Title: Evaluation of Global Wave Probabilities Consistent with Official Forecasts

Authors: Charles R. Sampson[1], Efren A. Serra[2], John A. Knaff[3], Joshua H. Cossuth[4]

Affiliations: [1]Naval Research Laboratory, Monterey, CA, USA

[2]DeVine Consulting, Naval Research Laboratory, Monterey, CA, USA

[3]NOAA STAR, CIRA, Fort Collins, CO, USA

[4]Office of Naval Research, Arlington, VA, USA

Corresponding author: Charles R. Sampson, NRL Monterey. Buck.Sampson@nrlmry.navy.mil

Abstract

The U.S. Navy is keenly interested in analyses and predictions of waves at sea due to their effects on important tasks such as shipping, base preparedness and disaster relief. U.S. Tropical Cyclone (TC) Forecast Centers routinely disseminate wind probabilities consistent with official TC forecasts worldwide, but do not do the same for wave forecasts. These probabilities are especially important at longer leads where TC forecast accuracy diminishes. This work describes global wave probabilities consistent with both the official TC forecasts and their wind probabilities. Real-time runs for 84 TCs between May 2018 and March 2019, with probabilities generated for 12-ft and 18-ft significant wave heights are used to calculate verification statistics. This results in 347, 319, 261, 214, 155, and 112 verification cases at lead times of 1, 2, 3, 4, and 5 days where each verification case consists of a 20x20 degree latitude longitude grid around the verifying TC position. When compared with wave probabilities generated solely by a global numerical weather prediction model, the wind probability-based algorithm demonstrates improved consistency with official forecasts and provides additional benefits. Those benefits include an improved capability to discriminate between 12-ft and 18-ft significant wave events and non-events. The verification statistics also shows that the wind probability-based algorithm has a consistent high bias. How these biases can be reduced in future efforts is also discussed.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-21-0037.1.

Significance Statement

The extreme wave heights associated with tropical cyclones are difficult to accurately forecast deterministically or probabilistically. To exacerbate matters, existing global ensemble systems cannot resolve the strongest winds in hurricanes and typhoons and provide input to wave models that is inconsistent with official forecasts. This paper describes an algorithm that provides ensemble winds wave products that are both more realistic and consistent with official forecasts from tropical cyclone forecast centers. We show that this method provides improved identification of extreme wave events, which should provide improved input for ship navigation and hazard avoidance that saves both lives and property.

1. Introduction

U.S. Navy operations are adversely impacted by high seas, especially those from tropical cyclones (TCs).   In particular, the U.S. Navy is concerned about significant wave heights and their effects on safely routing ships, routine and emergency ship sorties, and Human Assistance Disaster Relief activities. Traditionally, wave model ensembles are run with Numerical Weather Prediction (NWP) model surface winds to produce significant wave heights and wave height probabilities around TCs. However, the NWP models are generally inconsistent with official forecasts from the U.S. TC forecast centers and lack the resolution to adequately capture large gradients in TC structure specified in the official forecasts (e.g., Tolman et al. 2005). This is problematic for forecasters and downstream applications as the inconsistencies add confusion

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-21-0037.1.

to an already stressful situation. To address this issue, the U.S. Navy's Fleet Numerical Meteorology and Oceanography Center (FNMOC) implemented a deterministic global wave model forecast that uses post-processed winds from U.S. TC forecast centers as input to WAVEWATCH III ® (WW3; Tolman 1991, Tolman et al. 2002, NCEP 2020). This algorithm is named for the WAVEWATCH III model (WW3) and its input TC winds from the U.S. TC forecast centers (OFCL), thus named WW3TCOFCL (Sampson et al. 2013). Faced with deficiencies in both the forcing winds and resolution for forecasting TC generated waves in the Northwest Australian region, the Australian Bureau of Meteorology (Zieger et al. 2018, Aijaz et al. 2019) designed a post-processing method that correct wind distribution biases associated with TCs in the NWP model ensembles used to force their high resolution (8 km) wave model. For each ensemble member, the method constructs a synthetic vortex to replace the existing one, keeping the asymmetric flow in in the numerical model. An evaluation of operational real-time runs found improvements in both TC wind and TC-generated wave probabilities, and importantly they had consistency between the winds from the NWP ensemble and the waves. These consistency and resolution issues are important to operations, and as yet there is no operational global wave model ensemble consistent with U.S. TC forecast center forecasts, wind probabilities associated with TC forecasts (DeMaria et al. 2013), and deterministic wave forecasts derived from U.S. forecast center forecasts (Sampson et al. 2013).

To address both consistency and resolution issues, a post-processing algorithm has been developed that constructs and inserts realistic wind structure in the vicinity of TCs out to 120 h. These winds are consistent with the forecasts from the U.S. TC forecast centers, which are frequently quite different in track, intensity and/or structure from the NAVGEM or other

Accepted for publication in *Weather and Forecasting*. DOI 10.1175/WAF-D-21-0037.1.

numerical model forecasts. These differences between official U.S. TC forecast and NWP forecasts can cause confusion for forecasters, warning managers and the general public in a time when coordinated and clear communication is of the utmost importance. The post-processed winds can then be used in the Navy Global Environmental Model (NAVGEM, Hogan et al. 2015) global wave model ensemble to produce wave probability fields that are consistent with deterministic TC forecasts and wind probabilities generated at the U.S. TC forecast centers. The current incarnation of this algorithm is designed to run as a 20-member ensemble on a 0.25 degree global WW3 grid, the same as currently used at FNMOC. This is an intentional design to be consistent with the current NAVGEM global wave model ensemble so that implementation is simplified, extra computational resources are minimal, and the wind post-processing algorithm can be run independently of the NAVGEM global wave model ensemble. Sampson et al. (2016) demonstrated that more ensemble members would be beneficial, but computational restrictions may not allow for expanding the ensemble. NRL has implemented the post-processing algorithm with the WW3 ensemble, executed in real-time for over a year, and gathered runs for this evaluation. The algorithm, hereafter referred to as WW3TCOFCL Ensemble, is described in section 2. Section 3 provides a description of how the data is used to conduct our evaluation. The result of the evaluations is provided in section 4, where individual cases and probabilistic verification is presented followed by conclusions and discussion of future work.

2. Algorithm Description

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-21-0037.1.

The WW3TCOFCL Ensemble follows the algorithm published in Sampson et al. (2016), except that the number of ensemble members has been reduced to 20 (the same number as in the FNMOC operational WW3 ensemble run using NAVGEM Ensemble surface winds, hereafter referred to as the WW3NAVGEM Ensemble) from 128. The WW3TCOFCL Ensemble grid has also been expanded to a global 0.25x0.25 degree grid to match the operational WW3NAVGEM Ensemble. These changes are made so that the algorithm adheres to computing and other resource constraints at FNMOC, and so that the algorithm could also be implemented within the current WW3NAVGEM Ensemble job instead of as a completely separate algorithm. Expanding the application to a global grid and reducing the number of ensemble members to 20 introduced major changes to the algorithm with potentially adverse effects. Also, there have been important changes (new sensors and new methods) in wind structure analysis that occurred at the Joint Typhoon Warning Center since the original evaluation that could potentially change the performance of the WW3TCOFCL Ensemble. And finally, the global grid allows wave to propagate around the world as they do in the real world while the limited domains in Sampson et al. (2016) did not. All these changes require vetting since their overall effects on performance are uncertain.

To summarize the current WW3TCOFCL Ensemble algorithm: First, 20 forecast ensemble members from the original 1000 generated using the Wind Speed Probability (WSP) algorithm (DeMaria et al. 2013) are randomly selected. Each WSP ensemble member is made available to the WW3TCOFCL deterministic model (Sampson et al. 2013) independently to create each ensemble member. The ensemble member is essentially the same as an official forecast defined at 0, 12, 24, 36, 48, 72, 96, and 120 h with the extent of the circulation extending to 20 kt at the radius of outermost closed isobar specified in the TC analysis. Hourly TC forecast wind fields are created and interpolated to high-resolution hourly storm-scale gridded fields using O'Reilly and Guza (1993) tessellation. Then, NAVGEM Ensemble surface wind fields are post-processed by removing the NWP model TC vortex from each member's set of forecast fields. Location is determined by using predicted centers from the National Centers for Environmental Prediction (NCEP) vortex tracker (Marchok 2002). The entire area out to the analyzed radius of outermost closed isobar is removed at all forecast times. This is done to remove geographical displaced and structurally different NAVGEM Ensemble forecasts so that only the background field remains. The removed TC vortex is replaced with bilinear interpolated data from the borders of the removed area. The final step of the gridded surface wind processing is inserting the hourly storm-scale gridded fields (one for each active TC) into the NAVGEM 10 m winds (originally at 1 degree resolution) to a 0.25x0.25 degree global grid for WW3 v5.16 — the operational version at FNMOC during 2018 and 2019. Even this resolution is insufficient to resolve the highest winds and waves, especially with TCs that have small eyewalls. The resultant set of gridded surface wind field forecasts at 1-h forecast intervals provide the wind forcing for WW3 to generate ocean wave forecasts for each ensemble member. Those ensemble wave forecasts are then combined to yield significant wave height probability fields exceeding a threshold (e.g., 12 or 18 ft) on a 1 degree resolution grid, which has a resolution consistent with the current WW3NAVGEM Ensemble probabilities available from FNMOC for evaluation purposes. An example of the 12-ft significant wave height probabilities on the right side of Fig. 1. Since we are only running 20 members of the WW3 ensemble, the probability fields are generated on a 1x1 degree global grid to reduce graininess noted in Sampson et al. (2016). Still, this graininess is visible at longer lead forecast times such as the 96-h WW3TCOFCL Ensemble forecast probabilities shown in Fig. 1.

The entire 10-m wind field preparation process takes just a few minutes on a Cray XC-30, and an estimated 1 hour of wall-time to run both the wind field preparation and the 20 WW3 ensemble members using 16 processors per ensemble member. Although attempts are made to warm start the WW3TCOFCL Ensemble every 12 hours using the previous 12-h forecast, this was not feasible when NRL computer resources became unavailable for extended periods of time. In these instances, the WW3TCOFCL Ensemble was cold started with potentially adverse effects on seas and swell in the early forecast times. These effects become less important beyond 24 h, but they are worth noting as they are plainly visible in visual inspection.

3. Evaluation Data

The WW3TCOFCL Ensemble was run in real-time on 84 TCs that existed between May 2018 and March 2019. NRL was able produce forecast data in the vicinity of TCs in all regions of the globe. As with most non-operational real-time NWP systems, NRL had issues with data acquisition and unscheduled computer downtime. As a result of this computer downtime, the evaluation set has periodic gaps resulting in some artifacts from the many WW3 cold starts, some of which are visible in our evaluation. Since the WW3TCOFCL Ensemble was run on the same grid and has the same number of members as the WW3NAVGEM Ensemble, verification of head-to-head cases will provide insight into both ensembles.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-21-0037.1.

For ground truth we use the WW3TCOFCL deterministic model analysis of significant wave height in feet (ft; 1 ft = 0.3048m), as that is the parameter most commonly used in Navy operations. Noting again that the WW3TCOFCL deterministic model uses post-processed winds forecasted by U.S. TC forecast centers.   Since the U.S. Navy is most concerned about significant wave heights in ship routing, we chose to evaluate significant wave height probabilities. We present results using WW3TCOFCL deterministic model significant wave height analyses, but we also evaluated results against WW3NAVGEM deterministic analyses. The WW3NAVGEM deterministic model analyses assimilate altimeter data (Cummings and Wittmann 2009), but little difference was found between results using the WW3TCOFCL and WW3NAVGEM deterministic model analyses as ground truth. The 12-and 18-ft thresholds chosen for evaluation are not necessarily the thresholds used for operational forecasting, but span a reasonable range of significant wave heights associated with TCs and are routinely available for the WW3NAVGEM Ensemble.

To gather data with 12-and 18-ft significant wave heights, which are not common in the tropics, our verification was limited to a 20x20 degree box surrounding the verifying TC position. This area is likely larger than the TC wind field (Frank 1977) and also generally encompasses the extreme waves associated with TCs. In most cases a 20x20 degree box will include many cases of zero probabilities in both the forecast and verification data (null cases), which affects results and their interpretation. The verification impacts of null cases are discussed section 3.   We also attempted this evaluation using a 10x10 degree box around the verifying TC location, and found that this smaller area did not always encompass the TC-driven

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-21-0037.1.

waves and highest significant wave height probabilities at longer forecast leads. At these longer leads, the area of high significant wave height probabilities can be both larger and dislocated from the 10x10 degree box around the verifying position. Our evaluation was also limited to TCs with verifying intensities of 35 knot (kt; 1 kt =0.514 m s^{-1}) or greater intensity, which results in limiting the false alarm rates for both algorithms.

Although we verify WW3TCOFCL Ensemble probabilities against WW3NAVGEM deterministic model significant wave height analyses (which assimilate altimeter wave heights), we do not to attempt verification ensemble runs against buoys and/or altimetry data explicitly, other than anecdotally. These observations have coverage issues that hinder verification of steep gradients and rare events, and can yield misleading results (see Sampson et al. 2013).

Table 1 provides a summary of the cases used in the verification. Each 20x20 degree verification area represents 400 potential paired forecast and verification points, so the values in Table 1 are effectively 1/400th of the paired forecast points evaluated (minus an estimated 10% that verified over land and were removed from verification). Grid differences also accounted for minor differences in the matched pairs over water, 1 or 2 paired forecasts in approximately 10% of the cases. This represents differences of less than 0.1% and is ignored.

Summary statistics at the end of the Results Section are provided with significance using a 2-tailed Student’s t-test. To remove correlation issues within the data, each 20x20 degree (each with potentially 400 paired forecasts) is treated as a single case. Then the t-tests are provided for the summary statistics –Discrimination Distance, ROC AUC, and Brier Score. No

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-21-0037.1.

effort is made to account for the effects of serial correlation in the summary data, but the degrees of freedom are conservatively estimated using the number of cases rather than the number of matched pairs (i.e., counting every point in the 20x20 degree box as a case).

4. Results

To demonstrate significant wave height forecasts we present results in three ways. We first present two cases that exemplify our real-time assessment of the differences between WW3TCOFCL Ensemble and WW3NAVGEM Ensemble significant wave height probabilities. We then verify WW3TCOFCL Ensemble and WW3NAVGEM Ensemble against WW3TCOFCL deterministic model significant wave height analyses, and for completeness, against WW3NAVGEM deterministic significant wave height analyses. For objective probabilistic verification statistics generation, we use the Model Evaluation Tools (MET; Development Test Center 2020) grid verification tools. We employ MET parameters Reliability, Likelihood, Calibration, ROC, ROC AUC, and Brier Score to obtain a reasonably complete summary of performance characteristics of each ensemble. Each of these metrics is described in section 4c.

a) Typhoon Maria (WP102018)—intensifying to 140 kt

To highlight differences in the two algorithms (WW3 run with/without post-processing) in an intensifying TC, we choose the Maria (WP102018). Maria, the eighth named storm of the 2018 typhoon season, was a powerful tropical cyclone that affected Guam, the Ryukyu Islands, Taiwan, and East China in early July 2018. Here we examine 96-h forecasts valid July 9, 2018 at 00 UTC, initiated on 5 July 00 UTC when the storm was located southeast of Guam and forecast to intensify as it moved toward Okinawa. Figure 2 shows details of the WW3TCOFCL Ensemble (left column) and WW3NAVGEM Ensemble (right column) forecasts of 12-ft seas. Consistent among the TCs inspected (approximately 30 cases) are that the WW3NAVGEM Ensemble input forecast tracks (Fig. 2 top row) and intensities both have reasonably large spread, but that ensemble member intensities tend to be too low, with intensities, unrealistically peaking near 70 kt for all members (Fig.2 second row). In comparison, the WSP tracks and intensities appear to be well-calibrated with individual forecasts encompassing the forecast, and thus provide more realistic wind forcing input to WW3. In the case of Maria, this results in large areas relatively weak wind forcing input to the WW3NAVGEM Ensemble, and much lower 12-ft significant height probabilities when compared to those from the WW3TCOFCL Ensemble (Fig.2, third row)—the issue is even more pronounced for higher significant wave height thresholds (not shown). These differences are not isolated, but seen throughout the data set, especially for developing TCs.

b) Hurricane Ileana (EP112018)—maintaining intensity at 40-45 kt

The majority of TCs are not forecast to intensify beyond 70 kt. To highlight differences between a weaker TC that is not forecast to intensify, we choose Hurricane Ileana’s 48-h forecast valid August 8, 2018 at 00 UTC, initiated on 6 August 00UTC. Ileana was a remarkably small TC and the ninth tropical storm in the East Pacific in 2018 and during its lifecycle tracked parallel to the Mexican coast. At this time, NHC forecasted Ileana to remain weak as it approached the Baja California Peninsula. In this case, the initial intensities used in the

WW3NAVGEM Ensemble encapsulate the initial estimate from NHC (Fig. 3, second row). The forecast track (Fig 3, top row) and intensity spreads (Fig. 3, second row) are larger than those produced from the WSP algorithm. The 12-ft seas probabilities forecasts (Fig.3, third row) from WW3TCOFCL Ensemble are still noticeably higher probabilities in the vicinity of the highest observed wave heights (Fig. 3, bottom row). Much of the difference in 12-ft significant wave height probabilities generated from the WW3TCOFCL Ensemble and WW3NAVGEM Ensemble can be explained by larger forecast track spread in the WW3NAVGEM Ensemble input.

c) Objective Scores

Once the analyses are limited to 20x20 degree boxes centered on the TC best track position, the probability forecasts can be inter-compared using standard probability metrics such as Reliability (Fig. 4), Discrimination (Fig. 5), Relative/Receiver Operating Characteristic (ROC; Fig. 6), and summary or derivative metrics such as Discrimination Distance, Area Under ROC Curve, and Brier Score (Fig. 7). Each of these metrics answers a specific question that we discuss below. Again, our evaluation uses MET, which in turn cites Wilks (2011) for most of its statistical algorithms. Results shown here are for a homogeneous data set, meaning that the scores from the two different algorithms can be compared since they are for the same TCs on the same dates. For ground truth we again use analyzed significant wave heights from the WW3TCOFCL deterministic model (Sampson et al. 2013) as these have been shown to have realistic TC structure. We also performed the same tests using WW3NAVGEM deterministic model analyzed significant wave heights for verification, but somewhat surprisingly found consistent results in both statistical analyses for the metrics chosen. Finally, the evaluation was conducted for 0, 1, 2, 3, 4, and 5 day forecasts, but we limit presentation of the Reliability, Discrimination and ROC charts to 1, 3, and 5 days and the results to those using the WW3TCOFCL model deterministic analysis as ground truth for brevity.

i) Reliability

Reliability determines how well the probabilities compare to observed frequencies. On a Reliability Diagram, perfect reliability is a diagonal (1:1) line from lower left to upper right, biases are indicated by model reliability being below (high bias) and above (low bias) the 1:1 line, and forecast confidence is provided by the slope of model reliability relative to the 1:1 line, that is under-confident when the slope is less than and overconfidence when the slope is greater than one (Wilks 2011). Reliability for both 12-ft and 18-ft significant wave height probabilities is shown in Fig. 4. The reliability for WW3TCOFCL Ensemble 12-ft significant wave height appears high biased (over-forecasting in Wilks 2011) throughout. The WW3NAVGEM Ensemble appears to overestimate low probabilities and underestimate higher probabilities in shorter forecast leads (under-confident), and overestimate probabilities like the WW3TCOFCL Ensemble does at longer forecast leads. The number of cases drops precipitously for the 120-h 18-ft significant wave height probabilities above 80%, dropping to 400 head-to-head cases or one grid (SH112019 verifying Mar 7 2019 at 12:00 UTC). So the Reliability Diagrams at 120 h for 18-ft significant wave height at the highest probability thresholds have few verification cases, reflected in the erratic changes in the reliability.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-21-0037.1.

In the case of the WW3NAVGEM Ensemble (under-confident in short-term forecast leads, over-forecasting at longer-term forecast leads), the authors suspect that the ensemble is challenged by resolution in that circulations tend to be too large at longer forecast leads. In the case of WW3TCOFCL Ensemble, the authors suspect several potential issues. The first is that WW3 is likely more appropriately run with 10-minute mean wind speeds since it is developed to use NWP fields. This is in contrast to U.S. official forecast center specified TC winds and wind probability realizations, which are both considered 1-minute wind speed estimates. Operational forecasters use conversion rates such as .93 (Harper et al. 2010) to convert the 1-minute wind speeds to 10-minute wind speeds, and this conversion would likely reduce the high bias. Another potential source of bias is the statistical wind radius model (DRCL; Knaff et al. 2007 and Knaff et al. 2018) used in the wind probabilities. DRCL wind radii become more symmetric as the forecast progresses in time, and these symmetric forecasts could provide unrealistic durations for TC winds. DRCL will never emulate the large symmetry fluctuations seen in nature. A more appropriate treatment of the asymmetries, especially at longer forecast periods, could provide more realistic changes in fetch and duration of winds around TCs.

ii) Discrimination

Discrimination is the relative frequency with which a forecast can discriminate between events and non-events, where perfect discrimination would entail no overlap between distributions of forecast probabilities for events and non-events. Discrimination Diagrams show these frequencies, where superior discrimination is indicated by separation between the events and non-events. Figure 5 shows discrimination for probabilities from our two algorithms at 1, 3, and 5 days. One obvious trend is that the separation between events and non-events becomes smaller as forecast length increases, as seen by the lines of the same color converging towards each other. The ability to discriminate between events and non-events drops with forecast lead time for both algorithms.

iii) Discrimination Distance

An easier way to visualize and summarize the discrimination is to graph the Discrimination Distance (the difference between the average of the event and non-events) for all forecast leads on one graph (Fig. 7). The Discrimination Distances for the WW3TCOFCL Ensemble are lower than for WW3NAVGEM Ensemble probabilities out to approximately 24 h, then remain approximately 10% higher for the longer leads. Significant differences using a 2-tailed t-test at the 5% level are present at all but the 24-h time period for 12-ft probabilities, and at all but 24-h and 120-h time periods for the 18-ft probabilities. Discrimination Distances for 12-ft are about 10% higher than for 18-ft significant wave heights at all forecast leads, indicating more skill in discrimination of 12-ft significant wave heights. The Discrimination Distances also decay at longer leads, indicating less skill in discrimination between events and non-events at these forecast leads times.

iv) ROC

ROC is another measure of the ability of the forecast to discriminate between two alternative outcomes, thus measuring resolution. It is not sensitive to bias in the forecast, so says nothing about reliability. A biased forecast may still have good resolution and produce a good ROC curve, which means that it may be possible to improve the forecast through calibration (e.g., correcting the bias). ROC can thus be considered as a measure of potential usefulness (Development Test Center, 2020). A perfect ROC curve follows the y axis from 0 to 1, then across the top of the diagram to 1, 1. The ROC degrades for both algorithms as forecast time increases (Fig. 6). This is true for both the 12-ft and 18-ft thresholds. .

v) ROC Area Under Curve

The area under a ROC Curve (ROC AUC) is a convenient way to summarize how a forecast discriminates between event/non-event (Wilks 2011). Values can theoretically go from 0 to 1. A perfect score is 1, describing the area under a curve that passes from x=0, y=0, through x=0, y=1, to x=1, y=1). The ROC AUC for the no-skill diagonal is .5 (the area under a diagonal from x=0, y=0 to x=1, y=1 on a ROC Diagram). As expected, the ROC AUC (Fig. 7) for the WW3TCOFCL Ensemble probabilities is relatively low at analysis time due to the many cold starts in our data set. The WW3TCOFCL Ensemble ROC AUC improves until about the 48-h forecast time, then gradually drops off through 120 h. The WW3NAVGEM Ensemble ROC AUC drops gradually through the forecast and is approximately 15% lower than the WW3TCOFCL Ensemble between 72 and 120 h. Differences in the ROC AUC pass significance tests at all forecast periods except at 48 h for 12-ft, and at 0 h for 18-ft significant wave height. The numbers of cases (each case representing an entire 20x20 degree grid) for this ROC AUC at 48, 72, 96, and 120 h are all well below 200, so conclusions on significance tests 18-ft significant wave height should await more cases. Recall that the high bias in the WW3TCOFCL Ensemble is not penalized in either the ROC or the ROC AUC, and that the ROC AUC is only used to discriminate between the event and non-event. It is encouraging that the WW3TCOFCL Ensemble probabilities maintain high ROC AUC out to 120 h since high bias, not depicted in either the ROC or ROC AUC, can be corrected through adjustments in the algorithm.

vi) Brier Scores

Brier Scores are another standard skill score for probabilistic forecasts, and measure both reliability and resolution (the ability to distinguish an event from a non-event). The Brier Score measures the mean square error of probabilities. Here again we use the WW3TCOFCL deterministic model analyses as ground truth. Brier Scores range from 0 to 1, 0 being a perfect score. Brier Scores for both ensembles evaluated are shown in Fig. 7 and they are within 3% of each other for both 12- and 18-ft thresholds. These generally rise as forecast time increase, indicating skill drops with forecast lead. The uptick in the WW3TCOFCL Ensemble at analysis time is expected as this ensemble was frequently cold started throughout the testing period and the WW3TCOFCL Ensemble (and its input) has little spread at analysis time. The WW3NAVGEM Ensemble probabilities have slightly lower Brier Scores than the WW3TCOFCL Ensemble probabilities at all forecast times for the 12-ft significant wave height threshold, and scores from the two algorithms are within 3% of each other. Differences for 12-ft probabilities are significant at all forecast periods. Brier Scores for 18-ft significant wave height thresholds are within 1% of each other with the WW3NAVGEM Ensemble scoring lower (better). Differences are significant at 24 and 96 h, but just barely pass significance tests. In the case shown in Fig. 2, the Brier Score for WW3NAVGEM Ensemble (0.082098) is lower than for WW3TCOFCL Ensemble (0.13089). This may seem counterintuitive as the WW3TCOFCL Ensemble probabilities "look" to capture the 12-ft significant wave heights in the WW3TCOFCL deterministic model analysis from 96 hours later. But upon further inspection (Table 2), the distribution of probability forecasts for WW3NAVGEM Ensemble is skewed to lower probabilities so that it scores much higher in the large number of non-events than the

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-21-0037.1.

WW3TCOFCL Ensemble probabilities for this case. The Brier score becomes inadequate for very rare (or very frequent) events because it does not sufficiently discriminate between small changes in forecast that are significant for rare events (Benedetti 2010). Thus, Brier Score unfairly penalizes extremely rare (or common) event forecasts and can actually leads to conclusions that disagree with our intuition (Jewson 2008), such as indicating that the WW3NAVGEM Ensemble outperforms the WW3TCOFCL Ensemble for the case in Fig. 2. The Brier Scores are still useful in our evaluation as they confirm high bias in the WW3TCOFCL Ensemble that, if corrected, could decrease the Brier Scores. However, tuning specifically to Brier Scores is not advised as that could result in undesired reduction in extreme event prediction (described as under-confident in Wilks 2011). An analog to this would be tuning a TC wind intensity consensus (e.g., see Sampson et al. 2008) to minimize mean forecast error when the most impactful errors are associated with rare and difficult to forecast rapid intensification events.

5. Conclusions and future work

A post-processing algorithm for insertion of real-time operational TC surface wind forecasts into a .25x.25 degree global 20-member ensemble surface wind field is described. This algorithm was run twice a day (at 00 and 12 UTC) for approximately one year and included active TCs from all basins. Each set of post-processed wind fields was then used as wind input to WW3 in order to generate a 20-member ensemble of forecasted significant wave height fields out to 5 days. The resultant significant wave height fields from each ensemble member were then compiled to create significant wave height probabilities on a 1x1 degree global grid.

Evaluation was performed using 20x20 degree boxes around verifying positions of the TCs at each forecast day using the MET statistics package. Both WW3NAVGEM and WW3TCOFCL deterministic model analyses were used as ground truth for evaluation of the probabilities and little difference was found between evaluations with the two ground truth datasets. Case studies indicated large discrepancies frequently existed between input winds from the two algorithms. NAVGEM Ensemble tracks and intensities generally had large spreads, and certainly larger than those generated by the WSP algorithm that are used in the WW3TCOFCL Ensemble for weaker TCs. WW3NAVGEM Ensemble input intensities were generally low-biased for intense TCs as the NAVGEM Ensemble resolution was challenged to represent steep wind gradients in relatively small TCs. Large discrepancies also existed between significant wave height probabilities generated by each of the ensemble forecasts. The WW3NAVGEM Ensemble significant wave height probabilities tended to be more widespread and lower in magnitude than those from the WW3TCOFCL Ensemble.

In objective evaluation, Reliability Diagrams show that WW3NAVGEM Ensemble overestimated low probabilities and underestimated higher probabilities in short-range forecasts, then generally overestimated probabilities by 5 days. WW3TCOFCL Ensemble generally overestimated all probabilities throughout the entire forecast. Brier Scores for WW3NAVGEM Ensemble were a few percent better than WW3TCOFCL Ensemble at 12-ft significant wave height forecasting at all forecast lengths, but inspection of individual cases indicated that those scores were heavily influenced by forecasts of very low probability for non-events (no 12-ft or 18-ft significant wave height in ground truth). Brier Scores for 18-ft significant wave height were within about 1% at all forecast lengths. ROC curves and ROC AUC indicated that discrimination between events and non-events degrades with forecast period for both sets of probabilities, but that WW3TCOFCL Ensemble forecast generally appeared better at discriminating events from non-events beyond 24 h. These results are confirmed by the Discrimination Diagrams, Discrimination Distances, and significance tests for Discrimination Distances.

The WW3TCOFCL Ensemble high bias noted in the Reliability Diagrams is likely correctable. Whether by converting the WW3TCOFCL Ensemble input 1-minute to 10-minute mean winds that are more representative of NWP model winds, by replacing the Wind Radii CLIPER Model (DRCL) with more realistic wind distribution realizations, or by applying some combination of the above, the high bias can be addressed. Also, the validation package developed in this work could be modified to validate whether changes in algorithms upstream of the WW3 ensembles (e.g., the WSP algorithm and the NAVGEM Ensemble) adversely affect the significant wave height probabilities. Operational forecasts are certain to improve in the future through use of new sensors, improved NWP representation of the vortex, and more advanced post-processing in the wind probability algorithm — all of which can affect these ensembles. Construction of TC-specific significant wave height probability verification was time-consuming, but the process to achieve this is in place and could be used as is or improved upon to validate TC-specific wave probabilities in the future. And addition of Object-Based Diagnostic Evaluation (MODE) verification available in MET may compliment the evaluation done within this work as it follows features (e.g., TCs) and reports statistics different than those here when comparing the features. That evaluation would be similar to and hopefully more rigorous than the 12-ft sea radii evaluation against operational NHC estimates as done in Sampson et al. (2016).

Acknowledgments: This effort is dedicated to our late friend and colleague Paul Wittmann, WAVEWATCH expert and all around great person. Publication of this work was graciously funded by the Office of Naval Research, Program Elements 0602435N and 0603207N and NOAA/NESDIS base funding. We would like to thank Chuck Skupniewicz at Fleet Numerical Meteorology and Oceanography Center, who provided both encouragement and funding to keep the WW3TCOFCL Ensemble project going. Thanks also to Chris Landsea at the National Hurricane Center and his advocacy for this type of effort - it is greatly appreciated. We also thank John Gotway and the entire MET Team, without their help and software this evaluation would be significantly more difficult. Finally, kudos to Jim Hansen for suggesting expanding the deterministic WW3TCOFCL model to generate probabilities for sortie decisions. The views, opinions, and findings contained in this report are those of the authors and should not be construed as an official National Oceanic and Atmospheric Administration or U.S. Government position, policy, or decision.

Data Availability: Both sets of ensemble significant wave height probabilities have been archived and are available on request; however, a non-disclosure agreement public release approval may be required to provide data.

8. References

Aijaz, S., J. D. Kepert, H. Ye, Z. Huang, and A. Hawksford, 2019: Bias correction of tropical cyclone parameters in the ECMWF Ensemble Prediction System in Australia. *Mon. Wea. Rev.,* **147**, 4261–4285, doi: https://doi.org/10.1175/MWR-D-18-0377.1

Benedetti, R., 2010: Scoring rules for forecast verification. *Mon. Wea. Rev.*, **138**, 203–211, doi: https://doi.org/10.1175/2009MWR2945.1

Cummings, J. A., and P. Wittmann, 2009: Navy implements data assimilation capability for its wave forecasting model. *JCSDA Quarterly,* No. 28, Joint Center for Satellite Data Assimilation, Camp Springs, MD, 2–3. [Available online at https://static1.squarespace.com/static/5bad1a12c2ff616821035c9f/t/5d1bc190f87d390001d7f83e/1562100113337/200909JCSDAQuarterly.pdf]. Accessed 6/10/2021.

DeMaria, M., J. A. Knaff, M. Brennan, D. Brown, C. Lauer, R. T. DeMaria, A. Schumacher, R. D. Knabb, D. P. Roberts, C. R. Sampson, P. Santos, D. Sharp, K. A. Winters, 2013: Operational tropical cyclone wind speed probabilities part I: Recent model improvements and verification, *Wea. Forecasting*, **28**, 586–602, doi: http://dx.doi.org/10.1175/WAF-D-12-00116.1

Development Testbed Center, 2020: Model Evaluation Tools (MET) [available on line at https://dtcenter.org/community-code/model-evaluation-tools-met] Accessed 6/10/2021.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-21-0037.1.

Frank, W. M., 1977: The structure and energetics of the tropical cyclone. I: Storm structure. *Mon. Wea. Rev.*, **105**, 1119–1135, doi:10.1175/1520-0493(1977)105<1119:TSAEOT>2.0.CO;2

Harper, B. A., J. D. Kepert, and J. D. Ginger, 2010: Guidelines for converting between various wind averaging periods in tropical cyclone conditions, World Meteorological Society, 64 pp. [Available online at https://library.wmo.int/index.php?lvl=notice_display&id=135#.X6gp8mR7mUk] Accessed 6/10/2021.

Jewson, S., cited. 2008: The problem with the Brier score. arXiv:physics/0401046v1 [physics.ao-ph]. [Available online at http://arxiv.org/abs/physics/0401046v1] Accessed 6/10/2021.

Klotz, B. W., and D. S. Nolan, 2019: SFMR Surface wind undersampling over the tropical cyclone life cycle. *Mon. Wea. Rev.*, **147**, 247–268, doi: https://doi.org/10.1175/MWR-D-18-0296.1

Knaff, J. A., C. R. Sampson, M. DeMaria, T. P. Marchok,J. M. Gross,and C. J. McAdie, 2007b: Statistical tropical cyclone wind radii prediction using climatology and persistence. *Wea. Forecasting*, **22**, 781–791, doi: https://doi.org/10.1175/WAF1026.1

Knaff, J. A., C. R. Sampson, and K. D. Musgrave, 2018: Statistical tropical cyclone wind radii prediction using climatology and persistence: Updates for the western North Pacific. *Wea. Forecasting*, **33**, 1093–1098, doi: https://doi.org/10.1175/WAF-D-18-0027.1

Marchok, T. P., 2002: How the NCEP Tropical Cyclone Tracker works. *Preprints, 25th Conf. on Hurricanes and Tropical Meteorology*, San Diego, CA, Amer. Meteor. Soc., P1.13. [Available online at http://ams.confex.com/ams/pdfpapers/37628.pdf] Accessed 6/10/2021.

Meissner, T., L Ricciardulli, and F.J. Wentz, 2017: Capability of the SMAP Mission to measure ocean surface winds in storms, *Bull. Amer. Meteor. Soc*., **98**, 1660-1677, doi: 10.1175/BAMS-D-16-0052.1

Mouche, A., B. Chapron, J. Knaff, Y. L. Zhao, B. Zhang, and C. Combot, 2019: Copolarized and cross-colarized SAR measurements for high-resolution description of major hurricane wind structures: Application to Irma Category 5 hurricane, *J Geophys Res-Oceans*, **124**, 3905–3922.

NCEP, 2020: WAVEWATCH III® Model. [Available on line at https://polar.ncep.noaa.gov/waves/wavewatch/] Accessed 12/06/2020.

O'Reilly, W. C., and R. T. Guza, 1993: A comparison of two spectral wave models in the Southern California Bight. *Coastal Eng*., **19**, 263–282.

Reul, N., and co-authors (2017): A new generation of tropical cyclone size measurements from space, *Bull. Amer. Meteorol. Soc.*, **98**, 2367–2386, doi: 10.1175/BAMS-D-15-00291.1

Sampson, C. R., J. L. Franklin, J. A. Knaff, and M. DeMaria, 2008: Experiments with a simple tropical cyclone intensity consensus. *Wea. Forecasting*, **23**, 304–312.

Sampson, C. R., J. S. Goerss, J. A. Knaff, B. R. Strahl, E. M. Fukada, and E. A. Serra, 2018: Tropical Cyclone Gale Wind Radii Estimates, Forecasts, and Error Forecasts for the Western North Pacific, *Wea. Forecasting*, **33**, 1081–1092.

Sampson, C. R., P. A. Wittmann, E. A. Serra, H. L. Tolman, J. Schauer, and T. Marchok, 2013: Evaluation of wave forecasts consistent with tropical cyclone wind forecasts, *Wea. Forecasting*, **28**, 287–294. doi: http://dx.doi.org/10.1175/WAF-D-12-00060.1

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-21-0037.1.

Sampson, C. R., J. Hansen, P. A. Wittmann, J. A. Knaff, and A. Schumacher, 2016: Wave probabilities consistent with official tropical cyclone forecasts, *Wea. Forecasting*, **31**, 2035–2045, doi: 10.1175/WAF-D-15-0093.1

Tolman, H. L., 1991: A third-generation model for wind waves on slowly varying, unsteady, and inhomogeneous depths and currents. *J. Phys. Oceanogr*., **21**, 782–797.

Tolman, H. L., B. Balasubramaniyan, L. D. Burroughs, D. V. Chalikov, Y. Y. Chao, H. S. Chen, and V. M. Gerald, 2002: Development and implementation of wind generated ocean surface wave models at NCEP. *Wea.Forecasting*, **17**, 311–333.

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

Wilks, D., 2011: Statistical methods in the atmospheric sciences, Elsevier, San Diego. 627 pp.

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

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-21-0037.1.

Table 1. Numbers of WW3TCOFCL Ensemble and WW3NAVGEM Ensemble cases (each being a 20x20 grid) gathered from real-time execution from May 26, 2018 00:00 UTC to March 18, 2019 00:00 UTC with 84 TCs occurring around the world during that period. Each 12-ft and 18-ft case required to have both ensemble forecasts and verifying WW3TCOFCL deterministic model analysis.

Tau	**0**	**24**	**48**	**72**	**96**	**120**
12-ft	347	319	261	214	155	112
18-ft	347	319	261	214	155	112

Table 2. Contingency Table for WW3NAVGEM Ensemble (left) and WW3TCOFCL Ensemble (right) greater than 12-ft significant wave height probabilities for the 96-h forecast case shown in Figure 2. Observed Yes and Observed No for the 20x20 degree grid encompassing the verifying TC position in Figure 2.

	WW3NAVGEM Ensemble Matched Pairs		WW3TCOFCL Ensemble Matched Pairs	
Prob	**Ob Yes**	**Ob No**	**Ob Yes**	**Ob No**
0.05	0	82	0	56
0.15	0	58	0	37
0.25	0	51	0	38
0.35	0	42	0	37
0.45	10	32	1	33
0.55	16	18	1	35
0.65	14	1	5	29
0.75	28	2	14	12
0.85	32	0	15	5
0.95	14	0	78	2

Table 3. Numbers of cases for Reliability, Discrimination and ROC shown in Figures 4-6.

WW3NAVGEM Ensemble			WW3TCOFCL Ensemble		WW3NAVGEM Ensemble		WW3TCOFCL Ensemble	
24h 12ft			24h 12ft		24h 18ft		24h 18ft	
Prob	Ob Yes	Ob No	Ob Yes	Ob No	Ob Yes	Ob No	Ob Yes	Ob No
0.05	102	76240	1326	79715	59	99054	169	99583
0.15	169	7871	565	6578	129	3689	193	3299
0.25	319	4397	609	3330	246	1804	267	1432
0.35	517	3087	596	2083	344	876	306	747
0.45	813	2093	744	1429	412	560	335	513
0.55	1049	1499	870	1058	432	238	323	330
0.65	1484	923	960	799	428	119	384	235
0.75	1821	461	1082	671	392	53	329	165
0.85	2189	204	1357	560	353	21	355	89
0.95	5363	64	5717	645	430	0	564	46
72h 12ft			72h 12ft		72h 18ft		72h 18ft	
Prob	Ob Yes	Ob No	Ob Yes	Ob No	Ob Yes	Ob No	Ob Yes	Ob No
0.05	440	39933	134	28743	326	59769	123	52648
0.15	622	9051	278	11742	332	5526	140	8876
0.25	818	4929	548	7149	324	2156	216	3807
0.35	907	3019	661	4478	314	1132	320	1995
0.45	983	2031	798	3265	365	620	469	1259
0.55	1048	1324	878	2604	387	363	520	720
0.65	1261	751	1213	1631	347	170	477	336
0.75	1362	450	1498	1224	307	41	371	120
0.85	1463	252	1782	680	118	6	188	33
0.95	1934	71	3048	309	46	0	42	3
120h 12ft			120h 12ft		120h 18ft		120h 18ft	
Prob	Ob Yes	Ob No	Ob Yes	Ob No	Ob Yes	Ob No	Ob Yes	Ob No
0.05	438	20507	253	14547	316	30250	212	27093
0.15	603	4695	327	5347	298	3549	229	5250
0.25	542	2512	369	4099	208	1494	280	2306
0.35	457	1750	390	2803	189	879	261	1184
0.45	427	1187	560	2200	213	579	244	743
0.55	522	839	744	1481	147	313	153	390
0.65	557	648	650	985	95	195	118	349
0.75	688	438	819	790	61	98	78	118
0.85	621	216	784	408	61	42	26	71
0.95	1149	217	1108	354	13	13	0	23

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-21-0037.1.

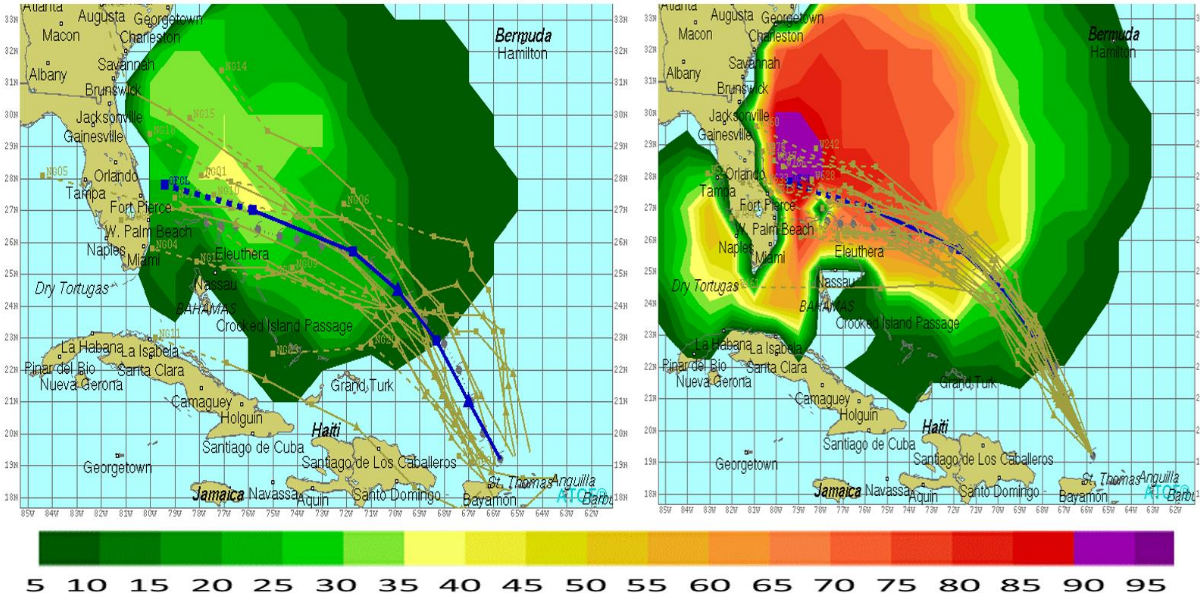

Figure 1. (left) WW3NAVGEM Ensemble and (right) WW3TCOFCL Ensemble 96-h forecast 12-ft sig wave ht probabilities for Dorian (AL052019) on Aug 29, 2019 at 00UTC. National Hurricane Center forecast track (blue) is shown for reference. Also, (right) NAVGEM Ensemble TC tracks and (right) wind probability realizations generated by the U.S. TC forecast center wind probabilities (brown) included. Probability (%) colorbar is shown at the bottom.

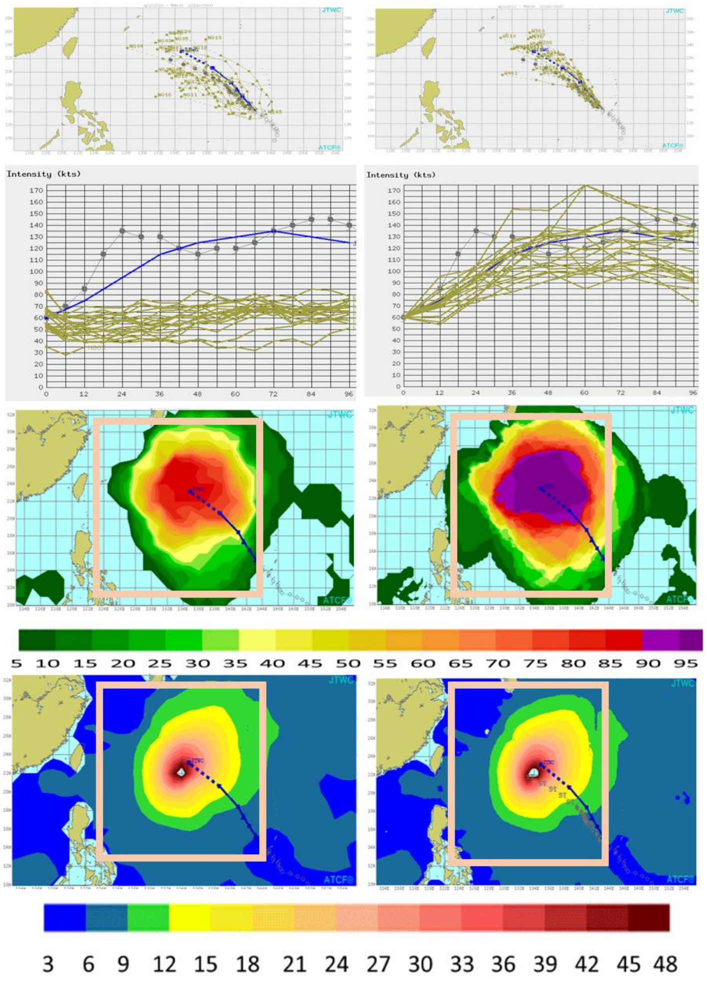

Figure 2. (left) WW3NAVGEM Ensemble and (right) WW3TCOFCL Ensemble. Input 96-h forecast tracks (top row), input forecast and verifying intensities (brown lines and black typhoon symbols, second row), 96-h forecast 12-ft sig wave ht probabilities (third row) and verifying significant wave height (ft) analyses (fourth row) for WW3NAVGEM deterministic model (left) and WW3TCOFCL deterministic model (right). Forecasts and analyses valid July 9, 2018 at 00 UTC for Maria (WP102018). Significant wave heights for this case are above the end of the color bar (48 ft). Joint Typhoon Warning Center forecast track and intensity (blue) is shown for reference. Verifying track labeled “ST” for Super Typhoon is shown (brown) in bottom right panel.

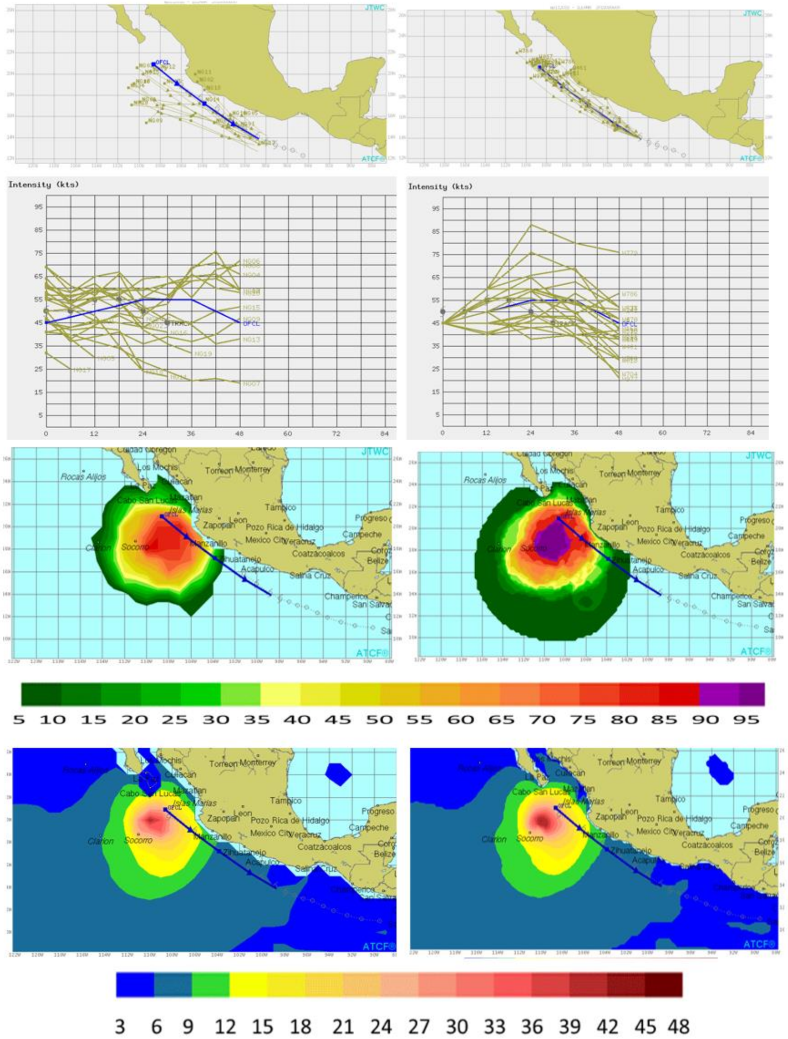

Figure 3. (left) WW3NAVGEM Ensemble and (right) WW3TCOFCL Ensemble. Input 96-h forecast tracks (top row), input forecast and verifying intensities (brown lines and black typhoon symbols, second row), 96-h forecast 12-ft sig wave ht probabilities (third row) and verifying significant wave height (ft) analyses (fourth row) for WW3NAVGEM deterministic model (left) and WW3TCOFCL deterministic model (right). Forecasts and analyses for Ileana (EP112018) 48-h forecast valid August 8, 2018 at 00 UTC. Significant wave heights for this case are above the end of the color bar (48 ft). National Hurricane Center forecast track and intensity (blue) is shown for reference.

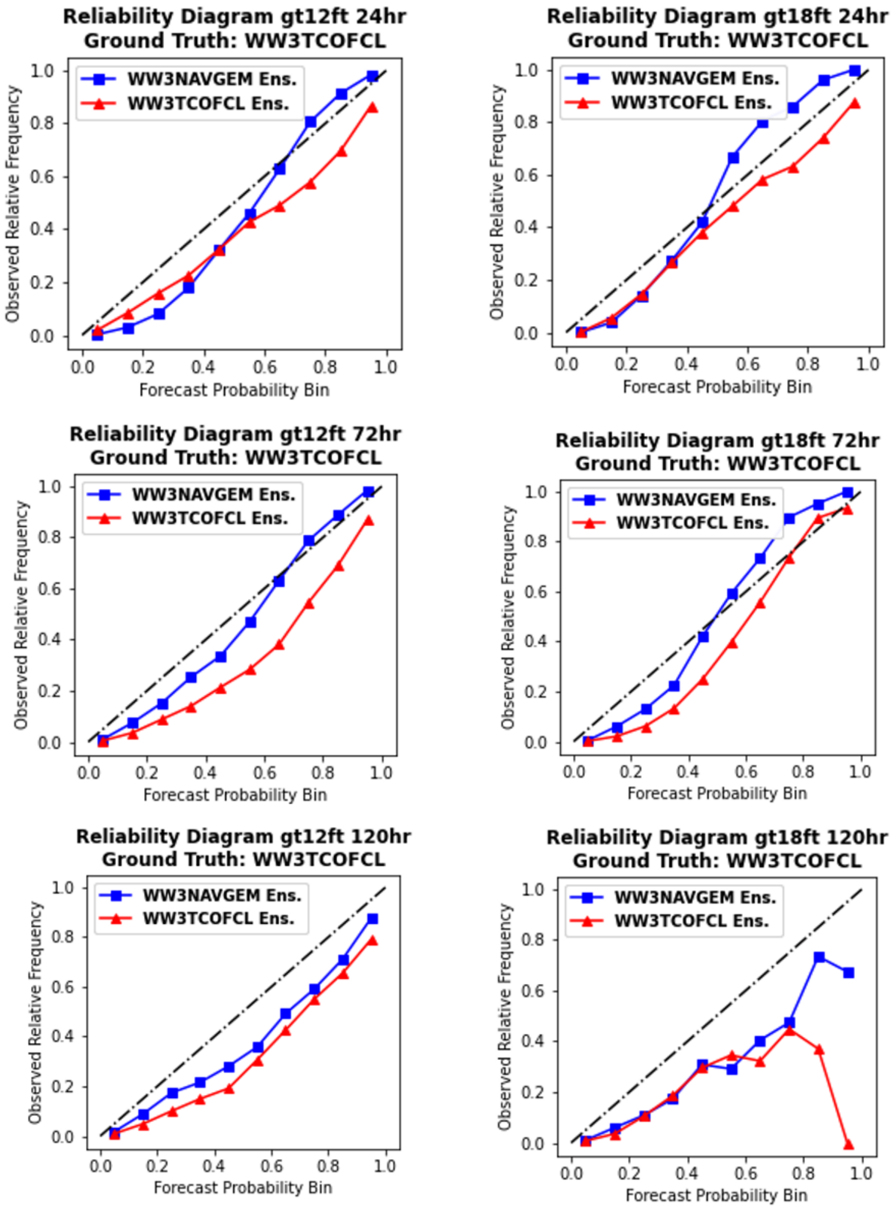

Figure 4. Reliability Diagrams for WW3TCOFCL Ensemble and WW3NAVGEM Ensemble 12-ft (left) and 18-ft significant wave height (right) with WW3TCOFCL deterministic model analysis employed as ground truth. Sequence progresses from (top) 24-h to (middle) 72-h to (bottom) 120-h forecast. See Table 3 for numbers of head-to-head cases. Dashed lines represent perfect reliability.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-21-0037.1.

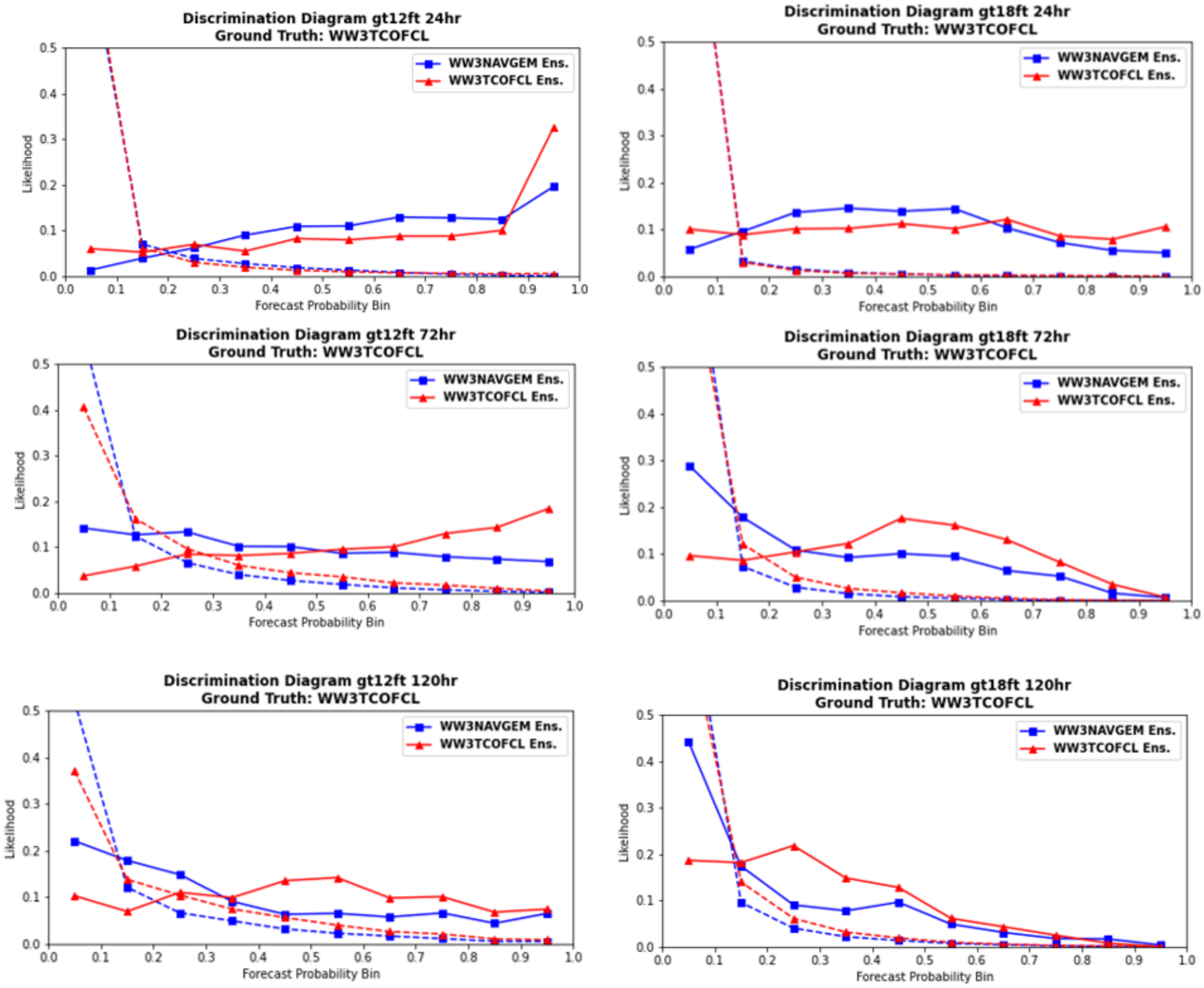

Figure 5. Discrimination Diagrams for WW3TCOFCL Ensemble and WW3NAVGEM Ensemble 12-ft (left) and 18-ft significant wave height (right) with WW3TCOFCL deterministic model analysis employed as ground truth. Solid lines indicate Observed Yes, dashed lines indicate Observed No distributions. Sequence progresses from (top) 24-h to (middle) 72-h to (bottom) 120-h forecast. See Table 3 for numbers of head-to-head cases.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-21-0037.1.

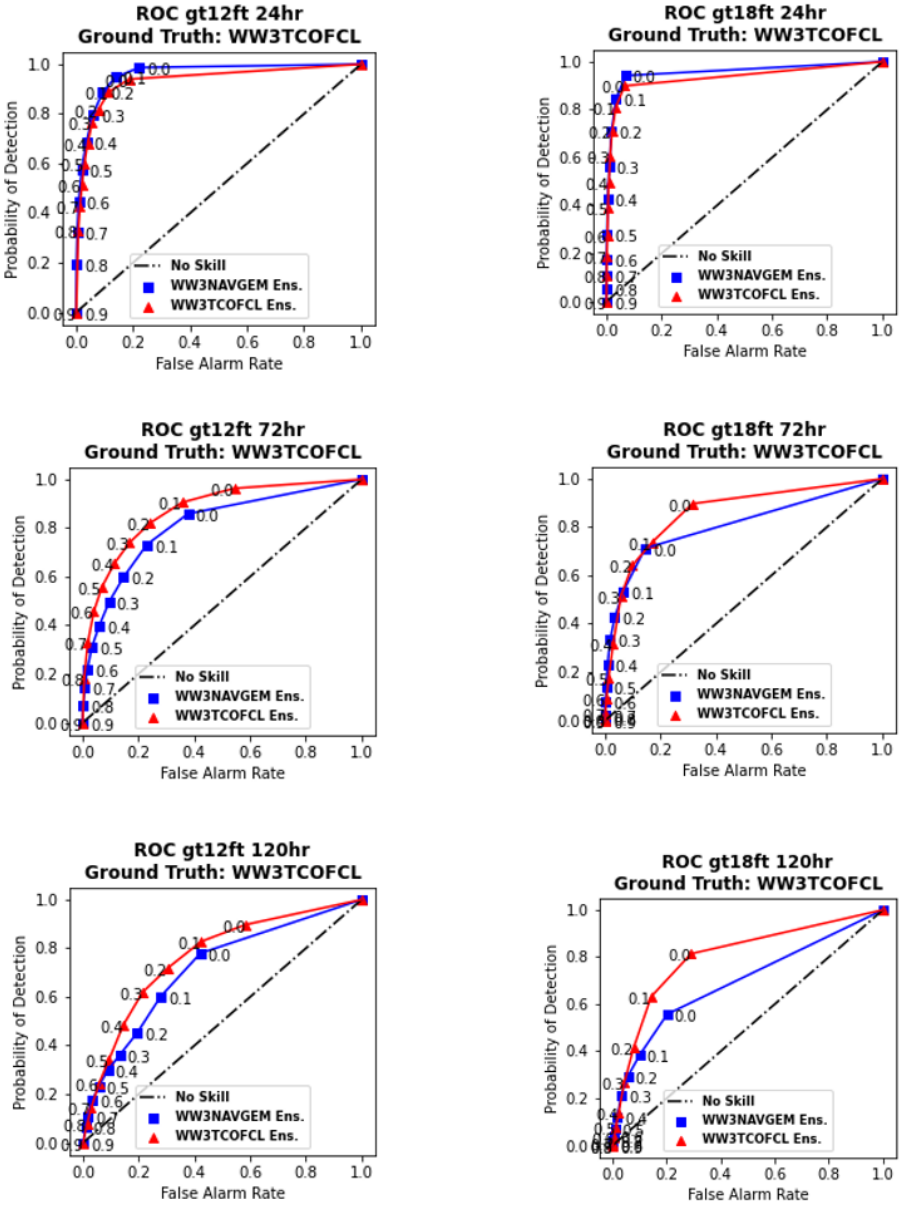

Figure 6. ROC Diagrams for WW3NAVGEM Ensemble (left) and WW3TCOFCL Ensemble (right) with WW3TCOFCL deterministic model analysis employed as ground truth. Sequence progresses from (top) 24-h to (middle) 72-h to (bottom) 120-h forecast. Dashed line indicates no skill. See Table 3 for numbers of head-to-head cases.

Accepted for publication in *Weather and Forecasting.* DOI 10.1175/WAF-D-21-0037.1.

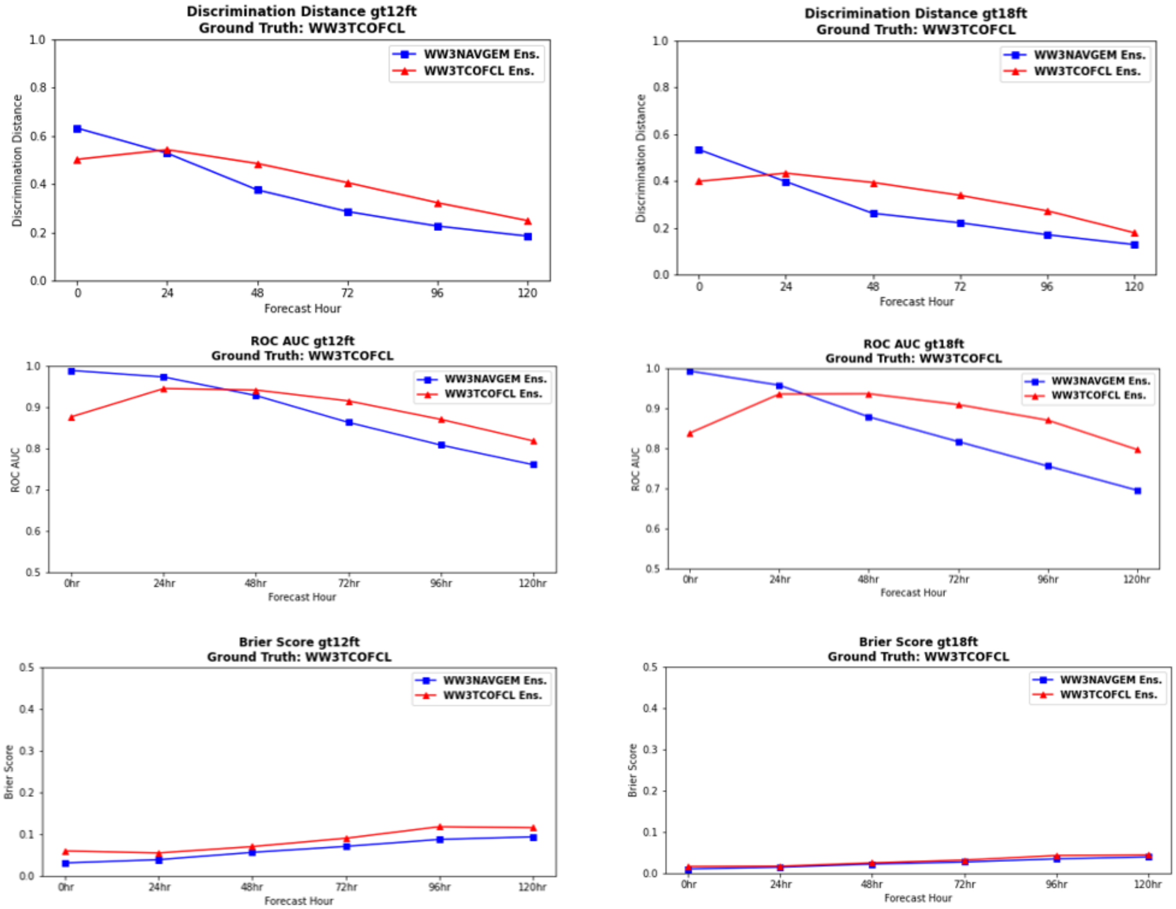

Figure 7. Discrimination Distances (top), ROC AUC (middle), and Brier Scores (bottom) for WW3TCOFCL Ensemble and WW3NAVGEM Ensemble. 12-ft (left) and 18-ft significant wave height (right) shown with WW3TCOFCL deterministic model analysis employed as ground truth. Sequence progresses from (top) 24-h to (middle) 72-h to (bottom) 120-h forecast. See Table 3 for numbers of head-to-head cases.