# Development and Validation of NOAA's 20-Year Global Wave Ensemble Reforecast

RICARDO M. CAMPOS,[a,b] ALI ABDOLALI,[c,d] JOSE-HENRIQUE ALVES,[e] MATTHEW MASARIK,[f] JESSICA MEIXNER,[g] AVICHAL MEHRA,[g] DARIN FIGURSKEY,[h] SAEIDEH BANIHASHEMI,[f] JOSEPH SIENKIEWICZ,[h] AND RICK LUMPKIN[b]

[a] *Cooperative Institute for Marine and Atmospheric Studies, University of Miami, Miami, Florida*
[b] *Atlantic Oceanographic and Meteorological Laboratory, National Oceanic and Atmospheric Administration, Miami, Florida*
[c] *U.S. Army Corps of Engineers, Engineer Research and Development Center, Vicksburg, Mississippi*
[d] *Earth System Science Interdisciplinary Center, College Park, Maryland*
[e] *Weather Program Office, Oceanic and Atmospheric Research, National Oceanic and Atmospheric Administration, Silver Spring, Maryland*
[f] *Lynker at Environmental Modeling Center, National Centers for Environmental Prediction, National Weather Service, National Oceanic and Atmospheric Administration, College Park, Maryland*
[g] *Environmental Modeling Center, National Centers for Environmental Prediction, National Weather Service, National Oceanic and Atmospheric Administration, College Park, Maryland*
[h] *Ocean Prediction Center, National Centers for Environmental Prediction, National Weather Service, National Oceanic and Atmospheric Administration, College Park, Maryland*

(Manuscript received 11 March 2024, in final form 14 August 2024, accepted 16 August 2024)

ABSTRACT: A new 20-yr wave reforecast was generated based on the NOAA Global Ensemble Forecast System, version 12 (GEFSv12). It was produced using the same wave model setup as the NCEP's operational GEFSv12 wave component, which employs the numerical wave model WAVEWATCH III and utilizes three grids with spatial resolutions of 0.2° and 0.25°. The reforecast comprises five members with 1 cycle per day and a forecast range of 16 days. Once a week, it expands to 35 days and 11 members. This paper describes the development of the wave ensemble reforecast, focusing primarily on validation against buoys and altimeters. The statistical analyses demonstrated very good performance in the short range for significant wave height, with correlation coefficients of 0.95–0.96 on day 1 and between 0.86 and 0.88 within week 1, along with bias close to zero. After day 10, correlation coefficients fall below 0.70. We found that the degradation of predictability and the increase in scatter errors predominantly occur in the forecast lead time between days 4 and 10, in terms of the ensemble mean and individual members, including the control. For week 2 and beyond, a probabilistic spatiotemporal analysis of the ensemble space provides useful forecast guidance. Our results provide a framework for expanding the usefulness of wave ensemble data in operational forecasting applications.

KEYWORDS: Oceanic waves; Wind waves; Ensembles; Numerical weather prediction/forecasting

## 1. Introduction

Ensemble forecasts of winds and waves offer important advantages over deterministic forecasts, playing a pivotal role in safeguarding lives at sea and supporting maritime operations. Some applications that benefit from accurate forecasts include ship routing, towing and maintenance work on oil rigs, construction of underwater pipelines (Saetra and Bidlot 2004), and offshore renewable energy. An ensemble forecast involves generating multiple independent model integrations concurrently, adding perturbations to either initial conditions, model parameters, or forcing fields. Kalnay (2003) described two primary advantages of ensemble forecasts. First, the averaging of ensemble members tends to smooth out uncertain components, which leads to better skill than a single deterministic forecast. Second, the spread of the ensemble members provides information on the forecast uncertainty. Both are crucial elements in operational forecasting.

The superior performance of ensemble forecasts over deterministic forecasts is quantitatively demonstrated in numerous studies dedicated to surface wind and wave prediction. Janssen et al. (2002) and Saetra and Bidlot (2004) highlighted the advantages of employing the ECMWF ensemble prediction system (EPS) for waves and marine surface wind forecasting, relying on buoy and altimeter data. Using the NOAA Global Wave Ensemble System (GWES), Alves et al. (2013, 2015) and Campos et al. (2018, 2020a) investigated the global improvements achieved by the arithmetic ensemble mean (EnsMean) compared to the control member. It was observed that scatter errors, typically at 5 m $s^{-1}$ for strong winds (U10) at midlatitudes, decreased to 3 m $s^{-1}$ for the ensemble mean—directly contributing to the forecast skill of significant wave height (Hs). Likewise, Roh et al. (2021) reported improvements of approximately 18% in the root-mean-square error of Hs for extreme wave conditions with a 3-day lead time compared to that of the deterministic model.

Further global and local studies examining the advantages of ensemble wave forecasts are found in Farina (2002), Chang et al. (2017), Zieger et al. (2018), Bell and Kirtman (2019),

Denotes content that is immediately available upon publication as open access.

*Corresponding author*: Ricardo M. Campos, ricardo.campos@noaa.gov

DOI: 10.1175/WAF-D-24-0043.1

and Valiente et al. (2023). The importance and quality of ensemble wave forecast products during tropical cyclone conditions have been discussed by Tolman et al. (2005), Xu et al. (2007), Sampson et al. (2011), Lazarus et al. (2013a,b), Pan et al. (2016), Zieger et al. (2018), Roh et al. (2021), and Abdolali et al. (2020, 2021). Regional and coastal ensemble wave forecasts were addressed in Pallares et al. (2015), Pezzutto et al. (2016), and Behrens (2015), focusing on aspects such as spatial resolution and computational cost. Finally, regarding practical applications, Luo et al. (2023) demonstrated that ship routing and speed optimization based on ensemble wave forecasts hold greater potential than that based on deterministic forecasts, leading to reduced ship fuel consumption and greenhouse gas emissions—further emphasizing the expected significance of ensemble systems in the future.

Despite the aforementioned benefits associated with wave ensemble forecast products, biases persist within ensemble systems, as reported by Bunney and Saulter (2015). This is a well-known limitation that requires postprocessing bias correction algorithms (e.g., Cui et al. 2012; Zieger et al. 2018; Campos et al. 2020b). The development of such corrections relies on extensive datasets of forecasts and observations. Additionally, for ensemble forecasts to be useful, rigorous verification against observations spanning multiple years globally is essential. To address these requirements and others, we have produced a 20-yr global wave ensemble reforecast based on the NOAA Global Ensemble Forecast System, validated against buoys and altimeters. By "reforecast," we refer to running the forecast model and configuration used operationally, with minor differences, applied retrospectively to past conditions, initiating from the year 2000. Our goal is to evaluate and discuss the wave ensemble forecast performance globally, ranging from calm to extreme conditions, while considering the variation in forecast lead times and spread. A companion paper is in production with further statistical analyses to provide additional validation of the probabilistic wave forecast.

In section 2, we describe the NOAA Global Ensemble Forecast System, version 12 (GEFSv12). Section 3 delineates the methodology and explains how the reforecast has been produced, while section 4 focuses on the validation against observations. The last section, section 5, contains the final discussion. Information about data access and public repositories is given at the end. We expect that this information, in conjunction with the methodology and results outlined here, will allow readers to have complete access to our publicly available work and data, thereby contributing to the promotion of open science on a global scale.

## 2. The NOAA wave ensemble forecast

The evolution of the National Centers for Environmental Prediction (NCEP/NOAA) ensemble forecast over the past three decades is detailed in Zhou et al. (2022), Alves et al. (2013), and Alves et al. (2024). Campos et al. (2018, 2020a) analyzed 2 years of operational archives of the NCEP Global Wave Ensemble System, discussing the performance in deep waters. The most recent version of the NCEP/NOAA GEFSv12 was launched in September 2020, and it has undergone substantial changes, as described in Zhou et al. (2022), Hamill et al. (2022), and Alves et al. (2024).

The ensemble forecast operates four times daily, with cycles at 0000, 0600, 1200, and 1800 UTC, providing forecast guidance to the U.S. National Weather Service. The horizontal resolution is approximately 25 km, and the forecast range is 35 days (16 for ocean waves). The ensemble generates 31 members per forecast cycle (30 perturbed members plus the control member, all with the same resolution). The wave component is based on a recent version of WAVEWATCH III (WW3DG 2019). It employs the input and dissipation source term ST4 (Ardhuin et al. 2010), nonlinear interactions discrete interaction approximation (DIA) (Hasselmann and Hasselmann 1985), third-order propagation scheme (UQ) with garden sprinkler effect (GSE) alleviation (PR3), and simple ice blocking (IC0; Tolman 2003). The WAVEWATCH III source terms' parameters were optimized utilizing the Cylc workflow engine (Cyclops v1.0; Gorman and Oliver 2018), in conjunction with a large set of observations.

The wave model operates across three distinct grids: the Southern Ocean (1/3°), Arctic Ocean (1/3°), and global core (1/4°). Alves et al. (2024) describe the integrated one-way coupling scheme, where the model obtains wind forcing from the atmospheric component every 1 h. The wave forecast outputs on a final single grid (0.25° × 0.25°). The spectral resolution of the wave model consists of 33 frequencies with wave periods ranging from 1.35 to 28.57 s and 36 wave directions. Unlike the 35-day atmospheric forecast range, the operational wave forecast system is limited to 16 days. Details of the optimization approach and values of optimized source-term parameters, resolution, and configuration of WAVEWATCH III utilized in GEFSv12 are available in Alves et al. (2024).

A reforecast product, as described in the next section, mimics the forecast system used operationally, with minor modifications (usually a smaller number of ensemble members) to reduce computational cost. The main difference between a reforecast and a reanalysis is that the reforecast preserves the forecast lead time and consecutive cycle runs. It holds two time dimensions: cycle time and forecast time, operating as a forecast simulation retrospectively. In contrast, a reanalysis holds only one time dimension, allowing data to be assimilated throughout the entire simulation. Therefore, analyzing the evolution of forecast error with lead time, which is important in operational forecasting, is only possible with reforecast data and not with reanalysis data.

## 3. Numerical simulations and reforecast features

Hamill et al. (2022) produced a 20-yr reanalysis dataset, not including the wave component, which served as the initial conditions to produce the GEFSv12 atmospheric reforecast. Hamill et al. (2022) explain that the reanalysis assimilates most of the observations used in the operational data assimilation system for initializing global predictions. This dataset spans from 2000 to 2019 and employs the same numerical

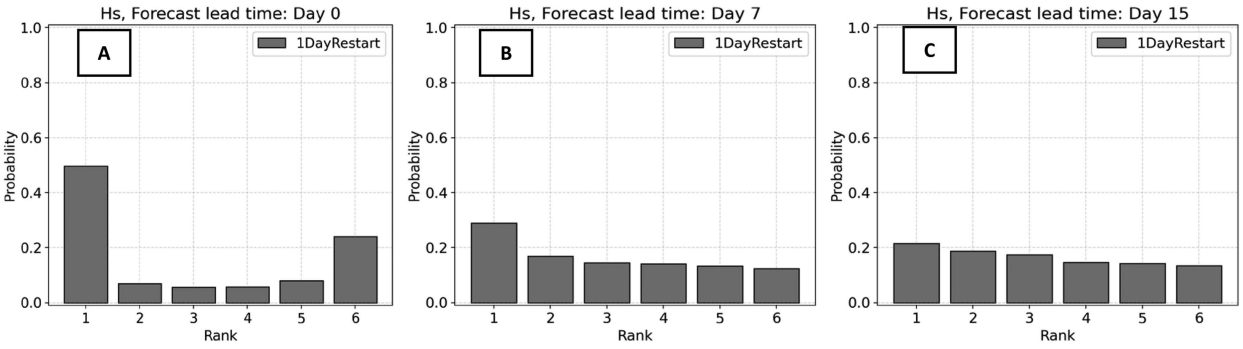


FIG. 1. Rank histograms for the initial experiment, from 24 Aug to 18 Oct 2016, of the GEFSv12 wave reforecast. The plots were generated using WAVEWATCH III point outputs worldwide, using the 24-h lead time of the preceding cycle. The rank histograms were calculated for different time leads: (a) the analysis, (b) 7-day forecast, and (c) 15-day forecast.

model and configuration as the operational atmospheric component. The primary difference from the operational system lies in the number of ensemble members and cycles per day. The ensemble reforecast comprises five members with 1 cycle per day, providing a forecast range of 16 days. Once a week (on Wednesdays), it extends to 35 days with 11 members. A recent validation conducted by Hamill et al. (2022) indicates that the quality of the GEFSv12 reanalysis is generally superior to that of NOAA's previous generation Climate Forecast System Reanalysis (CFSR; Saha et al. 2010).

In this paper, we describe the development of the wave reforecast forced by GEFSv12 atmospheric reforecast initialized by the reanalysis by Hamill et al. (2022). The wave model, source terms, parameters, grids, and resolution selected for the reforecast are the same as the operational GEFSv12 wave component described in Alves et al. (2024), as summarized in the previous section. The wave reforecast characteristics, in terms of ensemble size and forecast range, follow the atmospheric reforecast—expanding the wave forecast range to 35 days, which is not found in the operational implementation (restricted to 16 days). The wave model was run for the period from 1 January 2000 to 31 December 2019, keeping the three grids and output fields with a spatial resolution of 0.25° × 0.25° while adopting 3-h resolution for the grid outputs and 1-h resolution for the point outputs.

Initial simulations and tests were conducted for the period from 24 August to 18 October 2016, covering several meteorological systems from calm to extreme events, including Hurricane Matthew (category 5) and Hurricane Nicole (category 4). The validation was then expanded to cover the 20-yr period, presented in the next section. During these initial simulations, an important question emerged concerning the initial conditions (ICs) for WAVEWATCH III. Typically, the IC for each member in the wave ensemble uses a short-range forecast from a previous run of the same member in order to retain spread in initial conditions (Bunney and Saulter 2015). However, on Wednesdays, when the reforecast expands from 5 to 11 members, these additional six members do not have corresponding forecasts from the previous run to obtain an IC. Moreover, the initial simulations of the five members, using ICs generated from the previous cycle +24 h, showed a small spread in the short-term forecasts (Fig. 1).

The rank histograms of Fig. 1 show a small bias and indicate overconfidence of the wave ensemble in the first-day forecast lead time—suggested by the U-shape format of the plot. Hamill (2001) provided detailed information on the interpretation of rank histograms for verifying ensemble forecasts. Moving to the 7- and 15-day forecast (Figs. 1b,c), there is a noticeable increase in spread, appearing more suitable with flatter histograms. Consequently, selecting more advanced lead times for the IC might enhance the spread in short-term forecasts. On the other hand, this could also introduce larger scatter errors impacting the forecast quality and potentially deteriorating the initial conditions. Therefore, it is crucial at this stage to analyze the spread's evolution as a function of the forecast lead time as well as the influence of the wind input on the wave ensemble.

To investigate the spread in IC and identify the optimal approach for generating ICs for the reforecast, eight experiments were conducted. The initial five tests obtain ICs from forecasts from previous consecutive cycles, resulting in initial conditions with 1-, 2-, 3-, 5-, and 7-day lead times, progressively increasing the spread in the IC. Subsequently, test 6 was run by starting the model from rest, evaluating the propagation of the initial error throughout the forecast range. Test 7 applied the same IC (control member) for all the forecast members, starting with no spread and allowing the atmospheric spread to propagate to the wave field. Finally, test 8, but applied the same wind forcing (control member) for all the members. This test starts with a small spread that is progressively decreased to zero due to the lack of spread from the wind forcing.

The experiments were validated against NOAA/National Data Buoy Center (NDBC) and Copernicus buoys (Fig. 2), depicting the spread, the root-mean-square error (RMSE) of the ensemble mean, and the continuous ranked probability score (CRPS). Figure 2a shows the increase of the ensemble spread by using longer forecast lead times to generate the ICs (time lagging). However, this inflated spread diminishes rapidly within the first day, and by the third forecast day, the experiments converge (Fig. 2a). This suggests that the ensemble

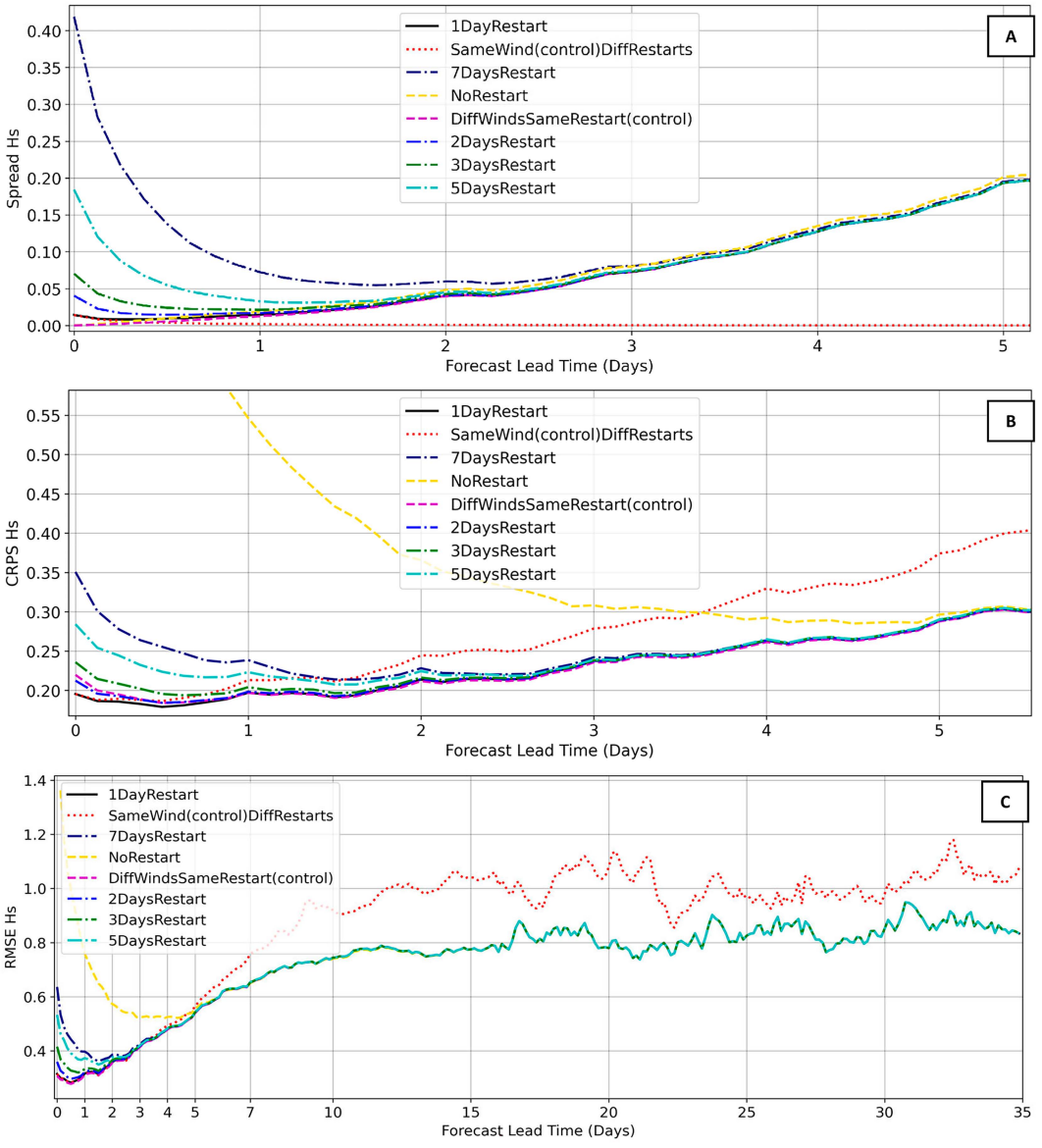


FIG. 2. Results of the IC experiments for the period from 24 Aug to 18 Oct 2016, validated against NDBC and Copernicus buoys in deep waters. Eight IC tests were analyzed to decide the best strategy to generate the WAVEWATCH III restart files for the GEFSv12 wave reforecast. The results are presented in terms of (a) spread, (b) CRPS, and (c) RMSE.

spread induced by the wind input plays a much more important role in the midrange and long range than the ICs in the wave model. More importantly, Figs. 2b and 2c reveal that artificially boosting the spread through time lag leads to higher RMSE and CRPS, compromising the ensemble's performance in the short range.

The results from the IC experiments collectively suggest that the impact of the initial condition in the wave ensemble forecast beyond 1 week is minimal, while the skill and spread of the atmospheric ensemble (wind inputs) are the most important features, as highlighted in Campos et al. (2023). Based on these conclusions, we decided to retain the shortest forecast from the previous run (+24 h) to produce the ICs for consecutive days. When the ensemble expands from 5 to 11 members, once a week, the five initial conditions from the preceding day are selected and randomized to force the additional six members.

The simulations were run on the Orion supercomputer equipped with multiple CPUs (2.4 GHz Intel Xeon Gold 6148 Skylake), each of those containing 20 cores, for a total of 40 cores per node, alongside 192 GB of memory (128 GB allocated for WAVEWATCH III simulations). Each job submission used 10 nodes (400 cores), processing seven consecutive cycles within an 8-h time frame. This configuration of nodes,

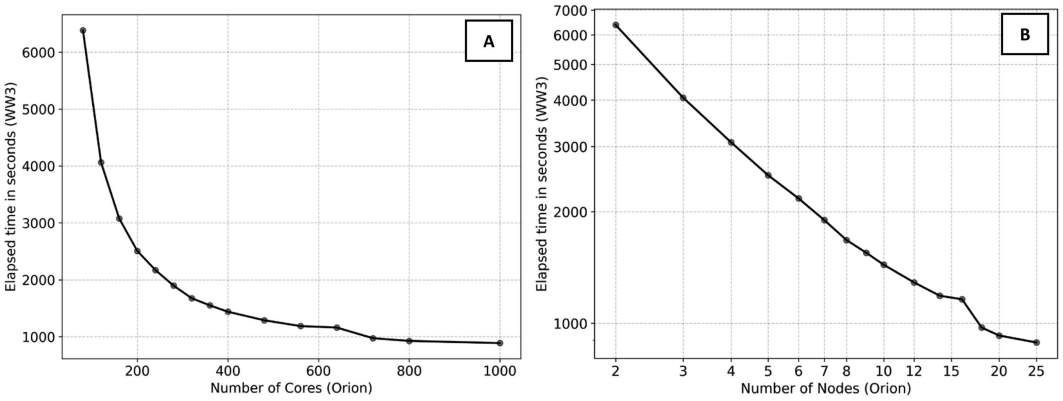


FIG. 3. Scalability test of WAVEWATCH III simulations of GEFSv12 reforecast using a different number of nodes, from 2 to 25, each one containing 40 cores.

cores, and reforecast cycles per job was achieved after conducting various tests, illustrated in the scalability plot of Fig. 3. The parallel computing was handled by the primary WAVEWATCH III executable ww3_multi, while pre- and postprocessing tasks were executed separately in a serial fashion. The 20 years of reforecast were split into four streams of 5 years, processed concurrently. The first month of the reforecast, January 2000, was used to spin up the model and is excluded from validations.

The wave variables in the field outputs were specifically chosen to maximize the wave information and usability while minimizing storage requirements, as outlined in Table 1. The field outputs were saved in grib2 (NCEP WMO 2023) format (gefs.wave.YYYYMMDD.EM.global.0p25.grib2), containing all the variables in the same file. A total of 658 point outputs have been selected, aligning with exact locations where observations are available (Fig. 4), and merged with the existing Global Forecast System (GFS) and GEFS point outputs used operationally. This blend actually led to 761 points, from which 658 correspond to wet/valid points on the GEFSv12 wave grid. Two formats of point outputs have been generated: a simple table containing the main wave parameters and the complete 2D wave spectrum, both saved in netCDF format (gefs.wave.YYYYMMDD.EM.tab.nc and gefs.wave.YYYYMMDD.EM.spec.nc).

TABLE 1. Wave variables of field outputs generated for the 20-yr GEFSv12 wave reforecast. The partitioned variables are composed of one wind-sea partition and three swell partitions.

| | |
|---|---|
| WND | 10-m wind speed |
| HS | Significant wave height |
| FP | Peak frequency |
| T01 | Mean wave period (Tm0, 1) |
| T02 | Mean wave period (Tm0, 2) |
| DIR | Mean wave direction |
| DP | Peak wave direction |
| SPR | Mean directional spread |
| PHS | Partitioned wave heights |
| PTP | Partitioned peak period |
| PDIR | Partitioned mean direction |

The GEFSv12 wave reforecast occupies a total disk space of 100T and is stored in the automatic weather station (AWS) cloud service, accessible publicly with web links provided at the end of this paper. The three output file formats, associated with global field outputs, point output (table), and point output (spectrum), accompanied by their respective sizes for one single forecast cycle are as follows:

- gefs.wave.YYYYMMDD.ENSM.global.0p25.grib2: 1.6G (16 days)/3.4G (35 days).
- gefs.wave.YYYYMMDD.ENSM.tab.nc: 8.3M (16 days)/19M (35 days).
- gefs.wave.YYYYMMDD.ENSM.spec.nc: 493M (16 days)/1.1G (35 days).

An automatic quality check was built and applied during the reforecast production, detecting issues throughout the simulations. Moreover, a visualization and validation tool, WW3-tools (Campos et al. 2022a), was developed in Python to facilitate the postprocessing of WAVEWATCH III data as well as altimeter and buoy data. Examples can be seen in Figs. 5 and 6. The automatic quality check flagged 35 cycles/members with problems, prompting their removal and subsequent rerun. Alongside the automated verification, a visual assessment of the results was conducted using a combination of plots in a panel format (Fig. 6) designed for detecting inconsistencies. Despite the time-consuming nature of the visual inspection, it successfully identified 8 cycles per member with unrealistic results, illustrated in Fig. 6, which were then deleted and rerun to build the final 20-yr reforecast dataset.

The initial assessments and quality checks of GEFSv12 wave reforecast files allowed for the identification of several interesting characteristics in the ensemble forecast. One example is illustrated in Fig. 7, showing an extreme event associated with an extratropical low and large wind fetch that resulted in waves up to 9 m of significant wave height (Hs). In the short range, within the first 5 days, Fig. 7b showcases a good agreement between forecast and observations for this event. However, the same event for a forecast lead time

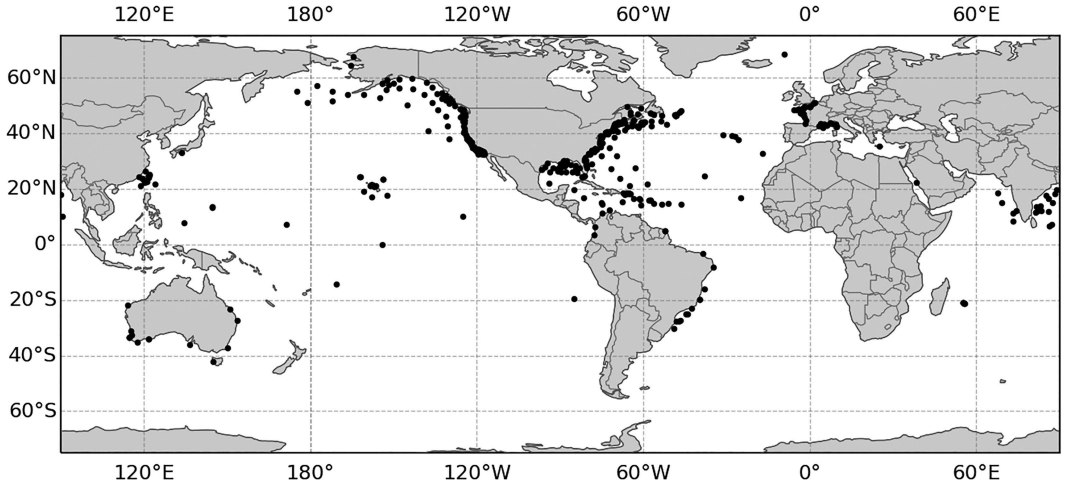


FIG. 4. Position of the point outputs of the GEFSv12 wave reforecast.

of 9 days, i.e., moving backward in the forecast cycles (Fig. 7a), reveals both the control member and the ensemble mean failed to accurately represent the peak of the storm—exemplifying a typical loss of predictability over time. Despite the poor performance of the ensemble mean and the control member nearly 10 days in advance of the event, the ensemble members between days 7 and 14 (week 2 forecast window) indicated a highly active period, which can be analyzed probabilistically to provide valuable forecast information in the long range. The scatterplots in Fig. 8 also illustrate the increased scatter error and loss of predictability in time. These aspects will be quantitatively analyzed in the following section, covering a 20-yr period.

## 4. Reforecast validation against satellite and buoy data

Validation of the GEFSv12 wave ensemble reforecast was performed against buoy and altimeter data, independently. The analyzed variables include wind speed at a 10-m height (U10), significant wave height (Hs), and mean wave period (Tm, from the buoy data only), with a primary focus on Hs. The spatial resolution and model optimization of GEFSv12 were not designed for coastal waters, potentially compromising forecast accuracy, as discussed by Chang et al. (2017) and Valiente et al. (2023). Similarly, altimeter data also exhibit limited accuracy in shallow waters near the coastline. To address these limitations, a grid mask was constructed to exclude coastal areas based on two criteria: water depth and proximity to the coast—following the validation methodology of Ribal and Young (2019) and Campos et al. (2020a, 2020c). Therefore, for the reforecast validation against both altimeters and buoy data, a minimum water depth of 80 m and a distance of 50 km to the nearest coast were imposed. These criteria were applied using NGDC/NOAA 1-min gridded elevations/bathymetry for the world (ETOPO1; Amante and Eakins 2009)

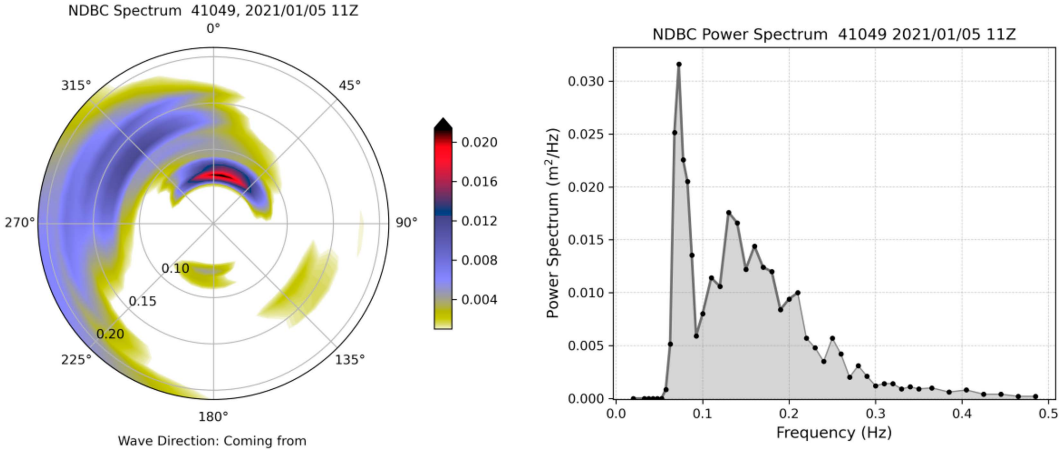


FIG. 5. Example of WW3-tools visualization. The plots show a wave spectrum measured on 5 Jan 2021 from NDBC buoy 41 049. The same visualization can be applied to WAVEWATCH III spectra.

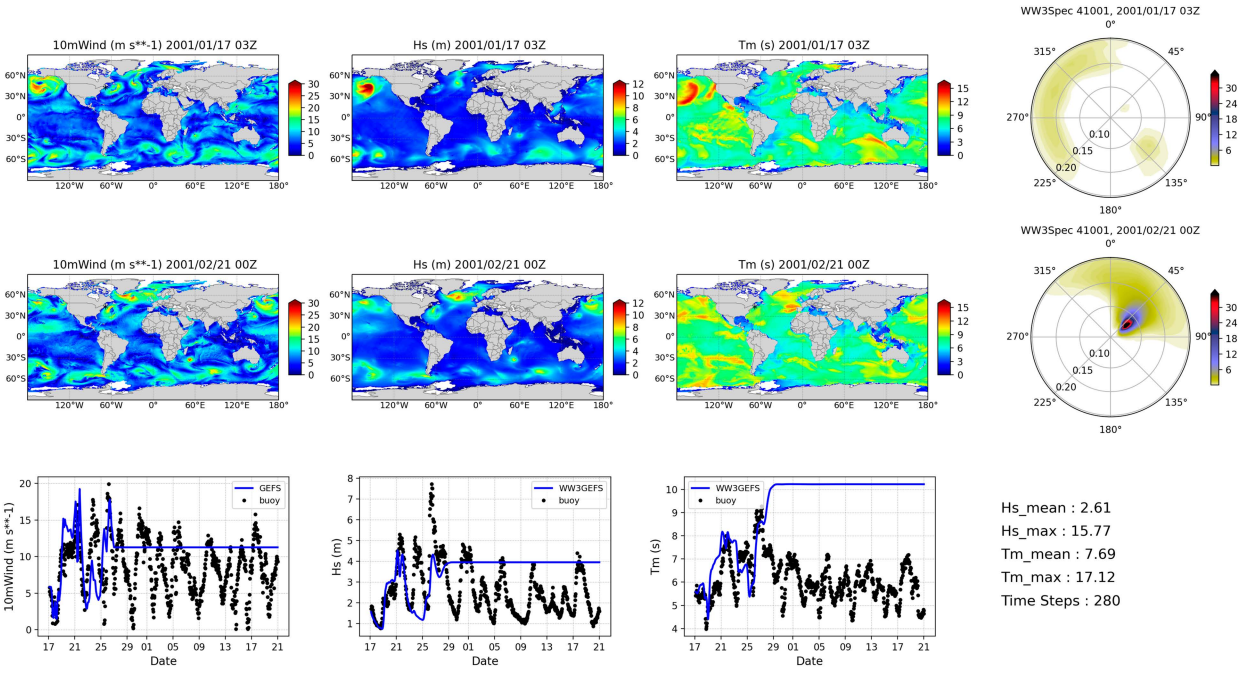


FIG. 6. Quality check panel utilized to inspect GEFSv12 wave reforecast files, analyzing one forecast cycle per figure. The panel plots (top) the first time step, (middle) the last forecast lead time, and (bottom) a time series for a predefined point (NDBC point 41001 in this case). The two directional spectra refer to this point. At the bottom right, basic statistics are calculated for the entire globe and cycle, also showing the number of steps available. This figure is an example of an error in the reforecast that could be captured by the visual inspection that was not detected in the automatic quality check.

and the information of distance to the coast from GSFC/NASA (0.04°), respectively.

One of our primary objectives is to assess the wave forecast's performance as a function of the forecast lead time, particularly advancing into longer forecast ranges. To expand the analysis on the lead time, we excluded the reforecast cycles limited to 16 days, focusing on forecast ranges covering 35 days containing 11 ensemble members. The validation statistics were inspired by the studies of Zhu and Toth (2008) and Willmott et al. (1985), and the error metrics were selected based on Mentaschi et al. (2013), who discuss the advantages of interpreting the systematic and scatter components (SCs) of the error separately. Furthermore, Mentaschi et al. (2013) recommend computing an additional metric (HH; Hanna and Heinold 1985) to address issues stemming from low values of RMSE and scatter index (SI), which can affect the statistics. A total of eight metrics [Eqs. (1)–(8)] were computed for the validation, where $x$ is the observation (altimeter or buoy data) and $y$ is the reforecast. The overbar in the following equations represents the arithmetic mean:

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

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

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

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

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

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

$$\text{HH} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i)^2}{\sum_{i=1}^{n} y_i x_i}}, \tag{7}$$

$$\text{CC} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}}. \tag{8}$$

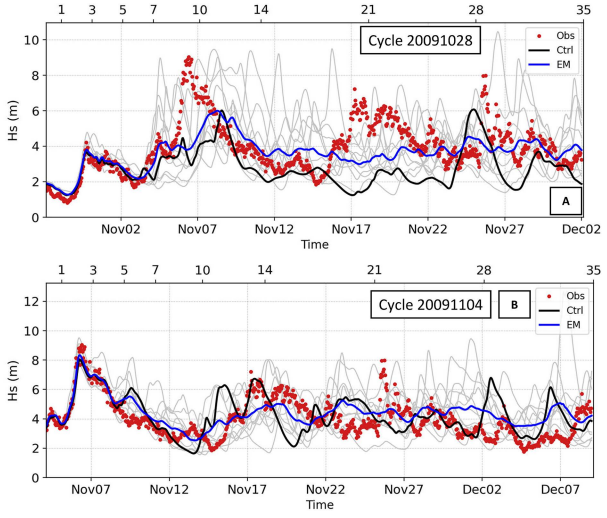


FIG. 7. Example of the GEFSv12 wave ensemble reforecast in the Pacific Ocean, for a position at 40.764°N, 137.377°W, compared with NDBC buoy measurements. The plots show a time series of significant wave height (Hs; m) of two forecast cycles where the date (month day) is presented in the bottom *x* axis and the forecast lead time (days) is presented in the top *x* axis. In the left-hand part of the graphics, it is expected better accuracy, whereas the right-hand portion of the plot shows larger scatter errors associated with more advanced lead time. The observations are shown in red, the control member is shown in solid black, the arithmetic EnsMean of the 11 members is shown in solid blue, and the individual members are shown in gray.

Those metrics as well as the following validation plots presented have been produced using WW3-tools.

### *a. Validation against buoy data*

The validation was performed using NDBC and Copernicus datasets, paired with WAVEWATCH III point outputs, creating data arrays with hourly resolution. After excluding coastal waters, the remaining data were quality controlled (NDBC 2015), and the buoys with longer and more consistently high-quality data, covering at least 1 year, were chosen for validation. This dataset comprises a total of 98 buoys in deep waters (Fig. 9b) which, over the 20-yr period, resulted in 5 044 272 observation–model matchups selected for statistical analyses.

The initial comparisons between reforecast and buoy data are outlined in Tables 2 and 3. They demonstrate very similar mean values of Hs between GEFSv12 and the observations, indicating an overall good calibration of the model. However, the ensemble mean, particularly for week 2 and beyond, presents a slight overestimation of Hs that will be further investigated. The variance, on the other hand, displays larger differences, with the buoy data showing higher variance than the reforecast for Hs. The differences become more notable for the ensemble mean from week 2 onward, where the reforecast severely underestimates the expected variance (Table 2) of the buoy data. This suggests that the underdispersed signal of Hs from GEFSv12 overestimates small waves and underestimates larger seas. The lower values of GEFSv12 95th and 99th percentiles align with this initial hypothesis.

Unlike Hs, the wave period, Tm, presented in Table 3, shows a large overestimation by the wave model, which also extends to the upper percentiles. The variance of Tm is consistently higher than the observations, impacting the ensemble mean of Tm for weeks 2–5. Tables 2 and 3 clearly highlight the forecast limitations beyond 7 days, showing a critical underestimation effect in the higher percentiles by the ensemble mean, smoothing out the severe events. The underestimation's magnitude for Hs reaches 25%–40% for the 99th and 99.9th percentiles, respectively.

The wave model's ability to represent observations, from calm to severe conditions, can be visualized using quantile–quantile plots (QQ plots). The initial plots in Fig. 10 demonstrate the great performance of GEFSv12 within the first 24-h forecast lead time regarding Hs, confirming the successful wave model's optimization. Both the control member and ensemble mean closely align with the main diagonal. The high quality of reforecast data and operational forecast archives is also discussed in Breivik et al. (2013) and Meucci et al. (2018). They argue that these sources offer better alternatives to traditional reanalysis, particularly for extreme value analysis (EVA) applications—an essential requirement within the

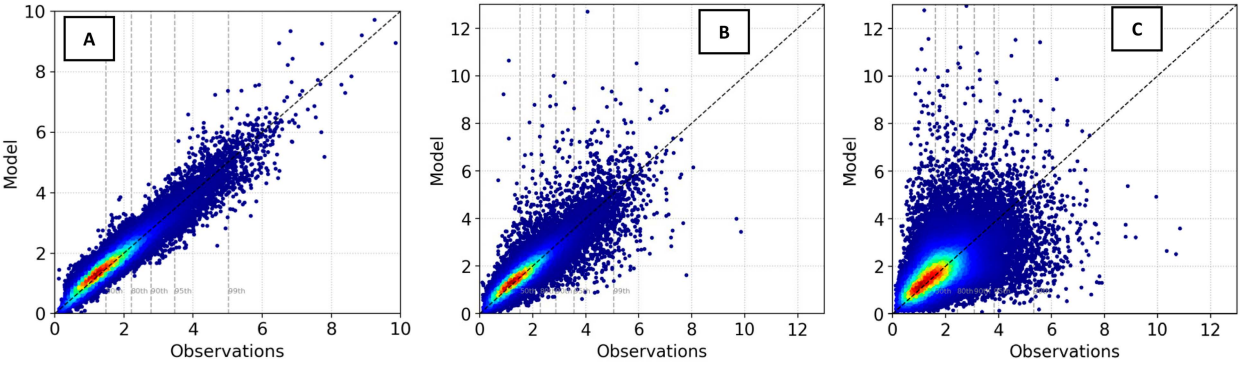


FIG. 8. Scatterplots of significant wave height (Hs; m) comparing the GEFSv12 wave reforecast (control member) point outputs with NDBC observation from 24 Aug to 18 Oct 2016. The plots display different lead time intervals: (a) day 1, (b) week 1, and (c) week 2.

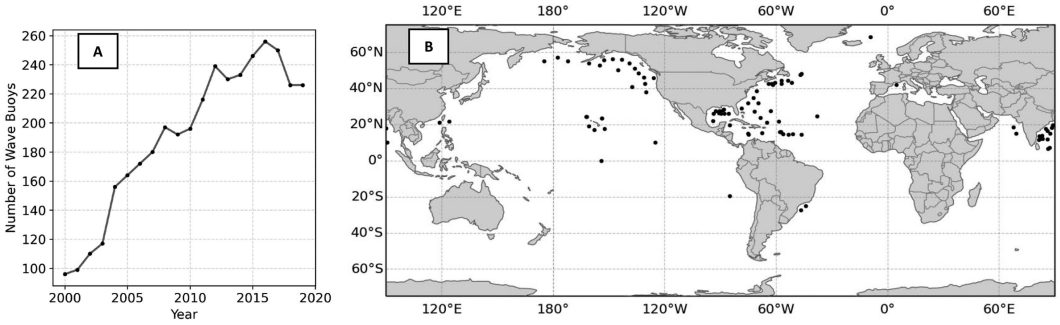


FIG. 9. Information on buoy data used for the reforecast validation. (a) Number of buoys available for each year, from 2000 to 2020. (b) Deep water buoys selected for validation.

marine industry. Nevertheless, while the quality of Hs remains, Tm shows an increasing overestimation, indicating the need for further analysis of the entire wave spectrum.

The decline in predictability over time is demonstrated through the progression shown in the QQ plots and scatterplots of Fig. 11. Initially, within the first week, the forecast agrees very well with observations (Fig. 11a) although some scattering is noticeable (Fig. 11e). However, as we move to week 2 and beyond, the ensemble mean starts to overestimate smaller waves while underestimating higher percentiles. For week 3 onward, only a few reforecast values of Hs exceed 6 m in the scatterplots (Figs. 11g,h), with most data falling below 5 m. In summary, while the scatter points of the control member appear symmetrically distributed around the main diagonal, the scatter points of the ensemble mean become increasingly clustered toward lower values of Hs on the *y* axis (reforecast). This feature is critical for operational applications and constrains the usefulness of the ensemble mean beyond week 2.

It is important to note that unlike scatterplots that directly display the matchups, QQ plots operate in the probabilistic domain, as the quantile function is the inverse of the cumulative distribution function. This aspect of QQ plots disregards differences in phase between observation and model signals, which are expected to diverge as forecast leads extend. In line with findings from Breivik et al. (2013), the validation results indicate that as the forecast lead time extends, the QQ plot of the control member continues to follow the main diagonal, whereas its scatter error rapidly increases. This leads to a deterioration of the ensemble mean, resulting in a curved and sloped QQ plot.

The assessment results using the error metrics from Eqs. (1) to (8) are presented in Tables 4 and 5. The results for day 1, considering forecast slices from 0 to 24 h, confirm the outstanding performance of GEFSv12 concerning Hs. The bias ranges from −1 to 3 cm only, accompanied by very high correlation coefficients (CCs) of 0.95. While there is room for improvement in the SI at 15%, the RMSE is reasonably low at 0.37. Moving from day 1 to week 1, the bias remains close to zero. As expected, there is a decrease in the correlation coefficient and an increase in scatter errors, yet the performance remains well preserved, with a CC of 0.86–0.88. Indeed, from week 1 onward, the benefits of the ensemble mean compared to the control member become evident, showing higher CC and reduced scatter errors. This reaffirms the advantage of ensemble forecasts over deterministic forecasts, as explained by Kalnay (2003) and echoed in previous NOAA wave ensemble validations (Alves et al. 2013; Campos et al. 2018, 2020a).

The most substantial degradation in wave forecasts, significantly impacting error metrics, occurs when transitioning from week 1 to week 2. The control member is the most affected, with CC dropping to 0.57 and scatter errors rising to 45%. Although the ensemble mean maintains better CC and

TABLE 2. Comparison between buoy observations and wave reforecast data for significant wave height (Hs; m), illustrating the control member and arithmetic EnsMean. The first three probabilistic moments are presented, followed by the 95th, 99th, and 99.9th percentiles. The results are categorized in forecast lead intervals, ranging from weeks 1 to 5 and covering a 35-day forecast range. The observations are in bold.

| | | Mean | Variance | Skewness | pctl95 | pctl99 | pctl99.9 |
|---|---|---|---|---|---|---|---|
| Hs | Obs | **2.11** | **1.43** | **1.64** | **4.43** | **6.25** | **8.80** |
| Week 1 | Control | 2.11 | 1.32 | 1.81 | 4.34 | 6.19 | 8.92 |
| | EnsMean | 2.16 | 1.24 | 1.63 | 4.36 | 6.01 | 8.37 |
| Week 2 | Control | 2.13 | 1.35 | 1.78 | 4.41 | 6.27 | 8.85 |
| | EnsMean | 2.20 | 0.89 | 1.17 | 4.07 | 5.18 | 6.48 |
| Week 3 | Control | 2.12 | 1.34 | 1.80 | 4.40 | 6.24 | 8.84 |
| | EnsMean | 2.19 | 0.73 | 0.94 | 3.89 | 4.70 | 5.55 |
| Week 4 | Control | 2.11 | 1.35 | 1.83 | 4.41 | 6.27 | 8.82 |
| | EnsMean | 2.18 | 0.71 | 0.92 | 3.87 | 4.62 | 5.35 |
| Week 5 | Control | 2.12 | 1.37 | 1.82 | 4.41 | 6.32 | 8.96 |
| | EnsMean | 2.18 | 0.72 | 0.92 | 3.89 | 4.61 | 5.38 |

TABLE 3. Comparison between buoy observations and wave reforecast data for mean wave period (Tm; s), illustrating the control member and arithmetic EnsMean. The first three probabilistic moments are presented, followed by the 95th, 99th, and 99.9th percentiles. The results are categorized in forecast lead intervals, ranging from weeks 1 to 5 and covering a 35-day forecast range. The observations are in bold.

| | | Mean | Variance | Skewness | pctl95 | pctl99 | pctl99.9 |
|---|---|---|---|---|---|---|---|
| Tm | Obs | **6.21** | **2.06** | **1.01** | **8.85** | **10.59** | **13.00** |
| Week 1 | Control | 7.77 | 4.32 | 0.54 | 11.52 | 13.47 | 15.64 |
| | EnsMean | 7.86 | 4.12 | 0.51 | 11.50 | 13.39 | 15.50 |
| Week 2 | Control | 7.85 | 4.48 | 0.52 | 11.69 | 13.62 | 15.65 |
| | EnsMean | 7.96 | 3.49 | 0.27 | 11.19 | 12.65 | 14.23 |
| Week 3 | Control | 7.89 | 4.57 | 0.48 | 11.72 | 13.62 | 15.66 |
| | EnsMean | 7.99 | 3.09 | −0.02 | 10.83 | 11.84 | 12.86 |
| Week 4 | Control | 7.88 | 4.57 | 0.48 | 11.71 | 13.59 | 15.67 |
| | EnsMean | 7.99 | 3.01 | −0.10 | 10.72 | 11.62 | 12.52 |
| Week 5 | Control | 7.89 | 4.66 | 0.47 | 11.77 | 13.60 | 15.60 |
| | EnsMean | 7.99 | 3.08 | −0.10 | 10.75 | 11.62 | 12.42 |

lower scatter errors, it comes at the expense of increased positive bias. Breivik et al. (2013) also investigated ensemble forecasts at extended lead times, reporting significant errors beyond 10 days with low CC of Hs (0.33) and large SI (0.68). The decline in predictability persists into week 3, revealing very large systematic and scatter errors. The validation results for weeks 4 and 5 are very similar to those for week 3, indicating a progressive increase in forecast errors stabilizing at week 3, with a pronounced lack of skill thereafter. Campos et al. (2024) describe a spatiotemporal methodology for generating probability maps that are more suitable for extended forecast ranges.

The error metrics for Tm, outlined in Table 5, indicate a weaker performance compared to Hs. The CC starts for day 1 with 0.77 and a scatter index above 20%. The large positive bias observed in Tm, presented in the previous plots, is confirmed by Table 5, which increases even more with the forecast lead time. Although the ensemble mean offers advantages over the control member, the model's reduced skill for Tm from the short-term range onward renders this advantage somewhat negligible when considering the overall larger errors.

Apart from bias, the results from Tables 2 to 5 quantitatively demonstrate the superior performance of the ensemble mean in comparison to the control member, which is more evident for week 2 onward. Nevertheless, the QQ plots in Fig. 11 highlight the problems of the ensemble mean in longer forecast ranges. Hence, we recommend utilizing the ensemble mean preferably within week 1 and possibly week 2 (excluding extreme events). This will become even more evident in the next section through the spatial validation using altimeter data. A more effective approach to visualize this effect is by combining various statistics into a single diagram. The Taylor diagrams (Taylor 2001) shown in Fig. 12 illustrate this interesting evolution of the ensemble mean compared to the control member as the forecast lead time progresses. While the ensemble mean exhibits better correlation coefficients, it consistently displays much lower standard deviations than the observations in longer forecast ranges. This trend is also evident in the time series of Fig. 7a for horizons extending beyond 10 days.

To better investigate the growth curves of the forecast error as a function of the forecast lead time, the statistical metrics were recalculated using 24-h segments from day 1 to day 35

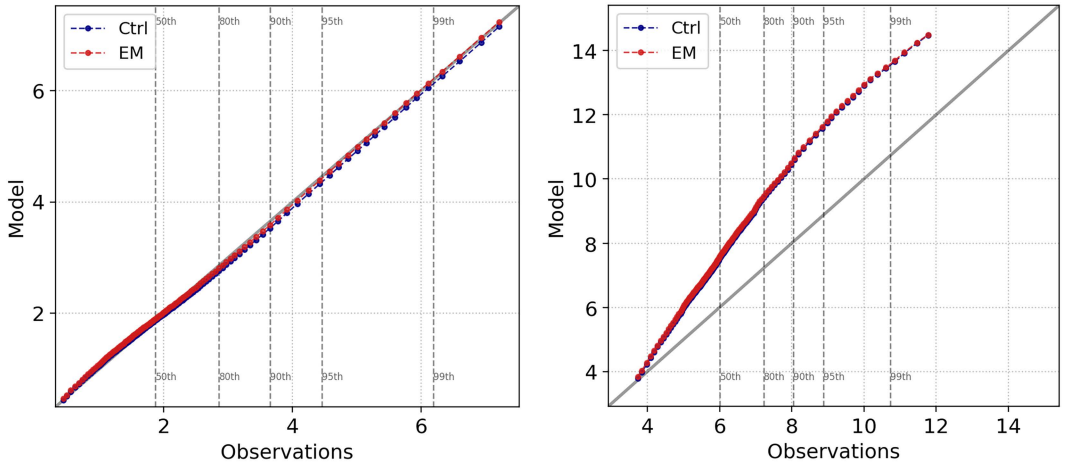


FIG. 10. QQ plots of (left) Hs and (right) Tm comparing the GEFSv12 wave reforecast with buoy observations. The plots were generated using time steps within day 1. The control member is depicted in blue, while the arithmetic EnsMean is shown in red.

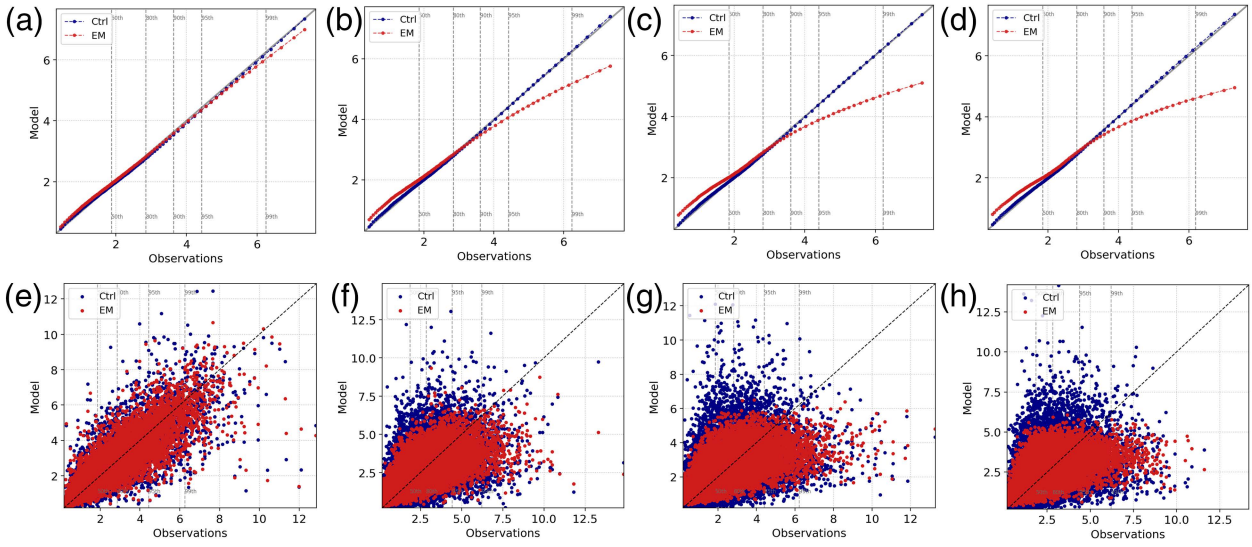


FIG. 11. Evolution of (top) QQ plots and (bottom) scatterplots of significant wave height (Hs; m) over forecast time, spanning from weeks 1 to 5. The results illustrate the comparison between the GEFSv12 wave reforecast and buoy data. The control member is shown in blue, while the arithmetic EnsMean is shown in red.

(Figs. 13 and 14). The correlation coefficient and HH highlight the point where the ensemble mean and control member begin to diverge progressively. In the initial 3 days, both exhibit quite similar results (refer to Fig. 7 for a practical example), but after day 3, the ensemble mean notably outperforms the control. Specifically, the GEFSv12 reforecast's ensemble mean maintains a CC above 0.9 for the first 4 days and above 0.8 for the initial 7 days for Hs, confirming the good quality of GEFSv12 in week 1. The RMSE below 0.6 within the first 5 days reinforces this conclusion.

The CRPS is a reliable tool that considers the entire ensemble forecast distribution, taking into account all members rather than relying solely on a single estimate (control) or the arithmetic mean (EM). The CRPS calculated for Hs is presented in Fig. 13c. By considering CC, HH, and CRPS altogether, it becomes evident that the loss of predictability predominantly happens between days 4 and 10. Beyond the 15-day mark, the error metrics exhibit relative stability, affirming the findings observed in Tables 4 and 5.

The bias plots (Figs. 14a,c) for both Hs and Tm show a tendency of gradual increase within the initial 10 days, especially the ensemble mean, presenting a positive bias (model overestimation) after day 7, being especially higher for weeks 2–5. These results align with Zhu et al. (2018), who evaluated the atmospheric forecast from GEFSv11, focusing on weeks 3 and 4. They reported a systematic bias while emphasizing the need for further calibration and bias correction when advancing to the subseasonal time scale. The magnitude of bias in Fig. 14 is relatively small for Hs but notably higher for Tm, reaching almost 2 s beyond the 10-day forecast period. Such a high positive bias of Tm directly affects the RMSE, as indicated in Fig. 14c.

TABLE 4. Results of the GEFSv12 wave reforecast validation against wave buoys, for significant wave height (Hs; m). The table shows eight error metrics calculated using Eqs. (1)–(8). The statistics are divided into forecast lead intervals, including day 1 and weeks 1–5.

| Hs | | Bias | RMSE | Nbias | NRMSE | SCrmse | SI | HH | CC |
|---|---|---|---|---|---|---|---|---|---|
| Day 1 | Control | −0.01 | 0.37 | −0.01 | 0.15 | 0.37 | 0.15 | 0.15 | 0.95 |
| | EnsMean | 0.03 | 0.37 | 0.01 | 0.15 | 0.37 | 0.15 | 0.15 | 0.95 |
| Week 1 | Control | 0.00 | 0.62 | 0.00 | 0.26 | 0.62 | 0.26 | 0.26 | 0.86 |
| | EnsMean | 0.05 | 0.57 | 0.03 | 0.23 | 0.57 | 0.23 | 0.24 | 0.88 |
| Week 2 | Control | 0.03 | 1.09 | 0.01 | 0.45 | 1.09 | 0.45 | 0.48 | 0.57 |
| | EnsMean | 0.10 | 0.87 | 0.05 | 0.36 | 0.86 | 0.36 | 0.37 | 0.70 |
| Week 3 | Control | 0.02 | 1.23 | 0.01 | 0.51 | 1.23 | 0.51 | 0.55 | 0.45 |
| | EnsMean | 0.10 | 0.94 | 0.05 | 0.39 | 0.94 | 0.39 | 0.41 | 0.62 |
| Week 4 | Control | 0.02 | 1.25 | 0.01 | 0.52 | 1.25 | 0.52 | 0.56 | 0.44 |
| | EnsMean | 0.10 | 0.95 | 0.05 | 0.40 | 0.95 | 0.40 | 0.42 | 0.61 |
| Week 5 | Control | 0.04 | 1.26 | 0.02 | 0.53 | 1.26 | 0.53 | 0.57 | 0.42 |
| | EnsMean | 0.11 | 0.96 | 0.05 | 0.40 | 0.95 | 0.40 | 0.42 | 0.60 |

TABLE 5. Results of the GEFSv12 wave reforecast validation against wave buoys, for mean wave period (Tm; s). The table shows eight error metrics calculated using Eqs. (1)–(8). The statistics are divided into forecast lead intervals, including day 1 and weeks 1–5.

| Tm | | Bias | RMSE | Nbias | NRMSE | SCrmse | SI | HH | CC |
|---|---|---|---|---|---|---|---|---|---|
| Day 1 | Control | 1.56 | 2.05 | 0.25 | 0.32 | 1.34 | 0.21 | 0.29 | 0.77 |
| | EnsMean | 1.62 | 2.10 | 0.26 | 0.33 | 1.34 | 0.21 | 0.29 | 0.77 |
| Week 1 | Control | 1.56 | 2.11 | 0.25 | 0.33 | 1.42 | 0.22 | 0.30 | 0.73 |
| | EnsMean | 1.65 | 2.14 | 0.27 | 0.34 | 1.37 | 0.21 | 0.30 | 0.74 |
| Week 2 | Control | 1.65 | 2.41 | 0.27 | 0.38 | 1.76 | 0.28 | 0.34 | 0.57 |
| | EnsMean | 1.76 | 2.30 | 0.28 | 0.36 | 1.48 | 0.23 | 0.32 | 0.63 |
| Week 3 | Control | 1.70 | 2.57 | 0.27 | 0.41 | 1.93 | 0.30 | 0.36 | 0.47 |
| | EnsMean | 1.80 | 2.35 | 0.29 | 0.37 | 1.51 | 0.24 | 0.33 | 0.57 |
| Week 4 | Control | 1.70 | 2.59 | 0.27 | 0.41 | 1.96 | 0.31 | 0.37 | 0.45 |
| | EnsMean | 1.81 | 2.36 | 0.29 | 0.37 | 1.53 | 0.24 | 0.33 | 0.55 |
| Week 5 | Control | 1.72 | 2.63 | 0.28 | 0.42 | 1.99 | 0.31 | 0.37 | 0.44 |
| | EnsMean | 1.82 | 2.39 | 0.30 | 0.38 | 1.55 | 0.24 | 0.34 | 0.54 |

An ensemble forecast validation must be accompanied by the analysis of its spread, calculating the dispersion among the 11 GEFSv12 ensemble members across different forecast lead times. The importance of correct spread in a wave ensemble is discussed by Saetra and Bidlot (2004). They describe how the ensemble spread measures the uncertainties in the predictions, being useful for determining the reliability of ensemble forecasts and for diagnosing errors. Mori and Hirakuchi (2004) demonstrated that the spread of ensemble members tends to monotonically increase with the forecast length, which is an important pattern that gives reliability to forecasts. The rank histogram serves as a suitable method for evaluating spread in terms of dispersion while also being sensitive to the bias of individual members. A full explanation of how to interpret rank histograms is found in Hamill (2001).

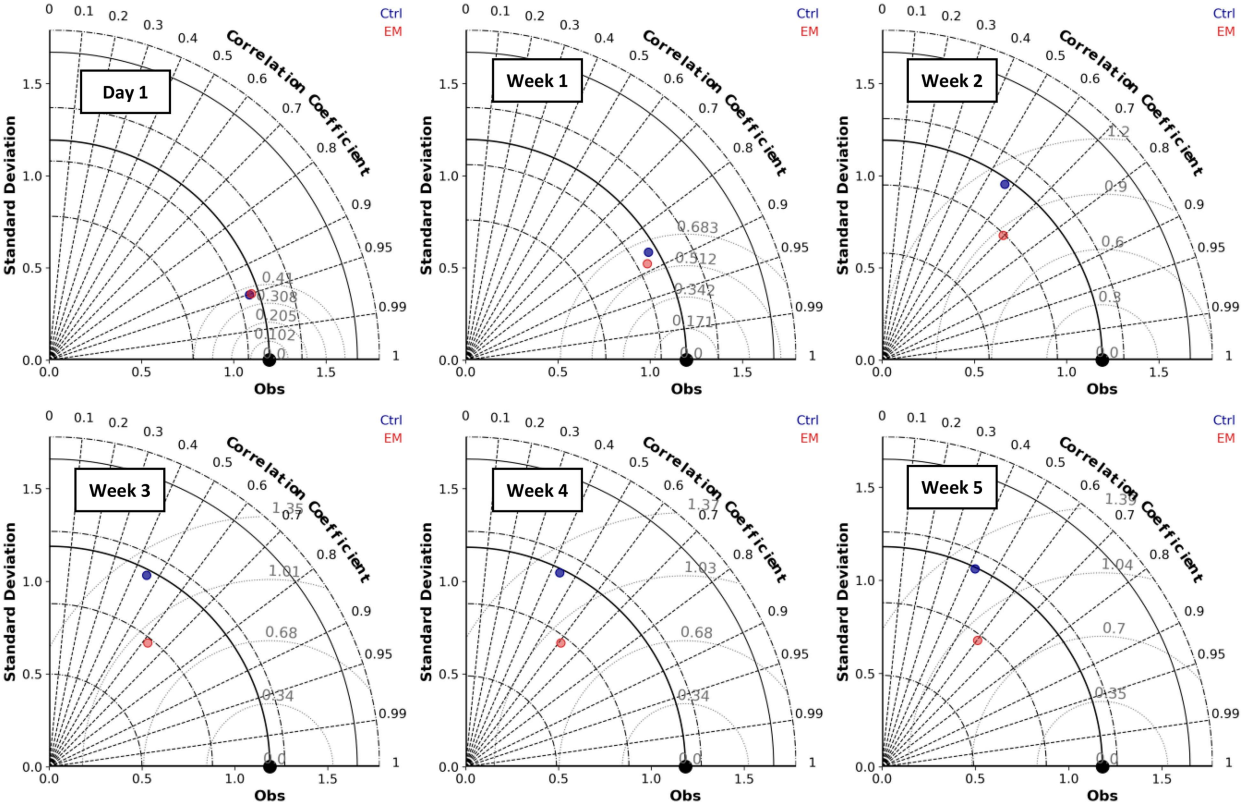


FIG. 12. Taylor diagrams showing the performance of GEFSv12 wave reforecast validated against buoy data, comparing the control member (blue) with the EnsMean (red). In each plot, the gray circular lines represent the RMSE. The model validation was conducted using buoy observations. Results are divided into forecast lead intervals, including day 1 and weeks 1–5. The solid centered line shows the standard deviation of the observations. Points on the left have lower standard deviations than the observations, usually overestimating small values and underestimating the top percentiles.

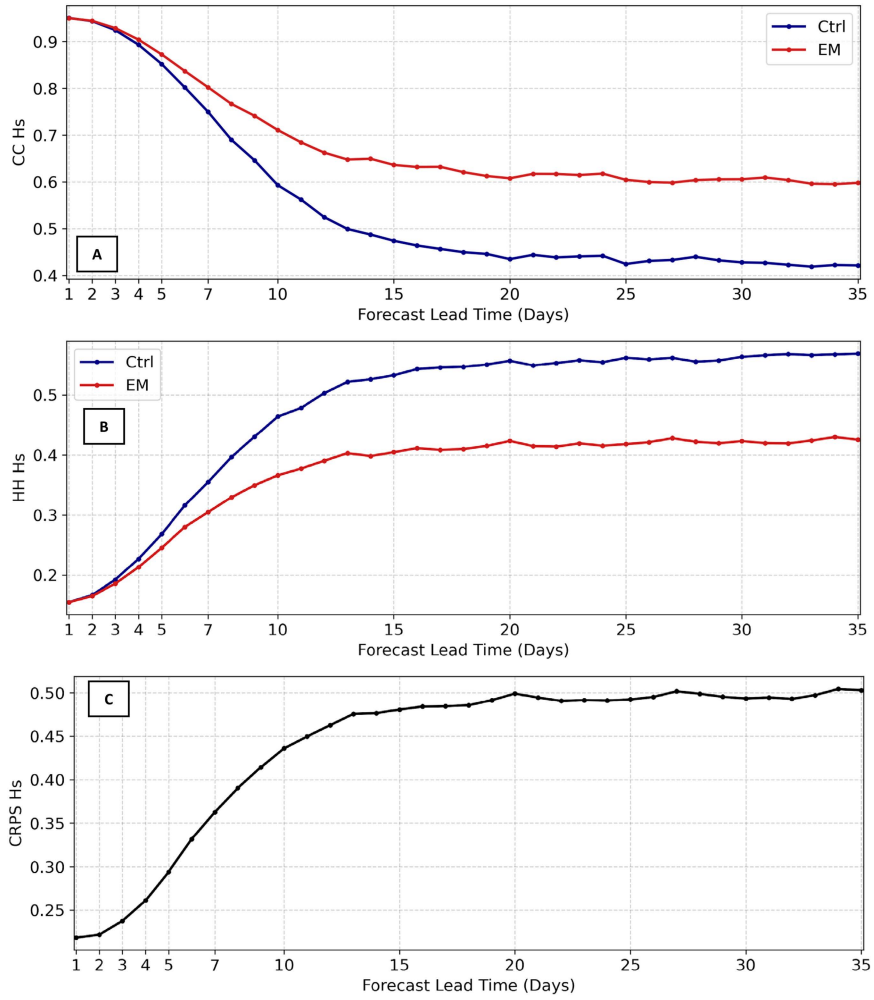


FIG. 13. (a),(b) Evolution of GEFSv12 wave reforecast error as a function of the forecast lead time for significant wave height (Hs). The model validation was conducted using buoy observations. The control member is plotted in blue, while the EnsMean is in red. (c) The CRPS, which measures the area between the cumulative distribution function of the ensemble forecast (using all members) and the observations.

The rank histograms computed for the GEFSv12 wave reforecast are presented in Fig. 15. Ideally, they should display a flat shape with bars containing similar probabilities or occurrences. A U shape indicates the ensemble is underspread, while an inverted bowl shape suggests the ensemble is overspread. Higher bar values on the left-hand side of the graphic may indicate positive bias, whereas the opposite may suggest negative bias. The results in Fig. 15 show that the ensemble is overconfident within the first week, especially in the short-range and day 1, suggesting underdispersion. The spread gradually increases across the forecast range, becoming more appropriate in weeks 2–5. Despite its small magnitude, the persistent positive bias impacts the rank histograms, evident in the asymmetry observed in the plots from Fig. 15. These findings align with the discussion in section 3 and corroborate with the results from Figs. 1 and 2.

### b. *Spatial validation using altimeter data*

Altimeter data offer several advantages over buoy data when it comes to model validation. This section explores two of them: (i) global spatial coverage and (ii) the provision of wind speed and significant wave height together. For the validation of the GEFSv12 wave reforecast, the quality-controlled and calibrated altimeter database from the Australian Ocean Data Network (AODN) was chosen. The reforecast period from 2000 to 2020 includes 11 satellite missions: TOPEX, *ERS-2*, *GFO*, *Jason-1*, *Envisat*, *Jason-2*, *CryoSat-2*, *HY-2A*, *SARAL*, *Jason-3*, and *Sentinel-3A*. Ribal and Young (2019) offer a complete description of the AODN dataset, providing details about each altimeter mission, uncertainties, estimated errors, and the calibration process. The altimeter calibration process is also discussed in Young et al. (2017).

The two wind and wave variables provided in the altimeter database, and therefore selected for validation, are significant

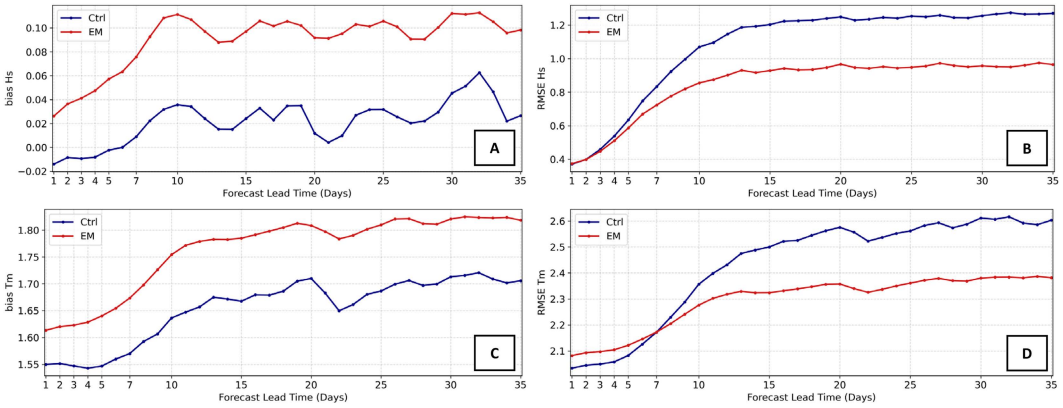


FIG. 14. Evolution of GEFSv12 wave reforecast error as a function of the forecast lead time for significant wave height (Hs) in the top plots and mean wave period (Tm) in the bottom plots. The model validation was conducted using buoy observations. The control member is represented in blue, while the EnsMean is plotted in red. The left side shows the bias, and the right side displays the RMSE.

wave height (Hs; m) and wind speed at 10-m height (U10; m $s^{-1}$). Data from the Ku band and Ka band (*SARAL*) were utilized, excluding the C-band data. The altimeter records were collocated into the 0.25° × 0.25° GEFSv12 wave grid using the methodology combined with spatial and temporal criteria suggested by Campos (2023). An inverse distance weighting using a linear function was applied to along-track altimeter records to produce the matchups of satellite/model data on the regular grid. Coastal water points were excluded, as previously reported, leading to a total of 34 646 126 matchups in deep waters. Figure 16 offers a global view of this dataset.

The same statistical analysis applied in the last section, including tables and plots, was applied to the validation against altimeter data. Additionally, to leverage the global data's inherent characteristics, the statistics have been recomputed for each grid point based on the methodology proposed by Campos et al. (2020a), which is based on Young and Holland (1996) and Sepulveda et al. (2015). This involved iteratively processing latitude and longitude grid points, pooling data within a 2° × 2° bin centered at each point. This approach yields global maps that are highly valuable for examining the regional distribution of forecast errors.

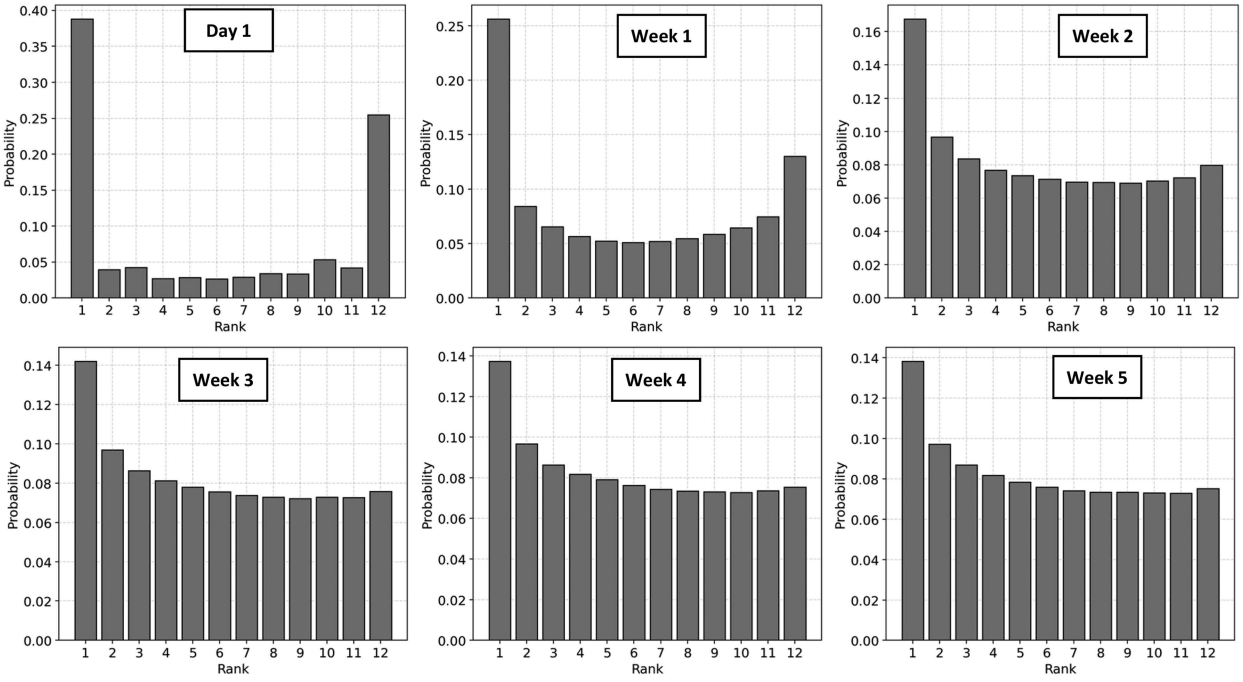


FIG. 15. Rank histograms of the GEFSv12 wave reforecast, for Hs. The model validation was conducted using buoy observations. Results are divided in forecast lead intervals, including day 1 and weeks 1–5.

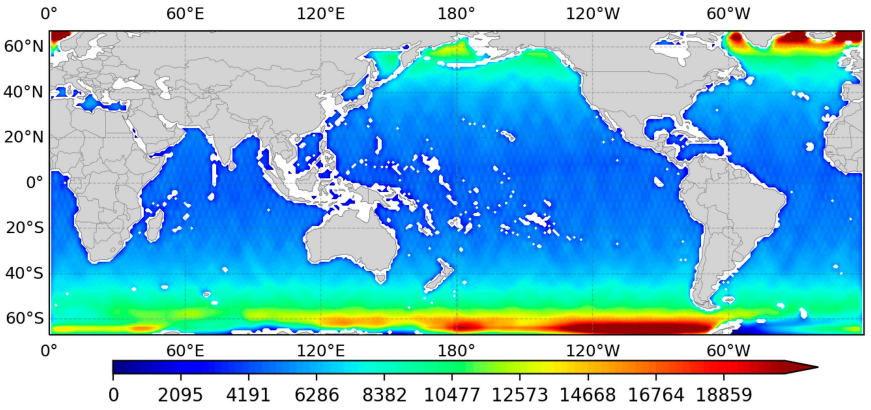


FIG. 16. Total count of AODN altimeter data distributed in space, on the GEFSv12 wave grid with a spatial resolution of 0.25° × 0.25°.

Before starting the reforecast assessment, a simple view of the collocated and binned altimeter data is presented in Fig. 17 in terms of arithmetic mean and 99th percentile of Hs and U10. The figure highlights regions with magnified intensity, illustrating where the escalation in severity in the 99th percentile mostly occurs. Latitudes north of 40°N and the entire Southern Ocean present the most extreme conditions. As expected, a direct correspondence is observed between the most active regions in U10 and Hs, confirming the strong relationship between surface winds and wave energy.

The initial results of RMSE considering the first 24 h of forecast are presented in Fig. 18. The global map of U10 illustrates the impact of the general atmospheric circulation, showing larger errors over the intertropical convergence zone (ITCZ), the western portion of semipermanent anticyclones, and cyclogenetic areas. Consistent with Campos et al. (2022b), the 10-m wind errors in Fig. 18 exhibit greater magnitudes in locations with warm currents. The RMSE pattern of Hs closely mirrors that of U10, especially in extratropical latitudes where the transfer of momentum from surface winds to ocean waves is more pronounced. However, unlike U10, equatorial locations at the ITCZ do not demonstrate significantly amplified RMSE for Hs, potentially due to the lower intensity of those winds and the influence of distant swells. Campos et al. (2020b) provided a global map illustrating the correlation coefficient between U10 and Hs, pointing to high correlations in the extratropics and lower correlations near the equator. The distribution of RMSE in Fig. 18 responds to this regional effect.

The RMSE has been decomposed into systematic and scatter errors, presented in Fig. 19. It shows the bias of U10 is positive (model overestimation) in equatorial regions and western portions of the Pacific and Indian Oceans. In contrast, the bias of U10 is negative (model underestimation) across great parts of the domain, especially in the Southern Ocean, where the altimeter winds are more intense than the GEFSv12 reforecast. The bias of Hs partially follows this spatial distribution but exhibits some differences. Although the

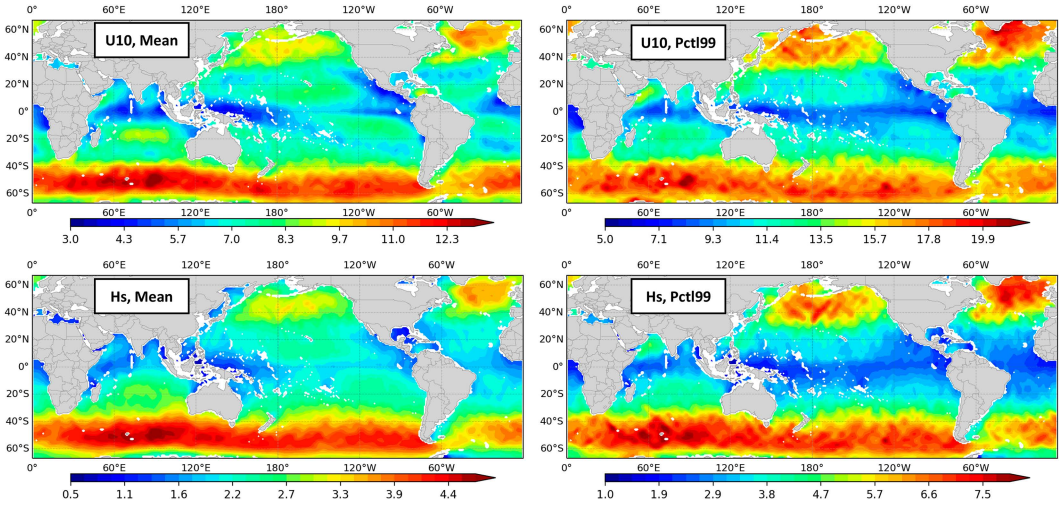


FIG. 17. Arithmetic mean and 99th percentile of the collocated altimeter data, computed using 2° × 2° bins centered at each grid point. (top) The 10-m wind speed (U10; m $s^{-1}$). (bottom) The significant wave height (Hs; m).

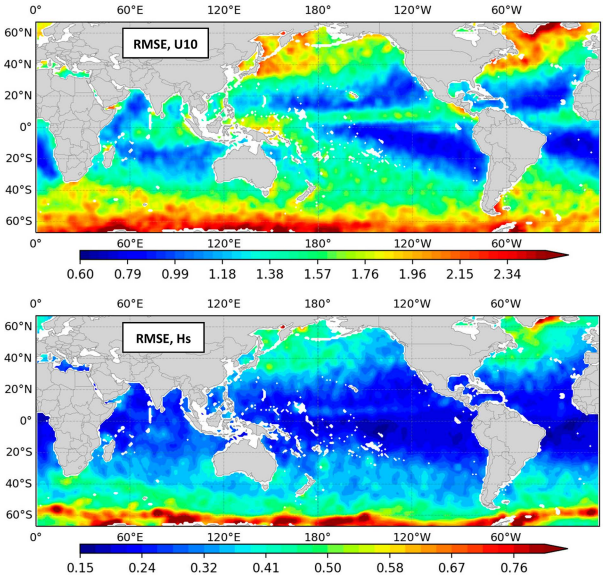


FIG. 18. Spatial validation of GEFSv12 wave reforecast using altimeter data. The global error maps display RMSE calculated using the first 24 h of forecast, for U10 (m $s^{-1}$) and Hs (m).

wave model is forced by wind speed that presents a reasonably homogeneous negative bias in the extratropics, the bias of Hs is mostly positive in the eastern portions of the Pacific Ocean and negative in the western portions. Similarly, the Atlantic Ocean has a more pronounced negative bias in the western longitudes. This pattern is more evident in the Pacific Ocean, associated with the largest basin and the longest swells.

It is known that the wave generation process requires persistent winds and large fetches, which are frequently present in the midlatitude meteorological flow from the west. Consequently, the results in Fig. 19 suggest that young waves tend to display a negative bias, while the mature swells, which accumulate uncertainties from wind and wave modeling, tend to exhibit a positive bias, on average. This effect occurs within a bias range that is generally very low, as indicated by the color bar scale in the plot. Storm-referenced assessments, following the meteorological systems in the Pacific Ocean, could further explore and validate this theory.

The scatter index in Fig. 19 highlights larger errors of U10 over warm ocean waters, directly affecting the waves and increasing the scatter errors of Hs. Across most regions globally, the scatter index of Hs ranges from 8% to 12%, while over relatively warm waters, it extends from 13% to 17%. This issue underscores the necessity for fully coupled systems, as evident from the influence of sea surface temperature on wave forecast performance. In conclusion, by examining both scatter index and bias concurrently, the eastern parts of the Pacific and Atlantic Oceans exhibit positive bias and low scatter errors, while the central-western regions display negative bias and higher scatter errors. These characteristics are crucial for consideration in future forecast system developments and postprocessing bias-correction algorithms.

The global maps show elevated errors in both U10 and Hs within polar regions near the Arctic and the Antarctic. This might be associated with model limitations in the wind–wave–ice interaction or associated with constraints in altimeter measurements near ice-covered areas. At present, no definitive conclusions can be drawn, emphasizing the need for dedicated validation specific to these regions.

The dataset and validation were segmented based on forecast lead time, from 1 to 35 days, employing the same methodology as the last section using buoy data. Figure 20 shows the correlation coefficient and normalized RMSE, now utilizing altimeter data with global coverage. The results closely resemble

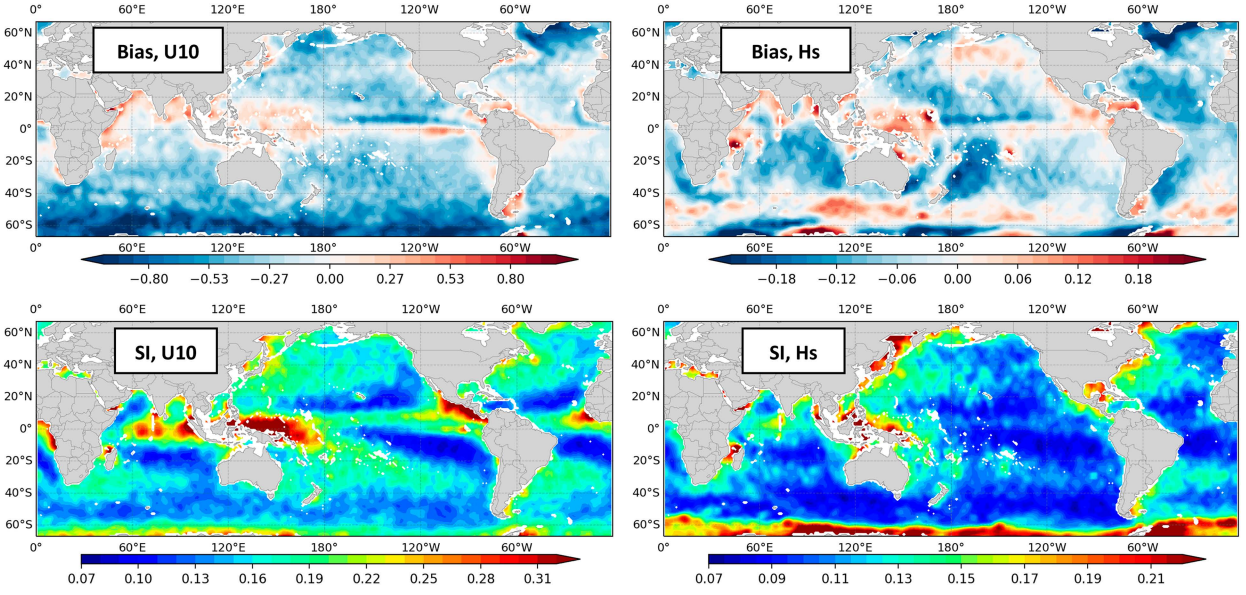


FIG. 19. Spatial validation of GEFSv12 wave reforecast using altimeter data. Global error maps showing (top) bias and (bottom) SI for U10 (m $s^{-1}$) and Hs (m). The SI is a unitless metric and can be interpreted as a percentage scatter error when multiplied by 100.

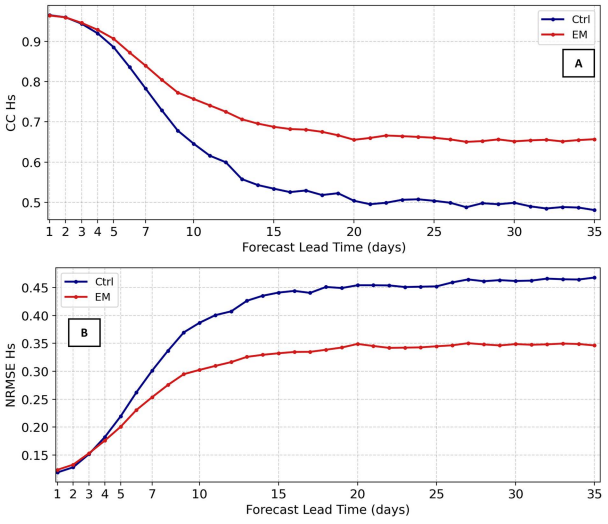


FIG. 20. Evolution of GEFSv12 wave reforecast error as a function of forecast lead time for significant wave height (Hs). The model validation was conducted using altimeter observations. The control member is plotted in blue, while the EnsMean is plotted in red.

those in Fig. 13, confirming the quantitative analysis and the observed error growth curves—an additional validation involving 34 million matchups from altimeters, which is a substantial increase from the 5 million matchups obtained from buoys. Nonetheless, slight differences are found between Figs. 20a and 13a concerning the CC values. In the reforecast validation using buoy data (Fig. 13a and Table 4), the CC begins at 0.95, declining to values below 0.9 by day 5. However, the validation using altimeter data (Fig. 20a) shows better performance, with CC starting at 0.96 and maintaining CC above 0.9 on day 5. Moreover, the minimum CC drops to 0.6 in the validation using buoy data, while the validation using an altimeter shows a minimum CC of 0.65. Regarding the increase in RMSE with forecast time, the results indicate RMSE remains below 20% within the initial 5 forecast days, emphasizing once more the outstanding performance of GEFSv12 in the short term.

Expanding the analysis of errors over forecast time using global maps allows us to identify areas where forecast deterioration is most prevalent. The global RMSE maps of U10 in Fig. 21 show a rapid increase in errors within the control member at extratropical latitudes. The most significant errors, above 4 m $s^{-1}$, are observed in the North Atlantic, Southern Ocean, and North Pacific. Meanwhile, the ensemble mean for week 1 demonstrates superior performance with lower RMSE, approximately 1 m $s^{-1}$ in tropical latitudes, and between 1.8 and 3.2 m $s^{-1}$ in the extratropics, above 35°N and below 35°S. The ensemble mean consistently outperforms the control member across all forecast ranges, albeit with smaller differences noted in week 1. The most substantial discrepancies emerge in week 2, where the control member shows considerable deterioration, while the ensemble mean maintains a lower RMSE. This illustrates the inadequacy of relying on deterministic forecasts for week 2 and beyond when ensemble forecasts are available.

The correspondence of Hs with U10 is very high, as illustrated in Fig. 22, indicating that RMSE in areas with high wind speeds significantly affects the quality of Hs. Given the efficient propagation of ocean waves with minimal dissipation, errors in Hs originating from U10 can propagate globally—transporting local wind errors to further distances. Therefore, unlike U10, localized errors in Hs may stem from forecast issues in remote locations. This becomes more critical considering that input and dissipation source terms, such as WAVEWATCH III ST4 (Ardhuin et al. 2010), rely on local wave spectra as input for their functions. For instance, if a swell is inaccurately represented due to substantial errors in U10 where it originated as wind sea, it can compromise the wave spectra across a vast area, consequently resulting in miscalculations in the local source terms—a cascade effect that expands the error.

It is interesting to note that there is not a substantial difference in RMSE between weeks 3, 4, and 5 for both U10 and Hs. However, while U10 experiences a significant deterioration in RMSE in the extratropical regions from week 1 to week 2, Hs appears to show a more gradual decrease in RMSE across weeks 1, 2, and 3. This could be associated with the prolonged propagation (or "memory") of mature swells in large basins such as the Southern Ocean and the Pacific Ocean. Finally, Figs. 20b and 22 confirm that the GEFSv12 wave ensemble overperforms the control member on a global scale, which is also discussed by Campos et al. (2020a) in their evaluation of GEFSv11.

## 5. Final discussion

In this paper, we described the development and validation of a new wave ensemble reforecast covering 20 years and extending to a 35-day forecast range. This represents a unique publicly available global dataset expected to support future research studies. The time frame of the GEFSv12 wave ensemble reforecast, from 2000 to 2019, is ideal for statistical analyses demanding large datasets, and it fortunately covers two decades of available, high-density observations—shown in Fig. 9a regarding the number of buoys and by Ribal and Young (2019) and in Fig. 16 in terms of altimeter data. The combination of a large reforecast dataset and extensive observations is essential for subsequent regional validations, model optimizations, the development of postprocessing bias-correction algorithms, and other demands in the marine industry such as extreme value analysis for design criteria. This paper provides detailed information on the reforecast construction that is expected to support upcoming system versions in the future.

Our first analysis of the GEFSv12 wave reforecast, concerning the ensemble spread and initial conditions, revealed that the spread of the wave ensemble members is primarily driven by the spread of the wind inputs, particularly noticeable beyond the initial 3-day forecast period. Similarly, the impact of initial conditions (ICs) on the wave ensemble forecast is predominantly limited to the first 5 days. Notably, we observed that the short-term wave forecasts of GEFSv12 were

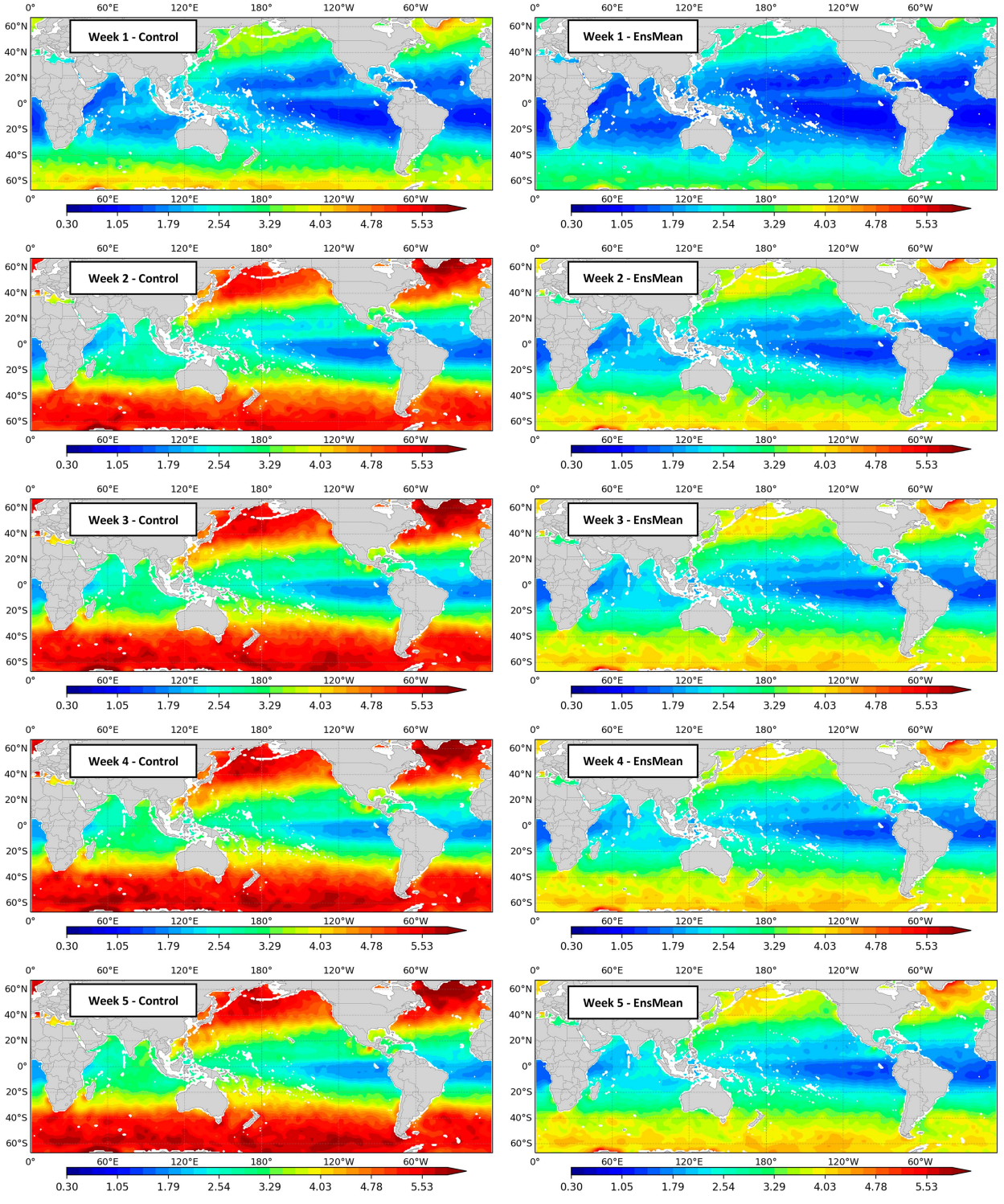


FIG. 21. Spatial validation of GEFSv12 wave reforecast using altimeter data. Global maps displaying RMSE for U10 (m $s^{-1}$) are presented across forecast time, ranging from weeks 1 to 5. (left) The control member. (right) The EnsMean. Consistent color palettes and ranges are maintained across all plots to allow direct comparison.

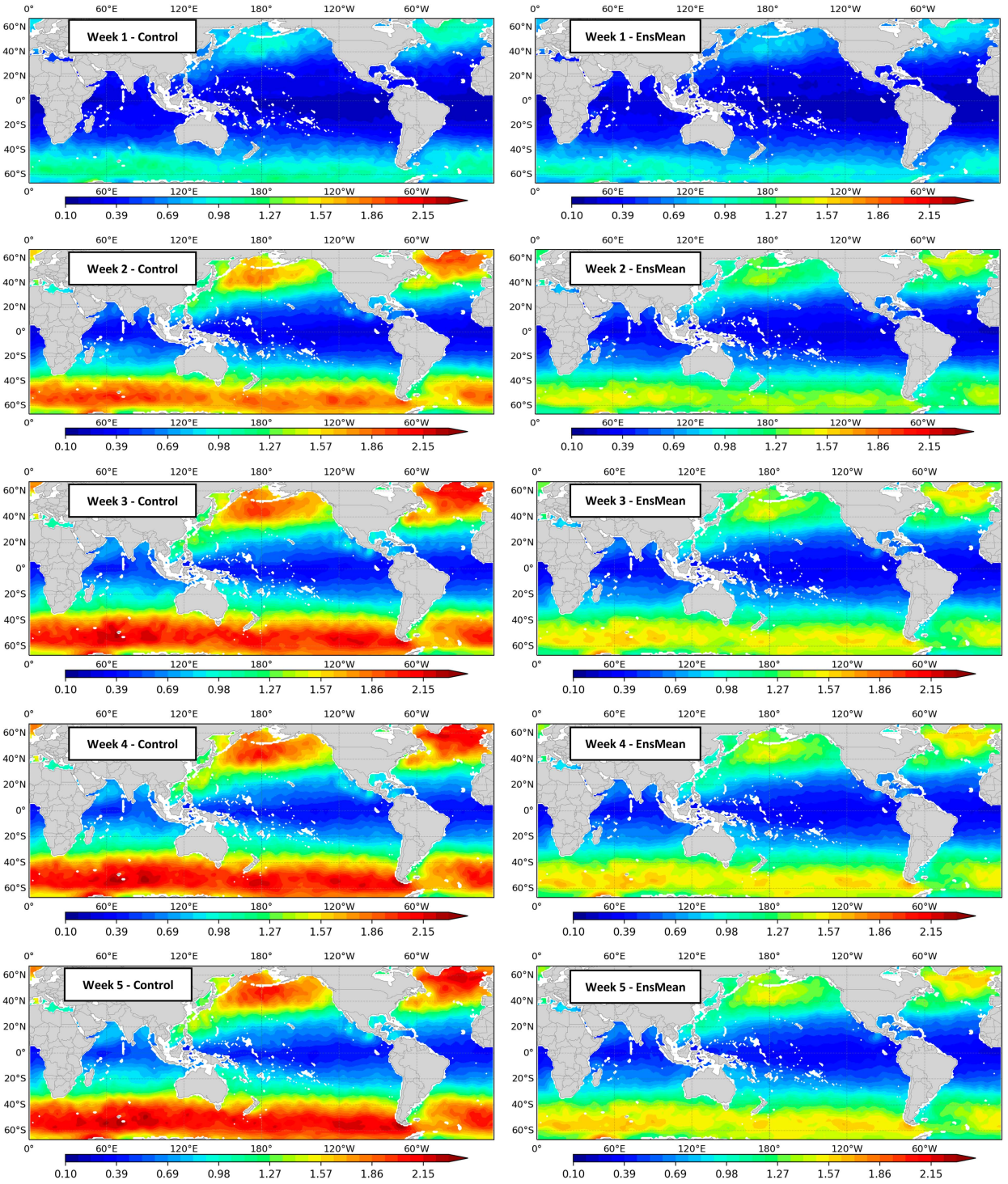


FIG. 22. Spatial validation of GEFSv12 wave reforecast using altimeter data. Global maps displaying RMSE for Hs (m) are presented across forecast time, ranging from weeks 1 to 5. (left) The control member. (right) The EnsMean. Consistent color palettes and ranges are maintained across all plots to allow direct comparison.

underspread within the first week, while for week 2 and beyond, the spread seems adequate. Interestingly, the assessment of the ensemble spread combined with the statistical validation suggests that the reduced scatter errors and higher correlation coefficients in the wave ensemble mean, compared to the control member, stem greatly from the atmospheric ensemble spread (wind inputs). Hence, the performance of the atmospheric component in GEFSv12, reported by Hamill et al. (2022) and

Zhou et al. (2022), has significantly contributed to the high-quality outputs of the wave reforecast.

In summary, the exceptional performance of wind inputs, combined with the successful optimization of WAVEWATCH III in GEFSv12, led to highly accurate results of significant wave height. It showed minimal bias close to zero, with a low RMSE of around 15%, and a correlation coefficient between 0.95 and 0.96 for day 1. The QQ plots illustrate the reforecast's high accuracy from low values to the upper percentiles (Figs. 10 and 11a). Although our validation did not focus on the wave spectrum, we observed that the performance of the wave period is significantly poorer compared to Hs. Considering the importance of wave period and directional spread for the industry and marine safety, we recommend a dedicated study focusing on improving these variables in the future.

The error growth curves and graphics indicate that beyond day 3, the ensemble mean significantly overperforms the control member, showing lower scatter errors and higher correlation coefficients. It demonstrates the superior performance of the wave ensemble forecast compared to deterministic forecasts while also illustrating the inadequacy of relying on deterministic forecasts for week 2 and beyond. Despite the improved quality of the ensemble mean, starting from week 2 forecasts, it tends to overestimate the smaller waves and underestimate more severe events above the 90th percentile. The evolution of error metrics as a function of forecast lead time, in Figs. 13, 14, and 20, indicates large scatter errors and low correlation coefficients beyond day 10, stabilizing from day 15 onward. This does not necessarily disqualify the ensemble application for midrange to long range but implies that ensemble forecast results should be processed differently. Rather than being restricted to the ensemble mean or individual members, a probabilistic approach using spatiotemporal analysis must be adopted.

The spatial validation using altimeter data revealed highly heterogeneous distributions in systematic and scatter errors. The amplified RMSE of U10 over warm currents directly compromised the wave model's performance, leading to increased Hs RMSE. This emphasizes the critical role of fully coupled systems in mitigating regional deficiencies. In a broader perspective, the global error maps indicated that the eastern parts of the Pacific and Atlantic Oceans exhibit positive bias and low scatter errors, while the central-western regions display negative bias and higher scatter errors. These errors can be reduced by postprocessing bias correction trained with altimeter data, incorporating latitude and longitude information to model spatial variability, as proposed by Campos et al. (2020b).

Finally, this research project and reforecast construction demanded substantial computational resources, storage, extensive scripting, and multidisciplinary scientific discussions. It demonstrated successful coordination among various NOAA centers, including the Atlantic Oceanographic and Meteorological Laboratory (AOML), Ocean Prediction Center (OPC), Environmental Modeling Center (EMC), and Climate Prediction Center (CPC). Within the same research project, a companion paper is in production, focusing on probabilistic wave forecasts for week 2 using a distinct validation approach based on fuzzy verification, proposed by Ebert (2008). Our future plans include (i) adding more observations to the validation and expanding the statistical analyses, (ii) further developing bias-correction algorithms using machine learning techniques, and (iii) designing new strategies dedicated to probabilistic wave forecasts for hurricane conditions. We believe these initiatives will address certain gaps and overcome current limitations in the GEFSv12 wave ensemble forecast outlined in this paper.

*Acknowledgments.* This work was funded by the Cooperative Institute for Marine and Atmospheric Studies (CIMAS), a Cooperative Institute of the University of Miami and the National Oceanic and Atmospheric Administration, Cooperative Agreement NA20OAR4320472. This study is part of the project entitled "Extending Marine Hazard Information to Week Two and Beyond." The WAVEWATCH III simulations and validations were conducted using the Orion supercomputer at Mississippi State University High Performance Computing (MSU-HPC). The authors acknowledge the support and expertise of Deanna Spindler and Todd Spindler for their efforts in the development and assessment of the GEFSv12 operational product. Additionally, the authors thank Richard Gorman for his contribution to the WAVEWATCH III optimization in GEFSv12. The authors would also like to thank Hendrik Tolman for the discussions on sources of error and spread.

*Data availability statement.* WAVEWATCH III main repository: https://github.com/NOAA-EMC/WW3. GEFSv12 20-year wave reforecast data: https://registry.opendata.aws/noaa-wave-ensemble-reforecast/, https://noaa-nws-gefswaves-reforecast-pds.s3.amazonaws.com/index.html, and https://github.com/NOAA-EMC/gefswaves_reforecast. GEFSv12 archive: https://registry.opendata.aws/noaa-gefs/ and https://noaa-gefs-pds.s3.amazonaws.com/index.html. AODN altimeter database: http://thredds.aodn.org.au/thredds/catalog/IMOS/SRS/Surface-Waves/Wave-Wind-Altimetry-DM00/catalog.html. NOAA National Data Buoy Center historical data: https://www.ndbc.noaa.gov/historical_data.shtml. Copernicus buoy data: https://catalogue.marine.copernicus.eu/documents/PUM/CMEMS-INS-PUM-013-045.pdf and ftp://my.cmems-du.eu/Core/INSITU_GLO_WAV_DISCRETE_MY_013_045/cmems_obs-ins_glo_wav_my_na_irr/history/MO/. Repository containing the scripts used for validation: https://github.com/NOAA-EMC/WW3-tools.

## REFERENCES

Abdolali, A., A. Roland, A. van der Westhuysen, J. Meixner, A. Chawla, T. J. Hesser, J. M. Smith, and M. D. Sikiric, 2020: Large-scale hurricane modeling using domain decomposition parallelization and implicit scheme implemented in WAVEWATCH III wave model. *Coastal Eng.*, **157**, 103656, https://doi.org/10.1016/j.coastaleng.2020.103656.

——, A. van der Westhuysen, Z. Ma, A. Mehra, A. Roland, and S. Moghimi, 2021: Evaluating the accuracy and uncertainty of atmospheric and wave model hindcasts during severe events using model ensembles. *Ocean Dyn.*, **71**, 217–235, https://doi.org/10.1007/s10236-020-01426-9.

Alves, J.-H., and Coauthors, 2013: The NCEP–FNMOC combined wave ensemble product: Expanding benefits of interagency probabilistic forecasts to the oceanic environment. *Bull. Amer. Meteor. Soc.*, **94**, 1893–1905, https://doi.org/10.1175/BAMS-D-12-00032.1.

——, N. Bernier, A. Chawla, P. Etala, and P. Wittmann, 2015: Probabilistic wave forecasting and ensemble-based data assimilation at the US National Weather Service. NCEP GWES, 2 pp., https://www.wcrp-climate.org/WGNE/BlueBook/2015/individual-articles/08_Alves_Jose-Henrique_etal_wave-ensembles.pdf.

——, and Coauthors, 2024: Development of a wave model component in the first coupled Global Ensemble Forecast System at NOAA. *Wea. Forecasting*, https://doi.org/10.1175/WAF-D-24-0048.1, in press.

Amante, C., and B. W. Eakins, 2009: ETOPO1 1 arc-minute global relief model: Procedures, data sources and analysis. NOAA Tech. Memo. NESDIS NGDC-24, 25 pp., https://www.ngdc.noaa.gov/mgg/global/relief/ETOPO1/docs/ETOPO1.pdf.

Ardhuin, F., and Coauthors, 2010: Semiempirical dissipation source functions for ocean waves. Part I: Definition, calibration, and validation. *J. Phys. Oceanogr.*, **40**, 1917–1941, https://doi.org/10.1175/2010JPO4324.1.

Behrens, A., 2015: Development of an ensemble prediction system for ocean surface waves in a coastal area. *Ocean Dyn.*, **65**, 469–486, https://doi.org/10.1007/s10236-015-0825-y.

Bell, R., and B. Kirtman, 2019: Seasonal forecasting of wind and waves in the North Atlantic using a grand multimodel ensemble. *Wea. Forecasting*, **34**, 31–59, https://doi.org/10.1175/WAF-D-18-0099.1.

Breivik, Ø., O. J. Aarnes, J.-R. Bidlot, A. Carrasco, and Ø. Saetra, 2013: Wave extremes in the northeast Atlantic from ensemble forecasts. *J. Climate*, **36**, 7525–7540, https://doi.org/10.1175/JCLI-D-12-00738.1.

Bunney, C., and A. Saulter, 2015: An ensemble forecast system for prediction of Atlantic–UK wind waves. *Ocean Modell.*, **96**, 103–116, https://doi.org/10.1016/j.ocemod.2015.07.005.

Campos, R. M., 2023: Analysis of spatial and temporal criteria for altimeter collocation of significant wave height and wind speed data in deep waters. *Remote Sens.*, **15**, 2203, https://doi.org/10.3390/rs15082203.

——, J.-H. G. M. Alves, S. G. Penny, and V. Krasnopolsky, 2018: Assessments of surface winds and waves from the NCEP Ensemble Forecast System. *Wea. Forecasting*, **33**, 1533–1564, https://doi.org/10.1175/WAF-D-18-0086.1.

——, ——, ——, and ——, 2020a: Global assessments of the NCEP Ensemble Forecast System using altimeter data. *Ocean Dyn.*, **70**, 405–419, https://doi.org/10.1007/s10236-019-01329-4.

——, V. Krasnopolsky, J.-H. Alves, and S. G. Penny, 2020b: Improving NCEP's global-scale wave ensemble averages using neural networks. *Ocean Modell.*, **149**, 101617, https://doi.org/10.1016/j.ocemod.2020.101617.

——, A. Abdolali, M. Masarik, and A. Mehra, 2022a: Visualization and validation methods applied to wave modeling. *The Unifying Innovations in Forecasting Capabilities Workshop (UIFCW)*, College Park, MD, Earth Prediction Innovation Center (EPIC), Unified Forecast System (UFS), and UFS Research to Operations (R2O) Project, 71, https://epic.noaa.gov/wp-content/uploads/2022/07/FINAL_UIFCW_Participant-Welcome-Packet_2022.pdf.

——, C. B. Gramcianinov, R. de Camargo, and P. L. da Silva Dias, 2022b: Assessment and calibration of ERA5 severe winds in the Atlantic Ocean using satellite data. *Remote Sens.*, **14**, 4918, https://doi.org/10.3390/rs14194918.

——, A. D'Agostini, B. R. L. França, A. L. A. Damião, and C. Guedes Soares, 2022c: Implementation of a multi-grid operational wave forecast in the South Atlantic Ocean. *Ocean Eng.*, **243**, 110173, https://doi.org/10.1016/j.oceaneng.2021.110173.

——, J. Meixner, A. Abdolali, S. Banihashemi, M. Masarik, and A. Mehra, 2023: Impact of ensemble initialization on an extended wave forecast system. *Third Int. Workshop on Waves, Storm Surges, and Coastal Hazards Incorporating the 17th Int. Waves Workshop*, Notre Dame, IN, University of Notre Dame, https://waveworkshop.nd.edu/.

——, D. Figurskey, and A. Mehra, 2024: Probabilistic wave forecast for week two and beyond based on the NCEP's Global Ensemble Forecast System. *12th Symposium on the Weather, Water, and Climate Enterprise*, Baltimore, MD, Amer. Meteor. Soc., J2.5, https://ams.confex.com/ams/104ANNUAL/meetingapp.cgi/Paper/433936.

Chang, H.-W., C.-C. Yen, M.-C. Lin, and C.-H. Chu, 2017: Establishment and performance of the ocean wave ensemble forecast system at CWB. *J. Mar. Sci. Technol.*, **25**, 11, https://doi.org/10.6119/JMST-017-0622-1.

Cui, B., Z. Toth, Y. Zhu, and D. Hou, 2012: Bias correction for global ensemble forecast. *Wea. Forecasting*, **27**, 396–410, https://doi.org/10.1175/WAF-D-11-00011.1.

Ebert, E. E., 2008: Fuzzy verification of high-resolution gridded forecasts: A review and proposed framework. *Meteor. Appl.*, **15**, 51–64, https://doi.org/10.1002/met.25.

Farina, L., 2002: On ensemble prediction of ocean waves. *Tellus*, **54A**, 148–158, https://doi.org/10.3402/tellusa.v54i2.12133.

Gorman, R. M., and H. J. Oliver, 2018: Automated model optimisation using the Cylc workflow engine (Cyclops v1.0). *Geosci. Model Dev.*, **11**, 2153–2173, https://doi.org/10.5194/gmd-11-2153-2018.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560, https://doi.org/10.1175/1520-0493(2001)129%3C0550:IORHFV%3E2.0.CO;2.

——, and Coauthors, 2022: The reanalysis for the Global Ensemble Forecast System, version 12. *Mon. Wea. Rev.*, **150**, 59–79, https://doi.org/10.1175/MWR-D-21-0023.1.

Hanna, S., and D. Heinold, 1985: *Development and Application of a Simple Method for Evaluating Air Quality*. American Petroleum Institute, 38 pp.

Hasselmann, S., and K. Hasselmann, 1985: Computations and parameterizations of the nonlinear energy transfer in a gravity-wave spectrum. Part I: A new method for efficient computations of the exact nonlinear transfer integral. *J. Phys. Oceanogr.*, **15**, 1369–1377, https://doi.org/10.1175/1520-0485(1985)015%3C1369:CAPOTN%3E2.0.CO;2.

Janssen, P. A. E. M., J. D. Doyle, J. Bidlot, B. Hansen, L. Isaksen, and P. Viterbo, 2002: Impact and feedback of ocean waves on the atmosphere. *Atmosphere–Ocean Interactions*, N. Perrie, Ed., WIT Press, 155–197.

Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 341 pp.

Lazarus, S. M., S. T. Wilson, M. E. Splitt, and G. A. Zarillo, 2013a: Evaluation of a wind-wave system for ensemble tropical cyclone wave forecasting. Part I: Winds. *Wea. Forecasting*, **28**, 297–315, https://doi.org/10.1175/WAF-D-12-00054.1.

——, ——, ——, and ——, 2013b: Evaluation of a wind-wave system for ensemble tropical cyclone wave forecasting. Part II: Waves. *Wea. Forecasting*, **28**, 316–330, https://doi.org/10.1175/WAF-D-12-00053.1.

Luo, X., R. Yan, and S. Wang, 2023: Comparison of deterministic and ensemble weather forecasts on ship sailing speed optimization. *Transp. Res.*, **121D**, 103801, https://doi.org/10.1016/j.trd.2023.103801.

Mentaschi, L., G. Besio, F. Cassola, and A. Mazzino, 2013: Problems in RMSE-based wave model validations. *Ocean Modell.*, **72**, 53–58, https://doi.org/10.1016/j.ocemod.2013.08.003.

Meucci, A., I. R. Young, and Ø. Breivik, 2018: Wind and wave extremes from atmosphere and wave model ensembles. *J. Climate*, **31**, 8819–8842, https://doi.org/10.1175/JCLI-D-18-0217.1.

Mori, N., and H. Hirakuchi, 2004: Short range wave forecasts using ensemble wave prediction system. *Coast. Eng.*, 1009–1021, http://doi.org/10.1142/9789812701916_0080.

NCEP WMO, 2023: NCEP WMO GRIB2 documentation version 31.0.0, https://www.nco.ncep.noaa.gov/pmb/docs/grib2/grib2_doc/.

NDBC, 2015: NDBC Web Data Guide. National Data Buoy Center – National Oceanic and Atmospheric Administration, 15 pp., https://www.ndbc.noaa.gov/docs/ndbc_web_data_guide.pdf.

Pallares, E., H. Hernandez, J. Moré, M. Espino, and A. Sairouni, 2015: Wave ensemble forecast in the Western Mediterranean Sea, application to an early warning system. *EGU General Assembly 2015*, Vienna, Austria, European Geoscience Union, EGU2015-8653, https://ui.adsabs.harvard.edu/abs/2015EGUGA..17.8653P/abstract.

Pan, S.-q., Y.-m. Fan, J.-m. Chen, and C.-c. Kao, 2016: Optimization of multi-model ensemble forecasting of typhoon waves. *Water Sci. Eng.*, **9**, 52–57, https://doi.org/10.1016/j.wse.2016.02.001.

Pezzutto, P., A. Saulter, L. Cavaleri, C. Bunney, F. Marcucci, L. Torrisi, and S. Sebastianelli, 2016: Performance comparison of meso-scale ensemble wave forecasting systems for Mediterranean Sea states. *Ocean Modell.*, **104**, 171–186, https://doi.org/10.1016/j.ocemod.2016.06.002.

Ribal, A., and I. R. Young, 2019: 33 years of globally calibrated wave height and wind speed data based on altimeter observations. *Sci. Data*, **6**, 77, https://doi.org/10.1038/s41597-019-0083-9.

Roh, M., H.-S. Kim, P.-H. Chang, and S.-M. Oh, 2021: Numerical simulation of wind wave using ensemble forecast wave model: A case study of Typhoon Lingling. *J. Mar. Sci. Eng.*, **9**, 475, https://doi.org/10.3390/jmse9050475.

Saetra, Ø., and J.-R. Bidlot, 2004: Potential benefits of using probabilistic forecasts for waves and marine winds based on the ECMWF Ensemble Prediction System. *Wea. Forecasting*, **19**, 673–689, https://doi.org/10.1175/1520-0434(2004)019%3C0673:PBOUPF%3E2.0.CO;2.

Saha, S., and Coauthors, 2010: The NCEP Climate Forecast System Reanalysis. *Bull. Amer. Meteor. Soc.*, **91**, 1015–1058, https://doi.org/10.1175/2010BAMS3001.1.

Sampson, C. R., E. A. Serra, J. A. Knaff, and J. H. Cossuth, 2011: Evaluation of global wave probabilities consistent with official forecasts. *Wea. Forecasting*, **36**, 1891–1904, https://doi.org/10.1175/waf-d-21-0037.1.

Sepulveda, H. H., P. Queffeulou, and F. Ardhuin, 2015: Assessment of SARAL/AltiKa wave height measurements relative to buoy, Jason-2, and Cryosat-2 data. *Mar. Geod.*, **38**, 449–465, https://doi.org/10.1080/01490419.2014.1000470.

Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.*, **106**, 7183–7192, https://doi.org/10.1029/2000JD900719.

Tolman, H. L., 2003: Treatment of unresolved islands and ice in wind wave models. *Ocean Modell.*, **5**, 219–231, https://doi.org/10.1016/S1463-5003(02)00040-9.

——, J.-H. G. M. Alves, and Y. Y. Chao, 2005: Operational forecasting of wind-generated waves by Hurricane Isabel at NCEP. *Wea. Forecasting*, **20**, 544–557, https://doi.org/10.1175/WAF852.1.

Valiente, N. G., A. Saulter, B. Gomez, C. Bunney, J.-G. Li, T. Palmer, and C. Pequignet, 2023: The Met Office operational wave forecasting system: The evolution of the regional and global models. *Geosci. Model Dev.*, **16**, 2515–2538, https://doi.org/10.5194/gmd-16-2515-2023.

Willmott, C. J., S. G. Ackleson, R. E. Davis, J. J. Feddema, K. M. Klink, D. R. Legates, J. O'Donnell, and C. M. Rowe, 1985: Statistics for the evaluation and comparison of models. *J. Geophys. Res.*, **90**, 8995–9005, https://doi.org/10.1029/JC090iC05p08995.

WW3DG, 2019: User manual and system documentation of WAVEWATCH III version 6.07. NOAA/NWS/NCEP/MMAB Tech. Note 333, 466 pp., https://raw.githubusercontent.com/wiki/NOAA-EMC/WW3/files/manual.pdf.

Xu, F., W. Perrie, B. Toulany, and P. C. Smith, 2007: Wind-generated waves in Hurricane Juan. *Ocean Modell.*, **16**, 188–205, https://doi.org/10.1016/j.ocemod.2006.09.001.

Young, I. R., and G. J. Holland, 1996: *Atlas of the Oceans: Wind and Wave Climate*. Elsevier, 241 pp.

——, E. Sanina, and A. V. Babanin, 2017: Calibration and cross validation of a global wind and wave database of altimeter, radiometer, and scatterometer measurements. *J. Atmos. Oceanic Technol.*, **34**, 1285–1306, https://doi.org/10.1175/JTECH-D-16-0145.1.

Zhou, X., and Coauthors, 2022: The development of the NCEP Global Ensemble Forecast System version 12. *Wea. Forecasting*, **37**, 1069–1084, https://doi.org/10.1175/WAF-D-21-0112.1.

Zhu, Y., and Z. Toth, 2008: Ensemble based probabilistic forecast verification. *19th Conf. on Probability and Statistics*, New Orleans, LA, Amer. Meteor. Soc., 2.2, https://ams.confex.com/ams/88Annual/techprogram/paper_131645.htm.

——, and Coauthors, 2018: Toward the improvement of subseasonal prediction in the National Centers for Environmental Prediction Global Ensemble Forecast System. *J. Geophys. Res. Atmos.*, **123**, 6732–6745, https://doi.org/10.1029/2018JD028506.

Zieger, S., D. Greenslade, and J. D. Kepert, 2018: Wave ensemble forecast system for tropical cyclones in the Australian region. *Ocean Dyn.*, **68**, 603–625, https://doi.org/10.1007/s10236-018-1145-9.