Global Assessments of the NCEP Ensemble Forecast System using Altimeter Data

Ricardo Martins Campos[1*], Jose-Henrique G.M. Alves[2], Stephen G. Penny[3,4], Vladimir Krasnopolsky[5]

[1]Centre for Marine Technology and Ocean Engineering (CENTEC), Instituto Superior Técnico, University of Lisbon

[2]SRG/EMC/NCEP / NOAA Center for Weather and Climate Prediction

[3]Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado Boulder

[4]Physical Sciences Division, NOAA Earth System Research Laboratory

[5]EMC/NCEP / NOAA Center for Weather and Climate Prediction

*Corresponding Author, e-mail address: riwave@gmail.com

ABSTRACT:

Forecasts of 10-m wind (U10) and significant wave height (Hs) from the National Centers for Environmental Prediction (NCEP) Ensemble Forecast System are evaluated using altimeter data. Four altimeter missions are selected for the assessment in 2017 that provide a total of 33,229,297 data points matching model state to altimeter measurement. This large quantity of data allows the investigation of the error as a function of forecast ranges, quantiles, and location. Special attention is given to the comparison between the arithmetic mean of the ensemble forecast and the deterministic forecast control run. Error metrics are selected to quantify and separate the systematic and scatter components of the error. Results indicate a large reduction of the scatter errors (SCrmse) in the ensemble mean compared to the control run; more evident for U10, where large SCrmse of 5 m/s associated with strong winds at mid-latitudes beyond forecast day 7 drops to 3 m/s for the ensemble mean. This benefit is transferred to Hs and the largest SCrmse of 1.8 m at the control run is reduced to 1.3 m for the ensemble mean. Although the overall forecast skill of the ensemble forecast is improved, the extreme quantiles of Hs and U10 beyond forecast day 5 tend to underestimate the observations. This implies a need for bias correction algorithms applied during post-processing of the NCEP ensemble products. We conclude that for reliable wind and wave forecasts beyond 7 days at mid and high latitudes, it is essential to use ensemble forecast products, especially when associated with extratropical areas in the Southern Hemisphere.

Keywords: model validation, ensemble forecasts, extreme winds, extreme waves, altimeter data.

1. Introduction

The demand for reliable global forecasts of surface winds and waves has rapidly increased worldwide. This demand has followed population growth in coastal cities, and growth in offshore industries such as renewable wind energy and offshore oil and gas. Ship traffic has increased 300% since 1992 and shows an average increasing rate of 10% per year according to Tournadre (2014). This sector, among others, requires accurate predictions at longer forecast ranges, since most ship journeys exceed 1 week in duration. A containership crossing the Atlantic Ocean for example, considering a range of sailing speeds (Psaraftis and Kontovas, 2014), takes from 1 week to 20 days to complete the journey. Higher-quality wind and wave forecasts are also an essential element in operational oceanography programs that have been established around the world (Le Traon et al., 2015).

The same need is valid for extreme weather forecasts, where a balance between time and accuracy is critical for issuing reliable alerts while allowing sufficient time to take safety actions. The use of ensemble forecasting approaches can extend model forecast skill to longer lead times, as discussed by Kalnay (2003). A usual approach to ensemble forecasting is to produce several numerical model integrations (members) simultaneously starting from perturbed initial conditions, which represent uncertainties in the initial model state. The arithmetic mean of the ensemble members has generally been proven to outperform deterministic simulations (i.e. a single control run). For the specific case of NCEP's wave ensembles, benefits are larger beyond the 4^{th} or 5^{th} forecast day (Campos et al., 2018a). The combination of ensemble forecasts from several centers and models have further provided evidence that by incorporating model uncertainties in probabilistic products there is a significant increase in predictability (Candille, 2009). Such results have been a great motivation for operational centers to invest in ensemble forecasts since the 1990s, and in the specific case of wave products since 1998 (Hoffschildt et al. 1999).

Our goal is to assess the NCEP Ensemble Forecast System, comparing the widely-used deterministic forecast with the ensemble approach. Although this was already attempted in previous studies, we expand those results by focusing on the spatial distribution of errors, in order to provide a global estimate of forecast skill for 10-m wind speed (U10) and significant wave height (Hs). We also extend the analysis using altimeter wave-height products from a constellation of four satellite missions, whereas previous studies were generally limited to using a smaller number of mission products. Therefore, our assessment exploits a large volume of data by using millions of pairs of model/satellite, which allows a multivariate analysis of the forecast errors and provide additional support for the construction of robust post-processing algorithms of bias corrections, such as Zieger et al. (2018), Harpham et al. (2016), Durrant et al. (2009), and new developments using machine learning techniques described by Boukabara et al. (2019).

The NCEP Global Wave Ensemble Forecast System (GWES; Chen, 2006; Alves et al, 2013) runs a 10-day forecast, four times per day, with space-time output resolution of 0.5° and 3 h. GWES contains 20 perturbed members plus a control member (deterministic run) of the WAVEWATCH III model (Tolman 2016), forced by the Global Ensemble Forecast System (GEFS) winds, and ice concentrations from the NCEP's automated ice analysis system (Grumbine, 1996). Zhou et al. (2017) provide a complete assessment of GEFS, while Cao et al. (2007), Alves et al. (2013), and Campos et al. (2018a) analyzed the wave products of GWES. These prior results indicate that after the 5th forecast day, the ensemble mean from a single model produces a reduced scatter component of the error compared to the traditional deterministic run.

In addition to the NCEP prediction system, Bidlot (2017) performed a review and assessment of wave forecasts from 16 operational centers, using 21 years of in-situ observations. The three wave forecasts with the best scatter indexes according to his study are the European Centre for Medium-Range Weather Forecasts (ECMWF), Météo France (METFR), and Service Hydrographique et Océanographique de la Marine (SHOM) – considering that METFR and SHOM both use winds from ECMWF. Besides, Bidlot (2017) discusses the evolution of wave forecast throughout time, highlighting the improvements over the last 10 years, with a reduction around 0.10 on the scatter indexes, depending on the in-situ station. Although the slightly better skill of ECMWF wave forecasts compared to NCEP according to Bidlot (2017), NCEP products have the advantage of being publicly available on global scale, with easy access, being widely used worldwide.

Bunney and Saulter (2015) analyzed the UK Met Office wave ensemble that is driven by hourly wind fields from MOGREPS (Bowler et al., 2008), quantifying the uncertainties in short range (up to 7 days) for the Atlantic Ocean and around the UK. The authors found virtually nil bias for the overall statistics at the whole Atlantic domain but reported regional biases present in the UK, which pose an impact on the verification of short range forecasts, with low spread. It highlights the importance of performing a spatial analysis of forecast errors, which is one of our main goals. Saetra and Bidlot (2004) studied the potential benefits of using an Ensemble Prediction System (EPS) for waves and marine surface winds, and concluded that ECMWF EPS over-performs the control (“deterministic”) forecasts, despite the small tendency for overconfidence in the wave probability forecasts for waves above 6 and 8 m (more pronounced in the Southern Hemisphere). Our evaluation provide direct comparisons between the ensemble mean with control run and ensemble members using several error metrics, in order to investigate the performance and differences among results.

2. Altimeter Data and Evaluation Method

The work of Campos et al. (2018a) provided a multivariate assessment of GWES using buoy data, studying the forecast error as a function of forecast days and severity. Smaller scatter errors were found in the arithmetic ensemble mean of GWES than in the deterministic forecast (control run), with a significant improvement of the predictability at longer forecast ranges. However, large errors were still present in GWES beyond forecast day 3, associated with winds above 14 m/s and waves above 5 m. Because the results of Campos et al. (2018a) are only representative of the specific buoy locations where error metrics were calculated, the present study aims at filling this gap by using altimeter data in the GWES assessment. Our present focus is on a single wave ensemble product from NCEP's GWES, which will be expanded in a future study to include combined wave products from multi-center ensemble systems such as those planned under the North Atlantic Ensemble Forecast System (NAEFS) framework (Alves et al., 2013) and multiple centers as addressed by Bidlot (2017).

Uncertainties in altimeter data have been investigated by Sepulveda et al. (2015) and Queffeulou and Croizé-Fillon (2017). They found the altimeter estimates of Hs are in agreement with buoys, containing standard deviations of the order of 0.3 m, depending on the satellite. The recent study of Ribal and Young (2019) provide a complete assessment for 13 altimeters covering 33 years of data, evaluated against buoy data from the National Oceanographic Data Center (NODC). The comparisons for U10 and Hs have been analyzed and, regarding the satellite missions selected in our present study, Ribal and Young (2019) found very small differences, limited to 0.5 m/s and 0.10 m respectively. Therefore, considering this level of uncertainty is much smaller than GWES errors, altimeter data can be directly applied to our forecast assessment, after a quick additional quality control.

The period of evaluation is 2017, when four satellite missions were selected from the AVISO and NESDIS databases: JASON 2, JASON 3, Saral, and Cryosat 2. Altimeter tracks were collocated into the regular GWES grid based on the methodology of Young and Holland (1996) and Sepulveda et al. (2015), where all satellite observations with a maximum space distance of 25 km and time distance of 0.5 hours are averaged and then allocated to each grid point (Lat/Lon) at a specific time. In fact, a Gaussian function is applied to weight altimeter records by distance to the center grid point, providing one altimeter value per Lat/Lon/Time matching the regular GWES grid of 0.5°X0.5°. We have decided to collocate the altimeter data into the GWES space and not the opposite for a number of reasons: (i) to include an average of 10 to 20 altimeter records to a single GWES value, which increases the statistical significance of observations and reduce the impact of rare, but still possible, outliers and spikes; (ii) the high resolution of satellite sampling captures time and space scales that are different from the 0.5°X0.5° model grid and would input a misleading comparison between datasets; (iii) to avoid several interpolations of GWES dataset to the satellite space and time; (iv) practical computational limitations involving the amount of data, which reduces the storage space and RAM memory use when collocating altimeter data into the GWES space.

Figure 1 shows the count of altimeter measurements at each grid point that are used for the GWES assessments. This represents a large increase in the observations available for the calculation of the error metrics when compared to buoy assessments presented in Campos et al. (2018a), which permits a study of the spatial distribution of the model skill and also increases the statistical relevance of the analyses. A total of 33,229,297 pairs of GWES model state estimates and altimeter measurements were compiled for the assessments detailed in the following sections.

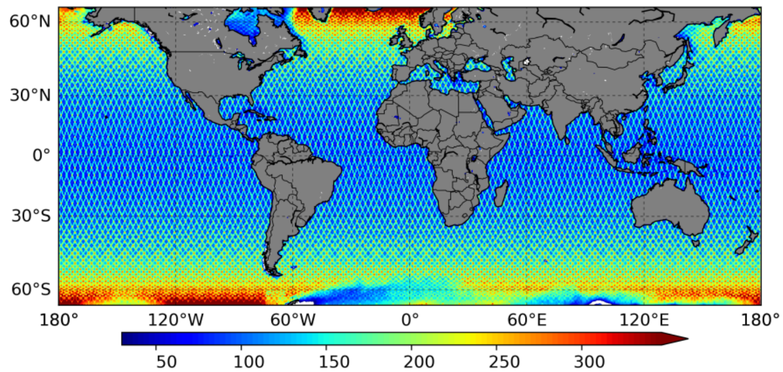

Figure 1 - Total count of altimeter measurements per GWES grid point for 2017.

Pairing buoy data with hindcast model states is reasonably simple and straightforward, since the buoy is at a fixed location and the hindcast consists of one instant in time, both having regular temporal resolution – and when this is not case, interpolation is still trivial. The task is more complex when pairing altimeter data and forecast models. The polar-orbit satellites do not measure at fixed locations, but rather they revisit a site once every 10–35 days (Cooper and Forristall, 1997). Furthermore, operational forecasts have two time dimensions, the first related to the forecast cycle (the specific time of the analysis), and the second related to the forward forecast leads. When pairing certain altimeter measurement with the first instant of the forecast model, by the time the next forecast step comes, the altimeter will be displaced to another location, which compromises the consistency of evaluating the whole forecast range with the same measurement.

The solution we use here is to make the forecast data selection for each altimeter measurement by moving backwards in time, instead of forward. The coordinate of the altimeter observation is used as a reference point (e.g. a certain longitude, latitude, and time) and matched with prior forecasts at various lead times all verifying at the same reference point. For example, we can select the 24-hour forecast starting from 1 day prior, the 48-hour forecast starting from 2 days prior, and so on. This procedure can also be applied with a temporal resolution of 6 hours, which is the time between consecutive GWES cycles (Figure 2).

The ensemble introduces another dimension to the forecast system. The result is a matrix of 21 members (20 plus the control run) times 10 days of forecast with 6-hour resolution (41 steps) at each model grid point. Each altimeter measurement allocated to the 0.5°X0.5° grid is paired to the 861 model results (see Figure 3C). With a perfect forecast simulation, the matrix should present a value close to the measurement and Figure 3C, regarding the difference from GWES to the observation, should be close to zero. However, with model error and uncertainty in initial and boundary conditions, predictability deteriorates and the ensemble spread increases with time.

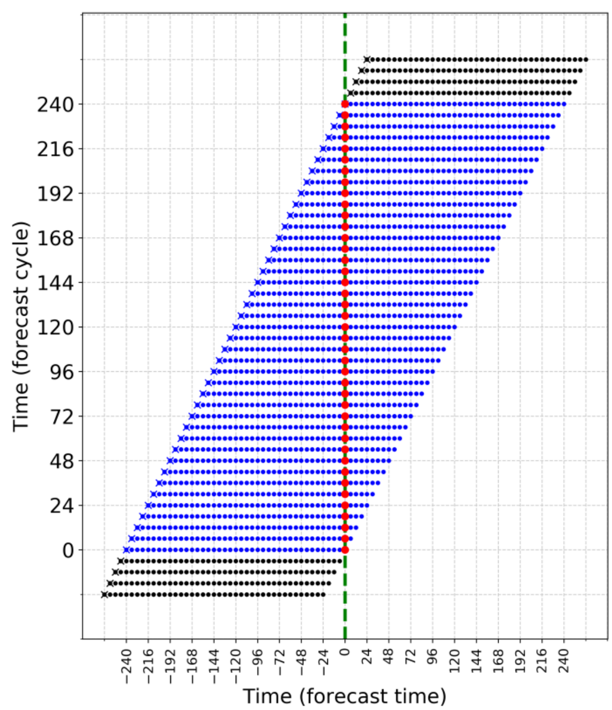

Figure 2 - Schematic of time and forecast cycle data selection (both in hours), for a specific time and location of the observation, centered at the satellite time (green dashed line). The y-axis shows the progress of forecast cycle (resolution of 6 hours) and the x-axis presents the forecast time, involving 240 hours (10 days) per cycle. The "x" sign at the beginning of each array illustrates the nowcasts; in black are the forecast cycles not used for the satellite/model matchup, and the blue color illustrates the 41 forecast cycles selected for the comparison. The 41 red dots are the exact values selected to match the single satellite observation, each one associated with a different forecast cycle but having the same time. When we include the 20 ensemble members to each of these 41 selected values, it is obtained the matrix illustrated by Figure 3C.

This backward scheme of model and observation pairing, illustrated by Figure 2, was applied to the whole year of 2017. The most extreme event of Hs, presented in Figure 3, occurred in the Labrador Sea with maximum Hs of 12.5 at 59.0°N / 52.0°W on Jan-25-2017 06Z. Figure 3A shows the evolution of the ensemble members together with the control run and arithmetic ensemble mean, and Figure 3B presents the same information but fitting an empirical distribution function to the 21 ensemble members of each forecast cycle. From Figure 3B and Figure 3A, it is possible to note that 10 to 8 days prior to the event, the forecast system did not foresee the upcoming extreme conditions. From forecast day 7, the ensemble members started to diverge and the spread increased, although the ensemble mean (EM) was still very low compared to the severity of the event. It suggests that some GWES members initially pointed to extreme conditions. From forecast day 4 towards the nowcast, Hs moved to much higher values and the spread decreased, indicating that GWES correctly captured the event so small upgrades were made until the instant of maximum of the storm. Figure 3C presents the same evolution described, and shows how the underestimation of GWES members was modified throughout the forecast cycles and the approach of the extreme event.

Figure 3 illustrates a successful prediction from GWES, at least considering the first seven forecast leads, and exemplified the high quality of wave forecast systems nowadays, also discussed by Bidlot (2017) through his historical analysis of evolution of forecast model skills. Another recent successful example of ensemble prediction was the Category 5 Hurricane Irma, in September 2017. The ensemble system of NCEP allowed forecasters and decision makers to issue the alert six days prior to the arrival of the event in the USA. Using one year of data covering the whole globe allow us to expand the assessment through a multivariate analysis using meaningful evaluation metrics.

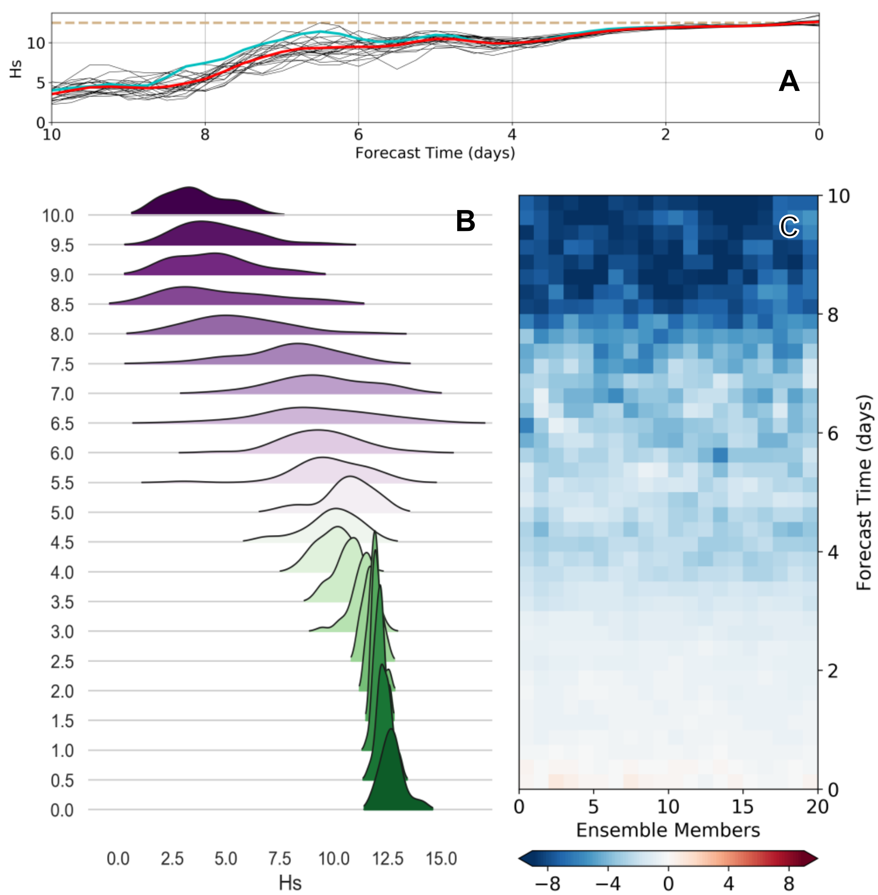

Figure 3 – Visualization of the most extreme value of Hs (m) measured by altimeters in 2017, at 59.0°N 52.0°W on Jan-25-2017, and the GWES performance for this time and location. In this event, Cryosat2 recorded the maximum Hs of 12.5 m at Labrador Sea. Panel (a) show the evolution of Hs for the control (cyan), ensemble members (black), and arithmetic ensemble mean (red) as a function of forecast time, associated with the same instant of maximum Hs, plotted as the dashed straight line (brown). Panel (B) presents the evolution of the empirical distribution functions of the 21 ensemble members for each forecast cycle, covering from the forecast 10 days prior to the event (top) until the nowcast (bottom); where the x-axis shows Hs and y-axis the forecast time. Panel (C) shows the difference of the GWES members minus satellite observation (fixed at 12.5 m) involving 10 forecast days (41 cycles) and 20 ensemble members, where blue colors represent underestimation of GWES and red colors overestimation.

Seven metrics are calculated to investigate the behavior of the GWES errors, described by equations 1 to 7; where x is the GWES forecast, y is the altimeter data, and the overbar indicates the arithmetic mean. Willmott and Matsuura (2005), Jolliff et al. (2009), and Mentaschi et al. (2013) discuss the limitations of using the root-mean-square error (RMSE) for model assessments. Chai and Draxler (2014), on the other hand, argue that just avoiding RMSE in favor of mean absolute error (MAE) is not the solution. Instead, Chai and Draxler (2014) suggest a combination of metrics beyond RMSE and MAE. Based on the study of Mentaschi et al. (2013) and the implementation of Campos et al. (2018a), we give special attention to the separation between the systematic error (equations 1 and 2) and the scatter component of the error (equations 5 and 6), as well as absolute (equations 1, 3, and 5) and normalized metrics (equations 2, 4, and 6), building a complete set of metrics to evaluate GWES. The correlation coefficient (CC) is also included (equation 7), where σ_x and σ_y are the standard deviations of the model and the observations respectively. Unlike other the metrics, CC values close to zero indicate poor results and the best models should be close to 1.

The normalized metrics (equations 2, 4, and 6) are divided by the observations and they are not divided by the total count of samples, n. Mentaschi et al. (2013) describe each error metric with more details. Therefore, NBias, NRMSE, and SI can be interpreted as ratios, or percentage errors when multiplied by 100. From equation 6, it can be seen that the scatter index (SI) is the normalized scatter component of the RMSE ($SCrmse$). Furthermore, equation (1) related to bias is the same as equation (1) of Chai and Draxler (2014), related to MAE. An additional discussion and guidance regarding forecast verification can be found at Ebert et al. (2013).

$$Bias = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i) \tag{1}$$

$$NBias = \frac{\sum_{i=1}^{n}(x_i - y_i)}{\sum_{i=1}^{n} y_i} \tag{2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} \tag{3}$$

$$NRMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{\sum_{i=1}^{n} {y_i}^2}} \tag{4}$$

$$SCrmse = \sqrt{\frac{\sum_{i=1}^{n}[(x_i - \bar{x}) - (y_i - \bar{y})]^2}{n}} = \sqrt{RMSE^2 - Bias^2} \tag{5}$$

$$SI = \sqrt{\frac{\sum_{i=1}^{n}[(x_i - \bar{x}) - (y_i - \bar{y})]^2}{\sum_{i=1}^{n} {y_i}^2}} \tag{6}$$

$$CC = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y} \tag{7}$$

3. Assessment Results

Due to the large number of altimeter data available in 2017, in our assessment of GWES we can afford to resample the altimeter/GWES pairs as a function of other variables that affect the forecast skill. Initially the assessment is performed as a function of forecast time and sea-state severity, and then as a function of the location, building global maps of GWES errors.

3.1 GWES wave-height error versus forecast time and percentile levels

The scatter component of the forecast error is presented in Table 1, where the deterministic forecast (control run) is compared with the arithmetic ensemble mean, EM. While the results for the first forecast days are similar, after the third day both SI and CC increasingly diverge, with the EM presenting much smaller errors than the control run. For U10, for example, the SI for the EM at day 10 is similar to the SI of the deterministic forecast at day 5 – a gain of five days in predictability of the wind speed. For the correlation coefficient, this gain is equal to four days. For the SI of significant wave height (Hs), there is a gain of three days. Table 1 highlights the importance of ensemble forecasting for those interested in longer forecasts ranges, especially after the fifth day. Table 1 also shows that the forecast for Hs present better results than for U10, for the whole forecast range.

The complete assessment of wave forecasts provided by Bidlot (2017), involving 16 operational centers, found SI from 0.13 to 0.20 for the nowcast and 0.30 to 0.37 on day-5. Although a direct comparison of Bidlot (2017) with Table 1 is not possible due to different observations utilized, it is interesting to note that the assessment of Hs from GWES for both the EM and the control run present smaller errors than reported by Bidlot (2017), where the SI of the GWES nowcast is 0.10 and day-5 is 0.20 to 0.23. It is worth to follow the next reports issued by the Lead Centre for Wave Forecast Verification (LC-WFV) that will probably provide a more suitable comparison involving recent data.

Table 1 – Scatter Index (SI) and Correlation Coefficient (CC) as a function of forecast time, from day 0 (nowcast) to day 10. For each variable and error metric, the control run is compared with the arithmetic ensemble mean (EM) of the 20 members. Results integrate the assessment of the whole globe using altimeter data.

	U10				**Hs**			
Forecast	**SI**		**CC**		**SI**		**CC**	
Day	Control	EM	Control	EM	Control	EM	Control	EM
0	0.146	0.143	0.940	0.942	0.105	0.108	0.971	0.970
1	0.170	0.157	0.920	0.930	0.114	0.113	0.969	0.969
2	0.202	0.176	0.888	0.909	0.132	0.127	0.961	0.963
3	0.239	0.199	0.843	0.880	0.159	0.149	0.943	0.949
4	0.283	0.226	0.780	0.840	0.193	0.176	0.915	0.928
5	*0.325*	0.252	0.709	0.794	0.231	0.206	0.876	0.899
6	0.362	0.273	0.638	0.749	0.269	0.232	0.829	0.869
7	0.396	0.292	0.568	0.706	*0.307*	0.258	0.775	0.836
8	0.419	0.305	0.515	0.672	0.337	0.278	0.730	0.806
9	0.435	0.315	0.472	0.645	0.356	0.289	0.686	0.781
10	0.449	*0.322*	0.438	0.622	0.377	*0.301*	0.645	0.758

Following the assessment structure of Hernandez et al. (2015), we complement the error metrics with the Taylor Diagram (Taylor, 2001) as it summarizes multiple aspects of model performance. Figure

4 confirms the increasing error with forward forecast leads and divergence of ensemble mean from the control run and ensemble members. For U10 this evolution leads the EM to progressively underestimate the observations. For Hs, the EM dots in the Taylor Diagram are also on the left of the control run and ensemble members, but without underestimation (on the right of the green curve). Both Table 1 and Figure 4 show very small correlation coefficients associated with forecast day 10, around 0.44 for U10 and 0.65 for Hs regarding the control run. These values are significantly improved to 0.62 and 0.76, respectively, when using the EM. The same increasing rate of improvement throughout forecast time of the EM compared to the control run is found in the RMSE, which can be easily noticed using the Taylor Diagrams.

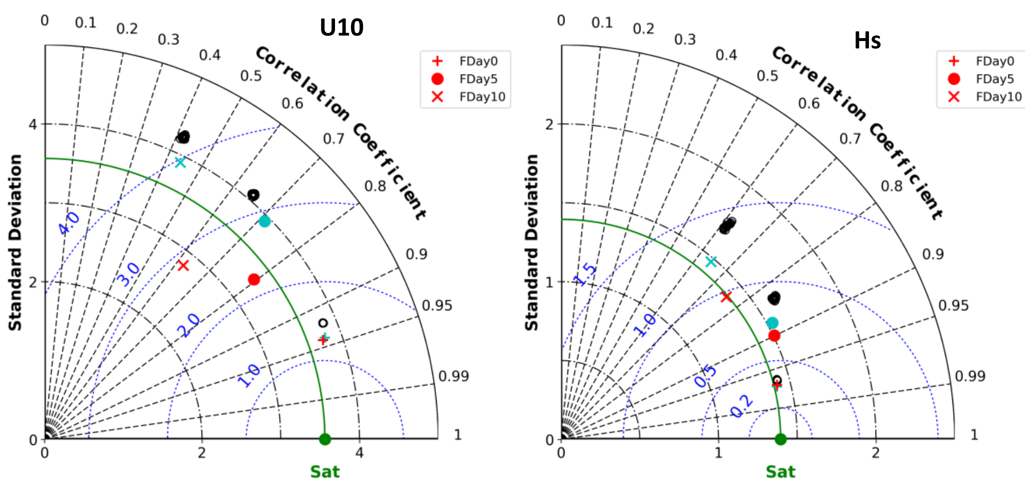

Figure 4 – Taylor Diagrams for U10 (left) and Hs (right) regarding three forecast ranges: day-0, day-5, and day-10. In terms of plot, the black dashed rays indicate the correlation coefficient, the dashed-dot black curves indicate the standard deviation (from which can be inferred a relative underestimation or overestimation of results), and the dotted blue curves are the RMSE. The green line presents the satellite observation as the reference. Concerning the results, the 20 ensemble members are plotted in black, the control run in cyan, and the EM in red. Markers on the right side of the green curve indicate overestimation of the model in regards to the satellite observation, whereas results on the left side indicate underestimation.

We next examine errors for changing severity of wave heights and wind speed. Each quantile, which is the inverse of the cumulative distribution, is calculated for increasing percentiles from 0 to

98%. In Figure 5 we visualize errors for wave conditions ranging from calm to extreme, together with the forecast time and compare the control run to the EM, with respect to the systematic (bias) and scatter errors ($SCrmse$). Typically, the largest errors are associated with longer forecast ranges and higher percentiles. These results indicate that the global assessment using altimeter data agrees with the previous results of Campos et al (2018a) using buoy data, where it was found that the largest errors occur after the fourth day of forecast under severe conditions. Systematic errors are similar between the deterministic and probabilistic forecasts, as expected. However, there is a large reduction of the scatter errors in the EM. This is more evident for U10, where the $SCrmse$ above 4 m/s associated with strong winds beyond forecast day 7 drops to values around 2 m/s. The benefit on the surface winds using the ensemble approach is propagated to the wave fields and the largest $SCrmse$ of 1.8 m is reduced to 1.3 m.

The problem of increasing bias with severity and percentiles is not addressed by the ensemble approach and requires investigation on model tuning and development of bias correction post-processing; which is out of the scope of our study. Regarding simplistic bias correction, for example, Reguero et al. (2012) based on Mínguez et al. (2011) suggested an efficient calibration of wave simulations with satellite altimetry data, while Campos et al. (2018b), based on Tolman (1998), used buoy and scatterometer data to calibrate surface winds and wave model parameters.

The systematic errors combined with low spread, usually at short-range forecasts, can be a problem as the ensemble spread does not properly represent the uncertainties of the prediction system - discussed by Bunney and Saulter (2015). Figure 5 suggests that this is not critical for GWES as the largest biases are found beyond forecast day 4. Nevertheless, Saetra and Bidlot (2004) found a small tendency for overconfidence in the wave probability forecasts for large waves above 6 and 8 m.

For this reason, we decide to include the estimation of the spread as a function of forecast time and percentile (Figure 6), as a complement to Figure 5. The largest spread for both U10 and Hs are found beyond forecast day 6 and associated with U10 above 10 m/s and Hs above 4 m. It matches the combination of percentiles and forecast ranges with large bias and $SCrmse$, representing the increased uncertainties of the NCEP ensemble prediction system.

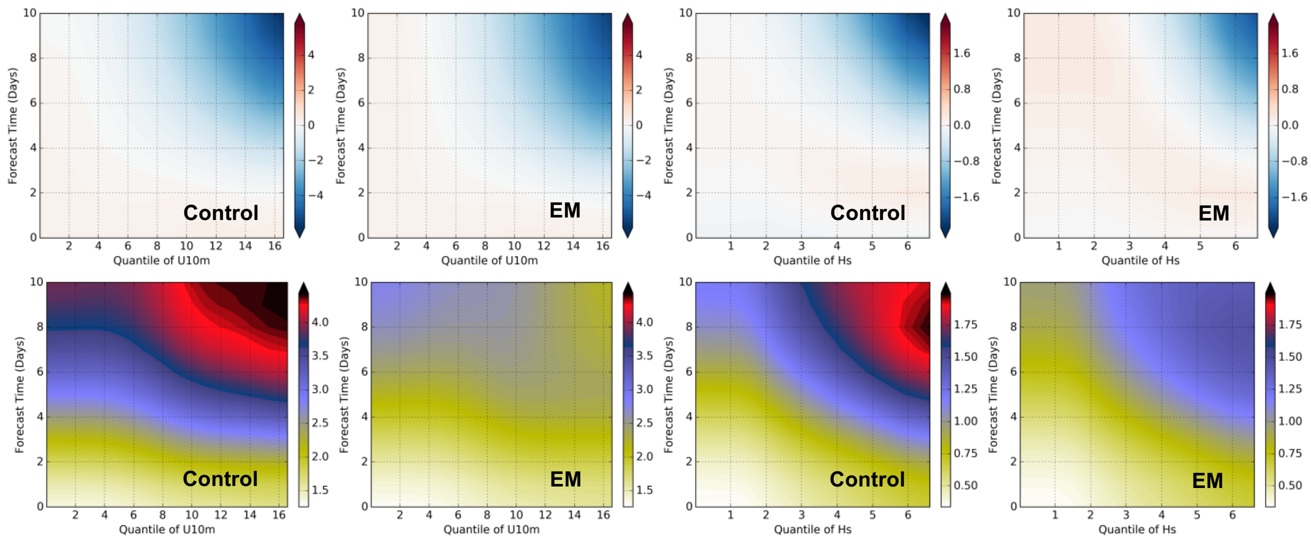

Figure 5 – **Bias** (top row) and $\boldsymbol{SCrmse}$ (bottom row) as a function of forecast time (y-axis) and quantiles (x-axis). For the bias plots, blue colors indicate that the model underestimates the observations, while red colors indicate the model overestimates the observations. The first two columns on the left are the wind speed at 10m (U10) in m/s, and the two columns on the right the significant wave height (Hs) in meters.

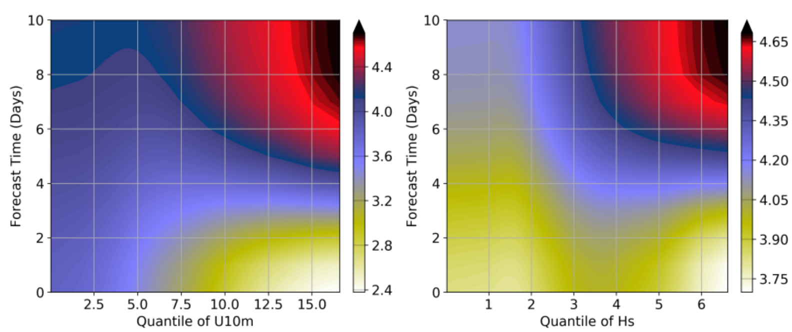

Figure 6 – Spread of the 20 ensemble members as a function of forecast time (y-axis) and quantiles (x-axis), for **U10** (left) and **Hs** (right).

The large number of observations in satellite databases relative to buoys, also allows a deeper investigation in the probabilistic domain so it can be verified if the forecast results can reproduce the distribution of observations. Same as performed by the ensemble assessment of Bunney and Saulter (2015), QQ-plots and probability distribution functions (PDFs) of U10 and Hs are presented in Figure 7, divided into three different forecast ranges.

The nowcast shows a good agreement between ensemble members, the arithmetic ensemble mean (EM), and the control run, with values close to perfect agreement. For the upper percentiles, the agreement of Hs from GWES with observations is better than the agreement for U10, where the strongest winds are slightly overestimated by the NCEP forecast. Moving to forecast day 5, the ensemble members and the control run start to diverge from the ensemble mean (EM). In the highest quantiles, particularly at longer lead times, the ensemble members and the control run tend to overestimate U10 and Hs compared to the observations, while the EM underestimates measurements of U10 at the longest lead times – confirmed by both QQ-plots and PDFs. The EM tends to overestimate measurements of U10 and Hs in calm and moderate conditions. The evolution of Hs quantiles closely follows U10, with Hs slightly shifted to higher values for the GWES in relation to altimeters, possibly due to tuning of the wave model parameters that control the transfer of momentum from surface winds to the wave spectra. Other explanation may be that altimeters under-sample more extreme sea states (Alves and Young, 2004), and spatial aliasing in model simulations may move the location of such cases into calmer regions depicted in the satellite data.

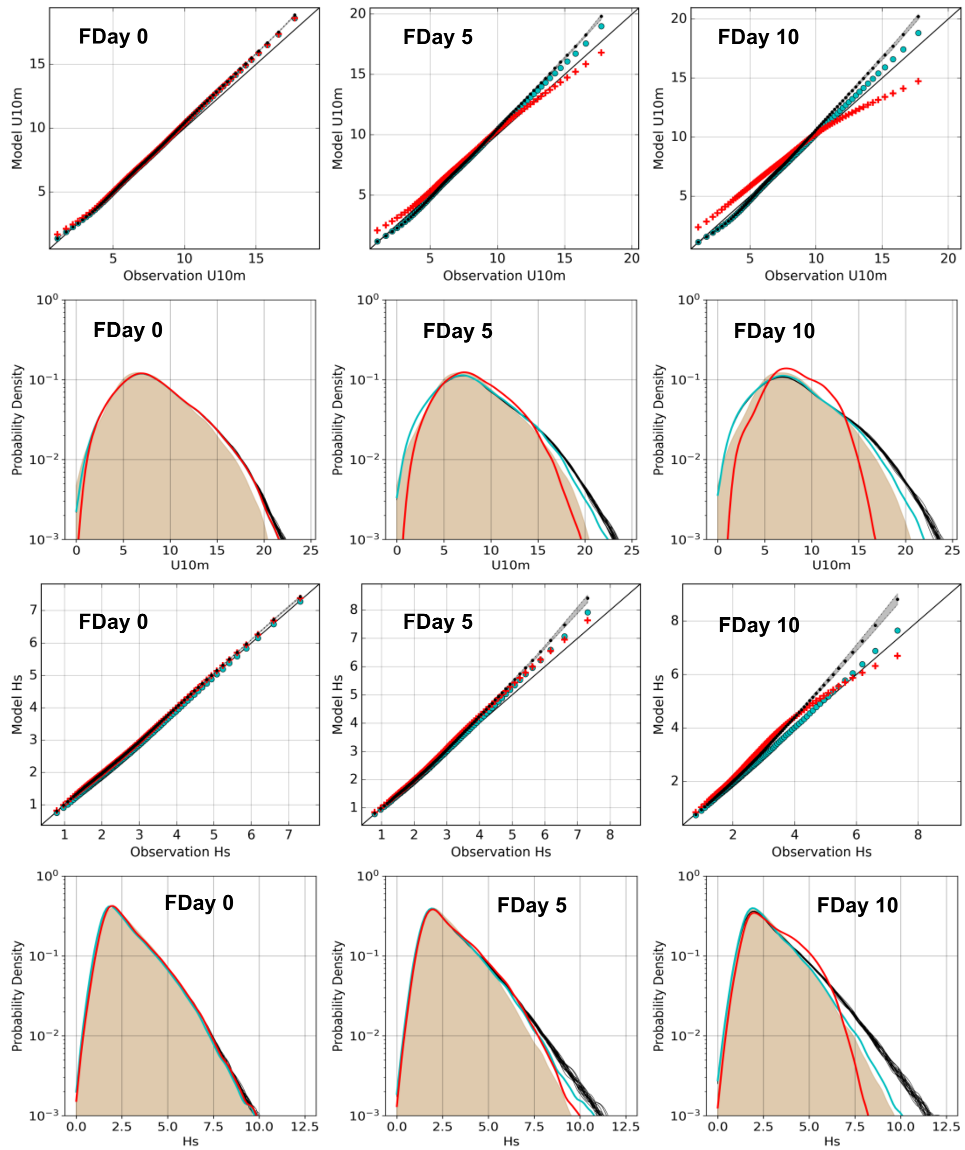

Figure 7 - QQ-plots and probability density functions (PDFs) for three different forecast ranges. The first two rows at the top of the figure show wind results (U10), in m/s, while the two bottom rows show results for wave heights (Hs), in meters. Black: ensemble members. Cyan: control run. Red: ensemble mean. The shaded brown at the PDF plots represents the empirical PDF, for the observations.

The PDF plots of Figure 7 corroborates with the results from the QQ-plots. They are also useful to indicate, through the density function, where in terms of intensity the bulk of the altimeter measurements (shaded brown) is concentrated, since they are invariant to the forecast time, as discussed before. The PDFs show most of the occurrence of U10 between 5 to 10 m/s and Hs between 1 to 4 m, which suggests that the discrepancy at larger quantiles should have a minor impact on the average statistics and error metrics, however, these discrepancies remain relevant. Figure 7 shows that the arithmetic ensemble mean (EM) of the ensembles deteriorates the tail of the PDF when compared to the observations, which can severely compromise the higher-order probabilistic moments and possible applications involving extrapolation and extreme value analysis (EVAs). In regards to the NCEP ensemble, this is more evident for U10 than Hs. This is an expected consequence of using the arithmetic EM, which eliminates higher wave-height values associated with ensemble member that may be closer to the "true" wave height. This result in itself justifies the development and use of alternative ways to determine ensemble means and probabilistic products in general, such as the proposed use of nonlinear means obtained via the use of neural networks made in a separate paper (e.g., Campos et al, 2019).

3.2 Spatial distribution of GWES errors

The construction of error maps was based on the methodology of Young and Holland (1996). After allocating the satellite tracks into the regular GWES grid of 0.5°X0.5° (section 2), the matchups of altimeter and GWES were selected within the radius of 2° to compute the error statistics for each location. Equations 1 to 7 were applied to calculate the metrics for given latitudes and longitudes, building the global maps of different types of errors. Once again, the emphasis will be on the interpretation of systematic and scatter errors separately.

Figure 8, Figure **9**, and Figure **10** present the main results of this paper, containing the maps of bias, $SCrmse$, and RMSE of GWES. It is now possible to clearly notice a strong spatial dependence of GWES errors, with the effect of the Atmospheric Circulation including the Hadley and Ferrel Cells, as well as the ITCZ and latitudes dominated by westerly winds. We can confirm the increase of GWES errors with longer forecast ranges; however, the rate is much larger at mid-latitudes than at tropical locations. This effect can be visualized in Figure 11 where the errors were integrated over the longitude to provide the errors versus Latitude.

First looking at the bias of the nowcasts (forecast day 0), both control and EM of U10 in Figure 8 present a small overestimation of wind intensities compared to the measurements. In extratropical areas this behavior increases when moving to forecast day 5 and 10 but the opposite occurs at the Equator, where GWES starts to underestimate the wind measurements. The bias of Hs, instead, shows a slight underestimation at the nowcast over the entire grid except in some extratropical locations in the Southern Hemisphere, more evident in the EM. On forecast days 5 and 10, the overestimation of Hs at mid-latitudes becomes much larger and non-symmetric in terms of Northern and Southern Hemispheres. For both U10 and Hs, the differences between the control run and EM increases mainly at extratropical locations with longer forecast ranges, confirmed by Figure 11, where the EM has larger bias than the control.

The scatter components of the errors ($SCrmse$) of U10 and Hs are small at the nowcast and very similar between the control member and EM. The $SCrmse$ increases at extratropical areas on forecast day 5 and 10, as well as the differences between the control and EM. In this case, the control member has much larger errors than the EM. The forecast day 10, for example, shows $SCrmse$ of U10 around 5 m/s for the control member and 3.5 m/s for the EM at mid-latitudes. Regarding Hs, the $SCrmse$ is 1.8 m for the control member and 1.3 m for the EM. It can be visualized by the global maps of Figure 8 and Figure **9**, as well as the error distribution over the latitudes of Figure 11.

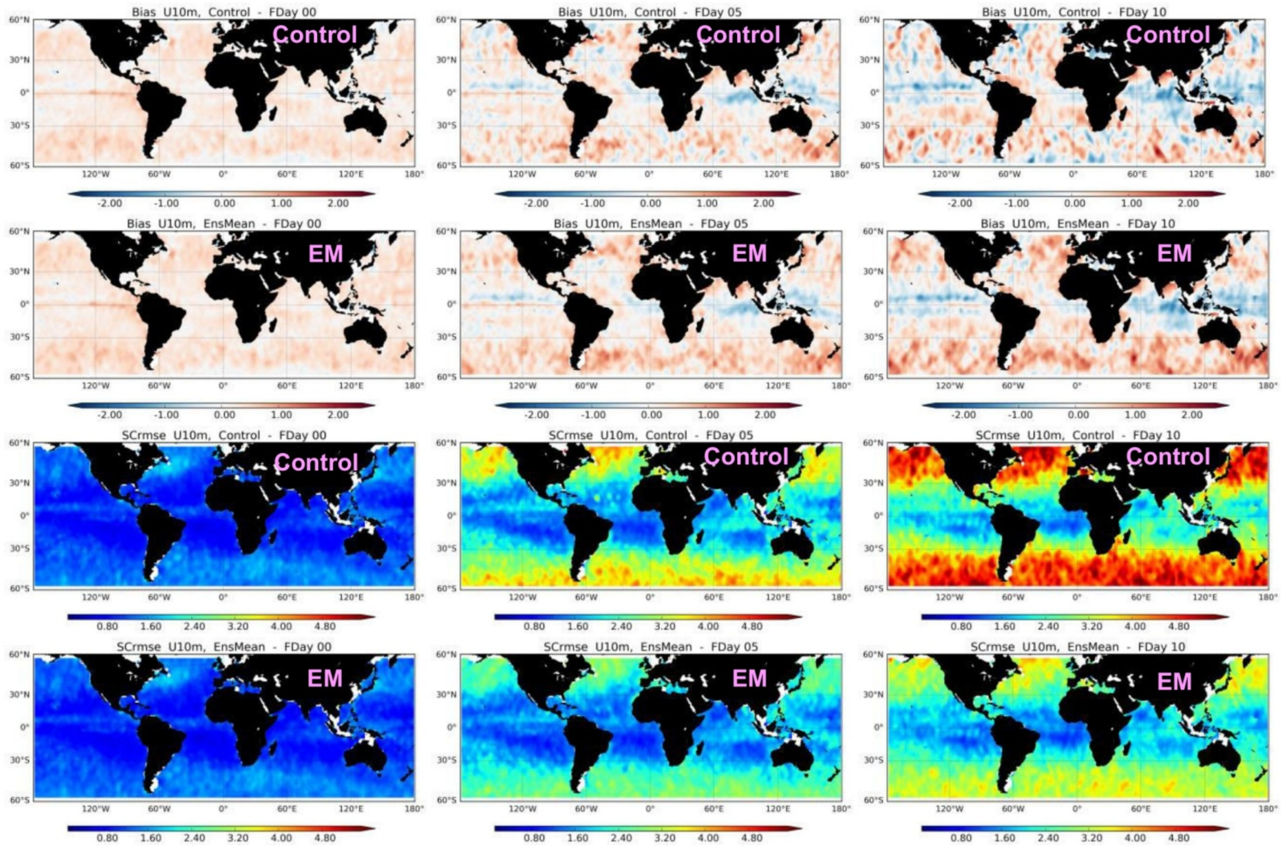

Figure 8 – Global maps of GWES error of **U10** (in m/s), comparing the control run (deterministic forecast) with the arithmetic ensemble mean (EM, probabilistic forecast). **Bias** in the first two top lines (red being overestimation of GWES and blue underestimation) and $\boldsymbol{SCrmse}$ in the last two bottom lines of plots. Columns represent different forecast times: left column the nowcast, center column day-5 forecast, and right column day-10 forecast.

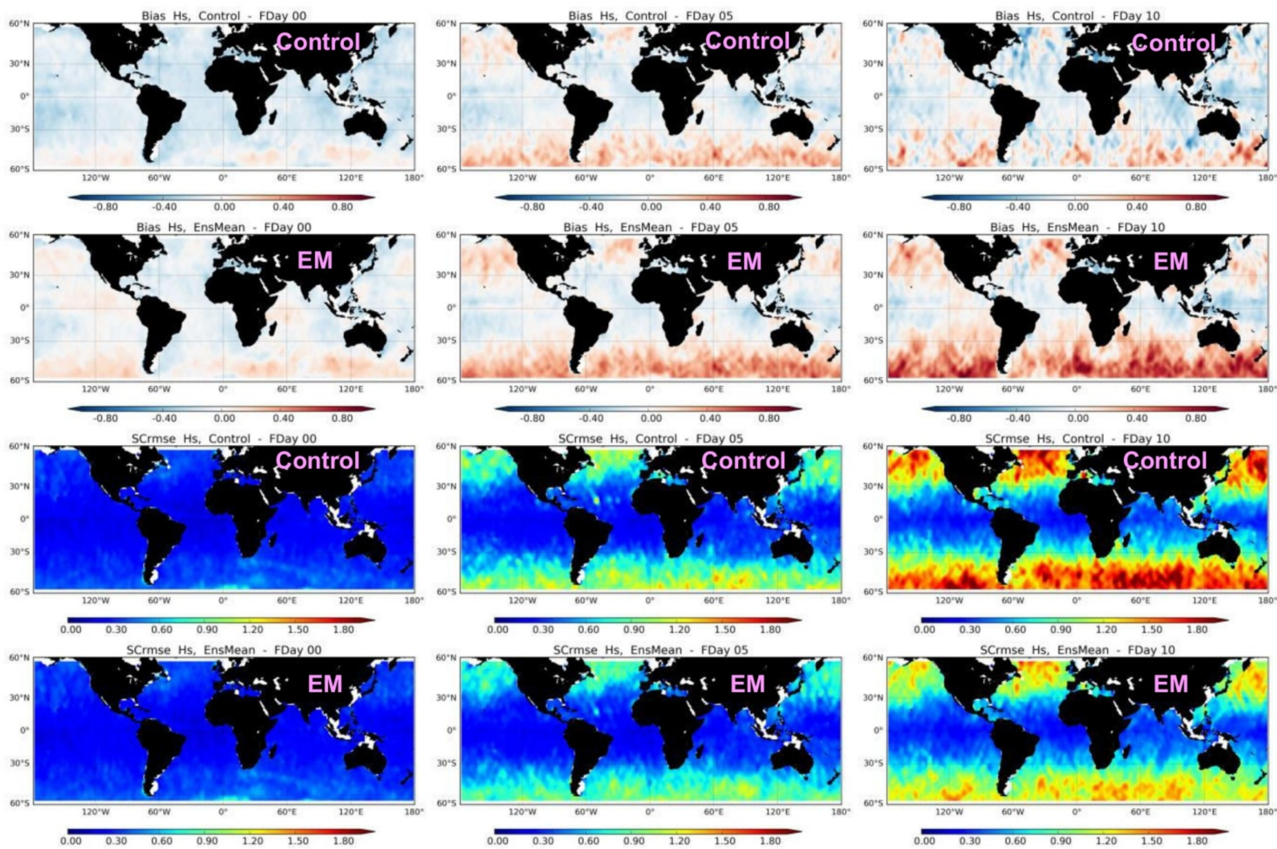

Figure 9 - Global maps of GWES error of **Hs** (in meters), comparing the control run (deterministic forecast) with the arithmetic ensemble mean (EM, probabilistic forecast). **Bias** in the first two top lines (red being overestimation of GWES and blue underestimation) and $\boldsymbol{SCrmse}$ in the last two bottom lines of plots. Columns represent different forecast times: left column the nowcast, center column day-5 forecast, and right column day-10 forecast.

The error maps of Figure 10 present the final results of RMSE, where it is possible to confirm, again, the dependence of wave height errors on the quality of surface wind speeds. As indicated by equation (5), the RMSE combines the systematic and scatter error. Jolliff et al. (2009) investigate how the bias contributes to the magnitude of the total Root-Mean-Square Difference. For our specific analysis, it has been verified that $SCrmse$ is at least twice the bias, and so the RMSE is influenced more by the increase of scatter errors than by the systematic errors. In general, at forecast day 10, the reduction of

RMSE of the EM compared to the deterministic run (control) varies from 20% to 30%, and smaller improvements are found at tropical locations.

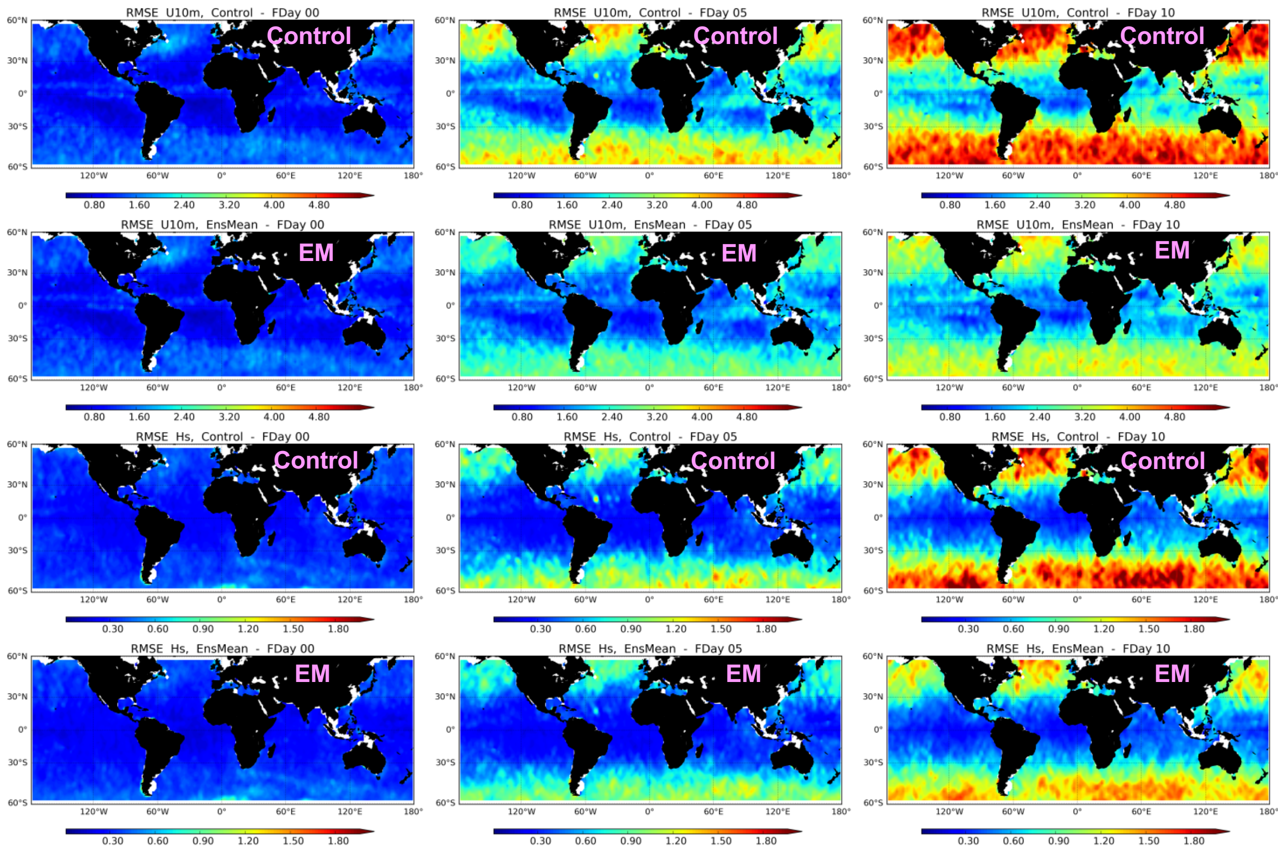

Figure 10 – Final Global maps of RMSE of **U10** (in m/s) in the two top row and **Hs** (in meters) at the two bottom rows, comparing the control run (deterministic forecast) with the arithmetic ensemble mean (EM, probabilistic forecast). Columns represent different forecast times: left column the nowcast, center column day-5 forecast, and right column day-10 forecast.

The forecast errors versus Latitude presented by Figure 11 partially present redundant information to Figure 8 and to Figure 10. However, the comparisons of curves as well as the correlation coefficient plots provide additional information regarding differences between Northern and Southern Hemispheres. The systematic errors of U10 and Hs at extratropical latitudes in the Southern Hemisphere are much larger than the same in the Northern Hemisphere – valid for the whole dataset including ensemble members, control run, and ensemble mean. At forecast day 10 the bias of Hs at 50°S is 0.50 m while at 50°N it is 0.15 m. For the wind speeds these differences are not as large as for

Hs but the bias of the EM of U10 in the Southern Hemisphere reaches 0.7 m/s while in the Northern Hemisphere it does not exceed 0.5 m/s. Such discrepancies are not very pronounced in the scatter errors but the correlation coefficients also point to worse performances in the Southern Hemisphere, especially in locations south of 50°S.

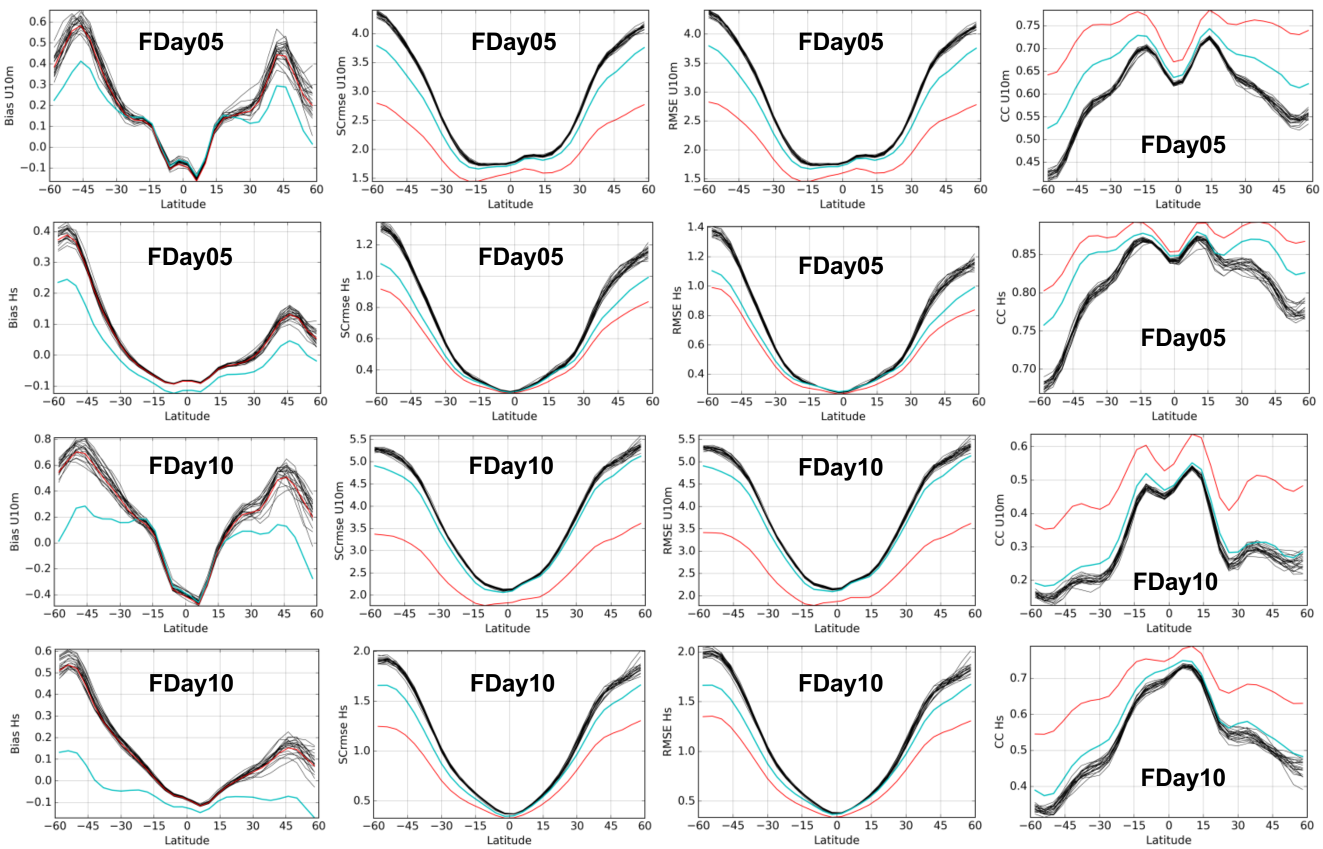

Figure 11 – GWES errors versus Latitude for **U10** (in m/s) and **Hs** (in meters). From left to right: Bias, SCrmse, RMSE, and CC (dimensionless). The top rows contain results for the forecast day-5 and the two bottom rows for forecast day-10. Black curves: ensemble members. Cyan curves: control run. Red curves: ensemble mean.

The unbalanced performance of NCEP ensemble forecasts of U10 and Hs between Hemispheres might be associated with the larger amount of continent and observations in the Northern Hemisphere. Moreover, the larger ocean basins in the Southern Hemisphere allow errors to propagate further distances and longer periods which can propagate and accumulate forecast errors. This is just a speculation and this subject requires more investigation since the Southern Ocean is known to be an extremely dangerous area to sail, and depends on the performance of global forecasts as the NCEP ensemble forecast system. Moreover, although the correlation coefficient plots of Figure 11 indicate better performances at tropical areas, they also show a small deterioration of the forecast at the Equator, which could be associated with mesoscale storms that are not properly simulated by the resolution of 0.5° of GWES. This might be the reason why the effect is more evident for U10 than Hs that respond much more to synoptic scale wind fetches.

Finally, Figure 11 also confirms an unexpected feature found in the previous figures, where Hs and U10 biases are larger for the EM than for the control run, especially at longer forecast ranges. It is well-known, as described before, that the ensemble approach reduces the scatter error and improves the correlation coefficient, and it is not meant to reduce bias. However, we expected similar values of bias of the EM compared to the control run and ensemble members, and not larger biases. This problem does not severely compromise the overall performance of the ensemble product since the greatest portion of the RMSE comes from the scatter component of the error ($SCrmse$), as concluded above.

4. Conclusions

The multivariate distribution of the NCEP Global Wave Ensemble System (GWES) errors has been investigated using altimeter data and seven error metrics, giving special attention to the comparison between the control run (deterministic forecast) and the ensemble mean. The first characteristic we observe, which confirms previous assessment studies including Cao et al. (2007) and Alves et al. (2013), is the reduction of the scatter errors of the ensemble forecast beyond the fifth day compared to the control run. Table 1 shows a gain of three to five days in predictability of Hs and U10. This is also in agreement with Saetra and Bidlot (2004) based on ECMWF products, who found that the arithmetic ensemble mean outperforms the control run. Figure 5 and Figure **7** add the increasing percentiles into the analysis and highlight the challenge of predicting extreme events using both ensemble and deterministic forecasts. The arithmetic mean of the ensemble members has smaller scatter error but shows underestimation of extreme events, which compromises the extremal tail of the PDFs.

As described by Jolliff et al. (2009), the “skill” portion of skill assessment may be mathematically defined, but the “assessment” will invariably rely upon the value judgments of the investigator. Based on our results, GWES users can judge and decide to use deterministic or ensemble forecasts, and have detail information of Hs and U10 errors for their specific locations and magnitude of interest. Considering the discussion of Willmott and Matsuura (2005), Jolliff et al. (2009), and Mentaschi et al. (2013), combined with our multivariate assessment and the whole set of results, we conclude that the arithmetic ensemble mean of GWES, derived from the probabilistic forecast, significantly outperforms the control run and the NCEP deterministic forecast.

Several studies have investigated the spatial behavior of wave models, as for example Stopa and Cheung (2014) and Campos and Guedes Soares (2016); however, this is the first work concerning the spatial distribution of the error of a global wave ensemble forecast. We identified similar systematic errors between the control and the EM calculated by integrating results over the entire globe. When the bias was calculated for each location, we see a heterogeneous distribution in space. In most locations, the EM has larger bias than the control member and this difference is larger for Hs than for U10, i.e., the bias of the EM of Hs is much higher than the control member, especially in the Southern Hemisphere. One possible explanation is the larger portions of water in the Southern Hemisphere, which makes the wave model to amplify small systematic errors. The analysis using maps of $SCrmse$ shows the great benefit of the ensemble approach mainly at mid-latitudes and longer forecast ranges.

Therefore, for reliable wind and wave forecasts beyond 7 days at mid and high latitudes, it is essential to use ensemble forecast products, however it is also essential to apply a geographically dependent bias correction.

The bias of the EM at longer forecast ranges is higher than the control run but the scatter errors of the EM are much smaller than the control. The discrepancies between them increase poleward of 20°N and 20°S. Therefore, if an efficient bias correction algorithm could be applied to the ensemble forecast in post-processing, this could maintain small scatter errors inherent to the ensemble approach while reducing the systematic errors of the GWES. Further than encouraging the use of probabilistic wave model products in support of wave guidance to marine weather forecasts, the results presented in this paper support the idea that the development of alternative methods to determine ensemble means is warranted. A step in that direction is discussed in a companion paper (e.g., Campos et al., 2019). Although our results are limited to products from a single wave ensemble system, it is believed that the benefits outlined here would also be sustained when assessing results from combined ensemble products, which will be the subject of work to be pursued in the near future.

Acknowledgments

This work has been funded by the US National Weather Service Office of Science and Technology (NWS/OST), NOAA award number NA16NWS4680011, with further support in the last stage of development from Fundação para a Ciência e a Tecnologia (FCT – Portugal) under the project EXWAV (RD0504) number PTDC/EAM-OCE/31325/2017. The authors would like to acknowledge Dr. Todd Spindler, the GEFS atmospheric ensemble team at NCEP, and the Department of Atmospheric and Oceanic Sciences (AOSC) at the University of Maryland.

Data sources

NCEP’s Global Wave Ensemble Forecast:

- ftp://ftpprd.ncep.noaa.gov/pub/data/nccf/com/wave/prod

Altimeters:

- ftp://avisoftp.cnes.fr/AVISO/pub/
- ftp://ftp.star.nesdis.noaa.gov/pub/sod/lsa/cs2igdr/

References

Alves, J.H.G.M., Young I.R., 2004. On Estimating Extreme Wave Heights using Combined Geosat, Topex/Poseidon and ERS-1 Altimeter Data. Applied Ocean Research, Vol. 25, No. 4, pp. 167-186.

Alves, J.H.G.M., Wittman, P., Sestak, M., Schauer,J., Stripling, S., Bernier, N.B., McLean, J., Chao, Y., Chawla, A., Tolman, H., Nelson, G., Klots, S., 2013. The NCEP–FNMOC combined wave ensemble product. Expanding Benefits of Interagency Probabilistic Forecasts to the Oceanic Environment. Bulletin of the American Meteorological Society, BAMS, December 2013.

Bidlot, J.R., 2017. Twenty-one years of wave forecast verification. ECMWF Newsletter, 150, 31-36.

Boukabara, S.-A., Krasnopolsky, V., Stewart, J.Q., Penny, S.G., Hoffman, R.N., Maddy, E., 2019. Artificial Intelligence May Be Key to Better Weather Forecasts. Earth & Space Science News: https://eos.org/opinions/artificial-intelligence-may-be-key-to-better-weather-forecasts

Bowler, N.E. , Arribas, A. , Mylne, K.R. , Robertson, R.B. , Beare, S.E. , 2008. The MOGREPS short-range ensemble prediction system. Q. J. R. Meteorol. Soc. 134, 703–722.

Bunney, C., Saulter, A., 2015. An ensemble forecast system for prediction of Atlantic–UK wind waves. Ocean Modelling, 96, 103–116.

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

Campos, R.M., Guedes Soares, C., 2016. Comparison and assessment of three wave hindcasts in the North Atlantic Ocean. Journal of Operational Oceanography, v. 9, p. 26-44.

Campos, R.M., Krasnopolsky, V., Alves, J.H.G.M, Penny, S.G., 2019. Nonlinear Wave Ensemble Averaging in the Gulf of Mexico using Neural Networks. Journal of Atmospheric and Oceanic Technology, 36, 113-127.

Campos, R.M., Alves, J.H.G.M., Penny, S.G., Krasnopolsky, V., 2018a. Assessments of Surface Winds and Waves from the NCEP Ensemble Forecast System. Weather & Forecasting, 33, pp. 1533-1564. DOI: 10.1175/WAF-D-18-0086.1

Campos, R.M., Alves, J.H.G.M., Guedes Soares, C., Guimaraes, L.G., Parente, C.E., 2018b. Extreme wind-wave modeling and analysis in the south Atlantic ocean. Ocean Modelling 124, 75–93.

Candille, G., 2009. The multiensemble approach: The NAEFS example. Monthly Weather Review, 137(5), pp.1655-1665.

Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geosci. Model Dev., 7, 1247–1250.

Chen, H.S., 2006. Ensemble prediction of ocean waves at NCEP. Proc. 28th Ocean Engineering Conf., Taipei, Taiwan, NSYSU, 25–37.

Cooper, C.K., Forristall, G.Z., 1997. The use of satellite altimeter data to estimate extreme wave climate. Journal of Atmospheric and Oceanic Technology, 14(2):254–66.

Durrant, T.H., Woodcock, F, Greenslade, D.J.M., 2009. Consensus forecasts of modelled wave parameters. Weather Forecast 24, 492–503.

Ebert, E., Wilson, L., Weigel, A., Mittermaier, M., Nurmi, P., Gill, P., Göber, M., Joslyn, S., Brown, B., Fowler, T., Watkins, A., 2013. Progress and challenges in forecast verification. Meteorol. Appl. 20, 130–139.

Grumbine, R.W., 1996. Automated passive microwave sea ice concentration analysis at NCEP. NOAA Tech. Note 120, 13 pp.

Harpham, Q., Tozer, N., Cleverley, P., Wyncoll, D., Cresswell, D., 2016. A Bayesian method for improving probabilistic wave forecasts by weighting ensemble members. Environmental Modelling & Software 84, 482-493.

Hernandez, F., Blockley, E., Brassington, G.B., Davidson, F., Divakaran, P., Drévillon, M., Ishizaki, S., Garcia-Sotillo, M., Hogan, P.J., Lagemaa, P., Levier, B., Martin, M., Mehra, A., Mooers, C., Ferry, N., Ryan, A., Regnier, C., Sellar, A., Smith, G.C., Sofianos, S., Spindler, T., Volpe, G., Wilkin, J., Zaron, E.D., Zhang, A., 2015. Recent progress in performance evaluations and near real-time assessment of operational ocean products, Journal of Operational Oceanography, 8-S2, 221-238.

Hoffschildt, M., J. Bidlot, B. Hansen, and P. A. E. Janssen, 1999. Potential benefits of ensemble forecasting for ship routing. ECMWF Tech. Memo. 287, 25 pp.

Jolliff, J.K., Kindle, J.C., Shulman, I., Penta, B., Friedrichs, M.A.M., Helber, R., Arnone, R.A., 2009. Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment. Journal of Marine Systems, 76, 64–82.

Kalnay, E., 2003. Atmospheric modeling, data assimilation and predictability. Cambridge University Press, 341pp.

Le Traon, P.-Y., Antoine, D., Bentamy, A., Bonekamp, H., Breivik, L.A., Chapron, B., Corlett, G., Dibarboure, G., DiGiacomo, P., Donlon, C., Faugère, Y., Font, J., Girard-Ardhuin, F., Gohin, F., Johannessen, J.A., Kamachi, M., Lagerloef, G., Lambin, J., Larnicol, G., Le Borgne, P., Leuliette, E., Lindstrom, E., Martin, M.J., Maturi, E., Miller, L., Mingsen, L., Morrow, R., Reul, N., Rio, M.H., Roquet, H., Santoleri, R., Wilkin, J., 2015. Use of satellite observations for operational oceanography: recent achievements and future prospects, Journal of Operational Oceanography, 8-S1, 12-27.

Mentaschi, L., Besio, G., Cassola, F., Mazzino, A., 2013. Problems in RMSE-based wave model validations. Ocean Modelling, 72, 53–58.

Mínguez, R., Espejo, A., Tomás, A., Méndez, F.J., Losada, I.J., 2011. Directional calibration of wave reanalysis databases using instrumental data. Journal of Atmospheric and Oceanic Technology. doi:10.1175/JTECH-D-11-00008.1.

Psaraftis, H. N., Kontovas, C. A., 2014. Ship speed optimization: Concepts, models and combined speedrouting scenarios. Transportation Research. Part C: Emerging Technologies, 44, 52-69. DOI: 10.1016/j.trc.2014.03.001

Queffeulou, P., Croizé-Fillon, D., 2017. Global altimeter SWH data set. Laboratoire d'Océanographie Physique et Spatiale IFREMER. Report available at ftp://ftp.ifremer.fr/ifremer/cersat/products/swath/altimeters/waves/documentation/altimeter_wave_merge__11.4.pdf

Reguero, B.G., Menéndez, M., Méndez, F.J., Mínguez, R., Losada, I.J., 2012. A Global Ocean Wave (GOW) calibrated reanalysis from 1948 onwards. Coastal Engineering 65, 38–55.

Ribal, A., Young, I.R., 2019. 33 years of globally calibrated wave height and wind speed data based on altimeter observations. Nature - Scientific Data, 6:77, 1-15.

Saetra, Ø., Bidlot, J.R., 2004. Potential Benefits of Using Probabilistic Forecasts for Waves and Marine Winds Based on the ECMWF Ensemble Prediction System. Wea. Forecasting, 19, 673–689.

Sepulveda, H.H., Queffeulou, P., Ardhuin, F., 2015. Assessment of SARAL AltiKa wave height measurements relative to buoy, Jason-2 and Cryosat-2 data. Marine Geodesy, 38 S1, 449-465.

Stopa, J.E., Cheung, K.F., 2014. Intercomparison of wind and wave data from the ECMWF Reanalysis Interim and the NCEP Climate Forecast System Reanalysis. Ocean Modell. 75, 65–83.

Taylor, K. E., 2001. Summarizing multiple aspects of model performance in a single diagram, J. Geophys. Res., 106(D7), 7183–7192.

Tolman, H. L., 2016. User manual and system documentation of WAVEWATCH III version 5.16. NOAA/NWS/NCEP MMAB Tech. Note 329, 326 pp.

Tolman, H.L., 1998. Validation of NCEP's ocean winds for the use in wind wave models. Global Atmos. Ocean Syst. 6 (3), 243–268.

Tournadre, J., 2014. Anthropogenic pressure on the open ocean: The growth of ship traffic revealed by altimeter data analysis. Geo-phys. Res. Lett., 41, 7924–7932.

Willmott, C., Matsuura, K., 2005. Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in assessing average model performance, Clim. Res., 30, 79–82, 2005.

Young, I.R., Holland, G.J., 1996. Atlas of the oceans: Wind and Wave Climate. Pergamon Press, New York, pp. 241.

Zhou, X., Zhu. Y., Hou, D., Luo, Y., Peng, J., Wobus, R., 2017. Performance of the New NCEP Global Ensemble Forecast System in a Parallel Experiment. Bulletin of the American Meteorological Society. https://doi.org/10.1175/WAF-D-17-0023.1

Zieger, S., D., Greenslade, Kepert, J.D., 2018. Wave ensemble forecast system for tropical cyclones in the Australian region. Ocean Dynamics, 68, 603–625.