate conditions, all forecast ranges were quite precise and accurate, while under extreme conditions, critical systematic errors were found after day 6, and large scatter errors after day 3. One striking source of general improvement to the outputs from the GEFS and GWES forecast systems was a reduction in biases in the GEFS short-range forecasts, which are systematically underestimated relative to the observations and have a negative impact on the GWES’s accuracy. An improved and robust bias correction algorithm could address this problem. As part of the current work, a method for correcting biases and scattering errors based on applications of neural networks has been developed; initial results were presented by Campos et al. (2017).

In terms of the scatter component of the error, the rate of deterioration with severity is larger for longer forecast ranges, especially for U10 followed by Hs. The worst errors are observed for waves above 4 m, winds above 13 m s⁻¹, and beyond the fourth day of the forecast. Figures showing the scatter component of RMSE (SCrmse) reveal the success of the EnKF method implemented by Zhou et al. (2017) on the wind speeds (U10), which benefits the wave forecast by directly improving the wave heights (Hs). The SCrmse of Hs related to day 10 using the ensemble mean is equivalent to the deterministic run of day 7, which represents a gain in forecast skill of about 3 forecast days—an improvement of great value for weather alerts and safety management. These features were confirmed by the analysis that recomputed the error metrics after removing the observational errors (Fig. 8), following the methodology and results of Bowler (2008), Ciach and Krajewski (1999), and Abdalla et al. (2011).

Among the three variables studied in this paper, Tp presented different features from U10 and Hs. It was shown in Fig. 2 that the skill of Tp is not significantly
improved at lower forecast ranges, which was further confirmed by Fig. 5, where large errors occur for very low and very high quantiles, while better results are verified for Tp from 8 to 15 s. Therefore, the operational GWES configuration of WAVEWATCH III tends to overestimate small wave periods and underestimate large ones. This behavior is slightly amplified at longer forecast ranges. This lower skill of Tp is expected, since it maps a totally different feature of the wave distribution, the peak of the power spectrum, instead of the zero-order spectral moment associated with Hs. 

The spectral domain.

Young (1995) discusses the estimates of the spectral peak frequency and, considering our results regarding integral parameters, we recommend future studies joining our methodology with assessments focused on the spectral domain.

Acknowledgments. This work has been developed jointly at the University of Maryland, and at the Environmental Modeling Center of NCEP, with funding from the National Weather Service Office of Science and Technology (NWS/OST), Award NAA16NWS4680011. Authors would like to acknowledge Dr. Todd Spindler for the support with the data manipulation and coding, and the atmospheric ensemble team at NCEP.

REFERENCES


Cao, D., H. S. Chen, and H. L. Tolman, 2007: Verification of ocean wave ensemble forecasts at NCEP. Proc. 10th Int. Workshop on Wave Hindcasting and Forecasting/First Coastal Hazards Symp., Oahu, HI, Environment Canada, G1.


Hs u ,S .A . ,E .A .M e i n d l ,a n dD .B .G i l h o u s e n ,1 9 9 4 :D e t e r m i n i n gth e
power-law wind-profile exponent under near-neutral stability


