DISTRIBUTION STATEMENT A. Approved for public release; distribution is unlimited.

A Multi-Model Ensemble for Western North Pacific Tropical Cyclone Intensity Prediction

Jonathan R. Moskaitis
Naval Research Laboratory
Monterey, CA 93943-5502
phone: (831) 656-4057 fax: (831) 656-4769 e-mail: jon.moskaitis@nrlmry.navy.mil

Award Number: N0001414WX20469

LONG-TERM GOALS

The long-term goal of the work begun in this one-year project is to develop a new Western North Pacific tropical cyclone (TC) intensity guidance product for the Joint Typhoon Warning Center (JTWC), based on an international multi-model ensemble of high-resolution dynamical TC prediction models. The guidance tool will provide a best estimate and quantitative uncertainty information for TC intensity, as well as for TC position and the extent of the TC surface wind field. This project will lay the groundwork for a more sophisticated future guidance product based on an "ensemble-of-ensembles", a collection of ensemble forecasts from the participating high resolution dynamical TC models rather than a collection of deterministic forecasts. The ensemble-of-ensembles will facilitate generation of state-dependent impacts-based probabilistic forecasts, such as the probability of exceeding hurricane-force winds at a given place and forecast lead time, for JTWC and military customers.

OBJECTIVE

The objectives for this one-year project are as follows:

1. Utilize existing deterministic forecasts of Western Pacific TCs from high-resolution dynamical TC prediction models to assess the potential of an ensemble mean intensity forecast (commonly known as a "consensus" forecast) to reduce intensity errors relative to its individual model components.
2. Leveraging complementary work conducted under the Hurricane Forecast Improvement Program (HFIP), develop visualization and validation techniques to evaluate the potential benefits of a combined ensemble of three high-resolution dynamical TC prediction model ensembles (i.e. an "ensemble of ensembles")

APPROACH

Objective 1: Collect existing real-time and retrospective intensity forecasts from the past three Western Pacific typhoon seasons (2012, 2013, and the ongoing 2014 season) from high-resolution dynamical TC prediction models. Assess the performance of the consensus intensity forecast relative to its individual components through validation of the forecasts with respect to the JTWC best track

observations. To more fully understand the results, further evaluate the intensity forecasts within the context of the theory of consensus prediction.

Objective 2: Retrospective ensemble forecasts for 132 Atlantic basin forecast cases were run with the HWRF model, GFDL model (these forecasts were conducted by our NOAA partners in HFIP), and COAMPS-TC model (run by NRL for HFIP). The approach here is to use these forecasts as a testbed to explore visualization techniques for an ensemble-of-ensembles and to validate the performance of the ensemble mean and ensemble distribution.

Key personnel: Jonathan Moskaitis (NRL), James Doyle (NRL), Melinda Peng (NRL).

WORK COMPLETED

Objective 1: Three models were identified as essential components of the consensus intensity forecast for Western Pacific TCs in 2012-2014: COAMPS-TC, HWRF, and GFDN. Intensity forecasts from three models are available for the vast majority of the 6-hourly watches for TCs in 2012-2014. Ideally, we would like to use forecasts from the same model configuration for all three years, but since we are depending on existing forecasts this was not possible. The GFDN forecasts, run operationally by FNMOC, were obtained from the ATCF a-decks. The HWRF forecasts, run in real-time for the Western Pacific basin under HFIP, were also obtained from the ATCF a-decks. For 2014, the COAMPS-TC forecasts used are the operational forecasts obtained from the a-decks. In June of 2014, FNMOC implemented an updated version of COAMPS-TC developed by NRL, replacing the version used operationally in 2013. For 2013 TCs, we did not use the operational COAMPS-TC forecasts, but rather retrospective forecasts conducted by NRL using the updated version of the model. For 2012, we used real-time COAMPS-TC forecasts run with an older version of the model. Using the aforementioned data set, the 3 models and their consensus were validated with respect to JTWC best track observations. The results and their relationship to consensus theory are presented in the next section.

Objective 2: Intensity products for the COAMPS-TC/HWRF/GFDL combined ensemble were developed. The individual model ensemble means and combined ensemble means were validated with respect to the NHC best track observations. The relationship between ensemble mean error and ensemble spread was assessed, and the reliability of probabilistic intensity predictions was validated.

RESULTS

Validation of 2012-2014 Western Pacific TC consensus intensity predictions

Figure 1 shows the key validation results regarding the 2012-2014 Western Pacific TC intensity forecast consensus. The figure shows intensity mean absolute error (MAE) and mean error (ME) as a function of lead time for COAMPS-TC, HWRF, GFDN and the consensus of the aforementioned three models. MAE is the standard metric for deterministic forecast accuracy for TCs, while ME is a measure of forecast bias. The statistics are calculated using a very large homogeneous sample of over 1200 forecasts (sample size shown in the bottom panel of Fig. 1), although these cannot be considered truly independent samples due to serial correlation of errors in 6-hourly forecasts for a given TC.

It can be seen in Fig. 1 that the MAE of the consensus forecast is lower that all three of its component models for all positive lead times. The difference between the MAE of the consensus and the MAE of

the best component model (HWRF or COAMPS-TC depending on the lead time) is typically 2 to 3 kt for lead times beyond 18 h. This is a substantial increase in performance for the consensus forecast relative to its component models.

Fig 2 shows the intensity MAE and ME for each individual year (2012, 2013, and 2014) rather than all three years combined. Model performance varies from year-to-year depending on the particular TCs that occur and changes that are made to the model before each northern hemisphere TC season. In 2012 and 2014, the consensus is clearly the most accurate forecast, and in 2013 it performs similarly to HWRF. The yearly results show that the good performance of the consensus is fairly robust despite year-to-year variations in the performance of its component models.

The theory underlying consensus prediction can help in diagnosing why the consensus intensity forecast outperforms its component models in Fig. 1. Consider the simplest possible consensus, consisting of two models, "model 1" and "model 2". Let $p_1(x)$, with mean $\mu_1$and variance σ_1^2, represent the intensity error distribution for model 1. Similarly, let $p_2(x)$, with mean μ_2 and variance σ_2^2, represent the intensity error distribution for model 2. The intensity error distribution for the consensus of model 1 and model 2, $p_c(x)$, has the following mean and variance:

$$\mu_c = \frac{1}{2}\mu_1 + \frac{1}{2}\mu_2 \qquad \text{Eq. (1)}$$

$$\sigma_c^2 = \frac{1}{4}\sigma_1^2 + \frac{1}{4}\sigma_2^2 + \frac{1}{2}\rho\sigma_1\sigma_2 \qquad \text{Eq. (2)}$$

The consensus mean error is just the average of the model 1 mean error and model 2 mean error. Note that if the model 1 mean error and model 2 mean error are of opposite sign, the consensus mean error can be closer to zero than both. In other words, the consensus forecast can have a better bias (mean error closer to zero) than either of its component models. This principle holds true for any number of component models, not just two. The variance of the consensus error distribution, shown in Eq. (2), involves three terms: the first term depends only on the model 1 error variance, the second term depends only on the model 2 error variance, and the third term depends the model 1 error variance, the model 2 error variance, and the correlation coefficient for the model 1 and model 2 errors, ρ. The consensus error variance can be lower than those of both component models depending on ρ and the relative values of σ_1^2 and σ_2^2; this situation is mostly likely to occur if σ_1^2 and σ_2^2 are similar in value and ρ is low, indicating low error correlation between model 1 and model 2. This line of reasoning holds true for any number of component models.
The mean squared error (MSE) is an alternative forecast accuracy measure to MAE, based on squared errors instead of absolute errors. The mean squared error of a forecast model can be written as

$$\text{MSE} = \mu^2 + \sigma^2, \qquad \text{Eq. (3)}$$

the sum of the square of the mean error plus the error variance (the MAE cannot be decomposed in this fashion). Following the discussion concerning Eq. 1 and Eq. 2, the MSE of the consensus forecast can be lower than the MSE of its component models. The following three factors promote a consensus forecast that outperforms all of its component models:

(1) Consensus mean error is closer to zero than the means errors of its component models.

(2) Error variances of the component models are similar.

(3) Errors of component models are poorly correlated (or, better yet, anti-correlated)

The three factors listed above can be evaluated for the models and consensus forecast shown in Fig. 1. The dashed lines in Fig. 1 show the mean errors; here the consensus mean error is further from zero than the mean errors for HWRF and GFDL, so this is not a helpful factor for the performance of the consensus. Fig. 3 examines the second and third factors listed above. The upper panel shows that the intensity error variances of COAMPS-TC, HWRF, and GFDN are fairly similar to each other at all lead times. The lower panel shows that the intensity errors of the three models are positively correlated with each other, but the correlation is well below one (generally between 0.5 and 0.7) for all three model pairs. Thus, it is the similarity of the component model error variances and the modestly positive error correlations amongst COAMPS-TC, HWRF, and GFDN that enable the consensus forecast to outperform all of its component models in Fig. 1.

Finally, let us examine the possibility of adding an additional model to COAMPS-TC, HWRF and GFDN in order to make a 4-model consensus forecasts rather than a 3-model consensus forecast. The additional model considered here is CHIPS, an axisymmetric dynamical intensity model. CHIPS is available in the ATCF a-decks about 80% of the time for 2012-2014 West Pacific TCs. The results for the 4-model intensity consensus are shown in Fig. 4. Like the 3-model consensus in Fig. 1, the 4-model consensus in Fig. 4 has a lower intensity MAE than all of its component models. Fig. 5 shows a comparison of the intensity performance for the 3-model consensus and the 4-model consensus, for the set of cases in which both versions of the consensus can be calculated (somewhat limited by the availability of CHIPS). The 4-model consensus has a lower intensity MAE than the three model consensus at all lead times, despite having a larger magnitude mean error due to the inclusion of CHIPS (due to negative bias of CHIPS at later lead times; see dashed orange line in Fig. 4). However, Fig. 6 shows that CHIPS has an error variance similar to that of the other three models, and CHIPS intensity errors have lower correlations with those of COAMPS-TC, HWRF and GFDN than COAMPS-TC, HWRF and GFDN have amongst each other. Even though CHIPS forecast have a substantial bias, because the error variance is similar to the other models and its errors are relatively independent of those of the other models, it makes a positive impact on the consensus performance.

The COAMPS-TC/HWRF/GFDN combined "ensemble of ensembles"

As far as we know, this is the first ever visualization/validation effort for an ensemble of high-resolution dynamical TC prediction model ensembles. As such, these results are preliminary as visualization and validation techniques will evolve with more experience.

Fig. 7 shows examples of a boxplot display for ensemble intensity forecasts. The four panels pertain to the combined COAMPS-TC/HWRF/GFDL ensemble, COAMPS-TC ensemble, HWRF ensemble, and GFDL ensemble for a particular forecast of Ernesto (05L, 2012). The boxplots make it easy to see the spread of the ensemble forecast, which is much larger for COAMPS-TC than for HWRF and GFDL. The envelope of ensemble solutions encompasses the observed intensity for COAMPS-TC (and also for the combined ensemble), but not for the HWRF ensemble or GFDL ensemble. This is just one case, but is representative of the performance of the combined ensemble and 3 individual ensembles in many other forecast cases.

Intensity verification summary statistics are shown in Fig. 8 for the 132-case Atlantic basin retrospective forecast sample. MAE and ME are displayed for the ensemble mean and ensemble

control (i.e. unperturbed forecast) for each component model. The MAE of the ensemble mean is nearly always lower than that of the ensemble control, particularly for COAMPS-TC and GFDL. Fig. 8 also displays results for the combined ensemble mean. The MAE of the combined ensemble mean is overall superior to the ensemble means of the three component models.

Beyond the deterministic forecast summary statistics described above, it is important to validate how well an ensemble represents the uncertainty of the intensity forecast. Fig. 9 shows comparisons of the average error of the ensemble mean ("skill") versus the average spread of the ensemble about its mean ("spread"). We would like to see the spread and skill roughly equal; typically ensembles have too little spread relative to skill ("underdispersive"). Of the three component ensembles, all are underdispersive (COAMPS-TC looks the best). Despite badly underdispersive forecasts from GFDL and HWRF, the combined ensemble spread-skill relationship is quite reasonable, although still clearly underdispersive. Rank histograms for the ensemble intensity forecasts are shown in Fig. 10. These diagrams diagnose the reliability of the ensemble. For a perfectly reliable ensemble, if it predicts an X% chance of an event, that event would occur X% of the time when such a prediction is made. A perfectly reliable ensemble would have a flat rank histogram. Similar to the spread-skill results, the rank histograms suggest that the COAMPS-TC and combined ensembles perform best.

IMPACT/APPLICATIONS

There is some potential for a useful consensus intensity forecast aid for Western Pacific TCs based on existing deterministic dynamical TC models, and perhaps including additional models from international partners in the future (particularly if their forecasts have errors independent of COAMPS-TC, GFDN, HWRF, and CHIPS). Further work is needed to assess the performance of the aforementioned consensus intensity aid against statistical intensity prediction models (e.g. LGEM, DSHP) and the statistical-dynamical model consensus. Based on the preliminary results from the COAMPS-TC/HWRF/GFDL combined ensemble in the Atlantic basin, it appears that the ensemble-of-ensembles concept for the Western Pacific also has operational potential further down the road.

TRANSITIONS

Future 6.2 and 6.4 programs would be needed to transition products to operations.

RELATED PROJECTS

The related projects include ONR 6.2 "Improvement of High Resolution Tropical Cyclone Structure and Intensity Forecasts using COAMPS-TC" and the Hurricane Forecast Improvement Program (HFIP).

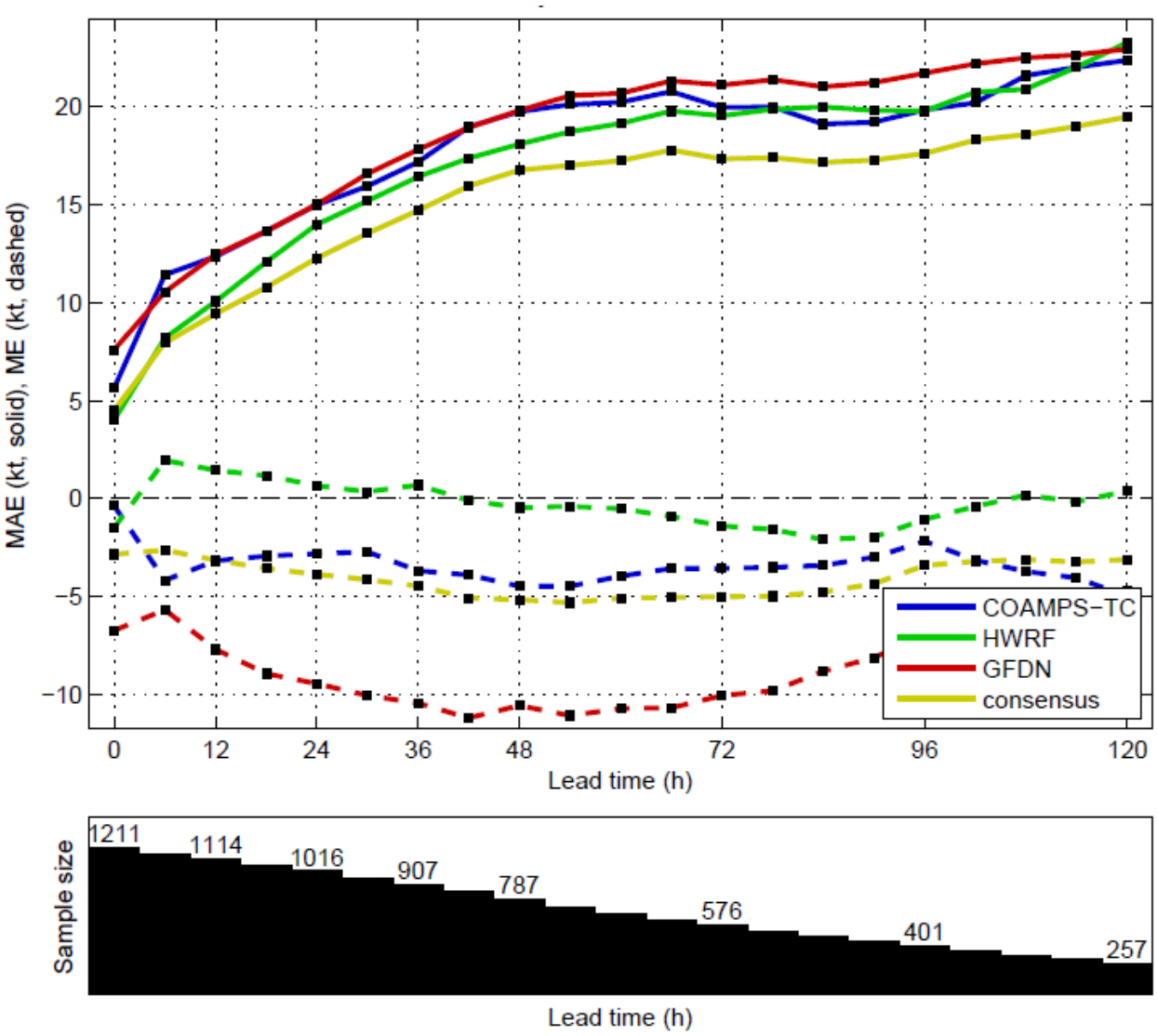

Figure 1: Intensity mean absolute error (MAE, solid) and intensity mean error (ME, dashed) as a function of lead time for 2012-2014 Western Pacific tropical cyclone forecasts from COAMPS-TC (blue), HWRF (green), GFDN (red), and the consensus of the 3 aforementioned models (yellow). The lower panel shows the sample size as a function of lead time for this homogeneous comparison.

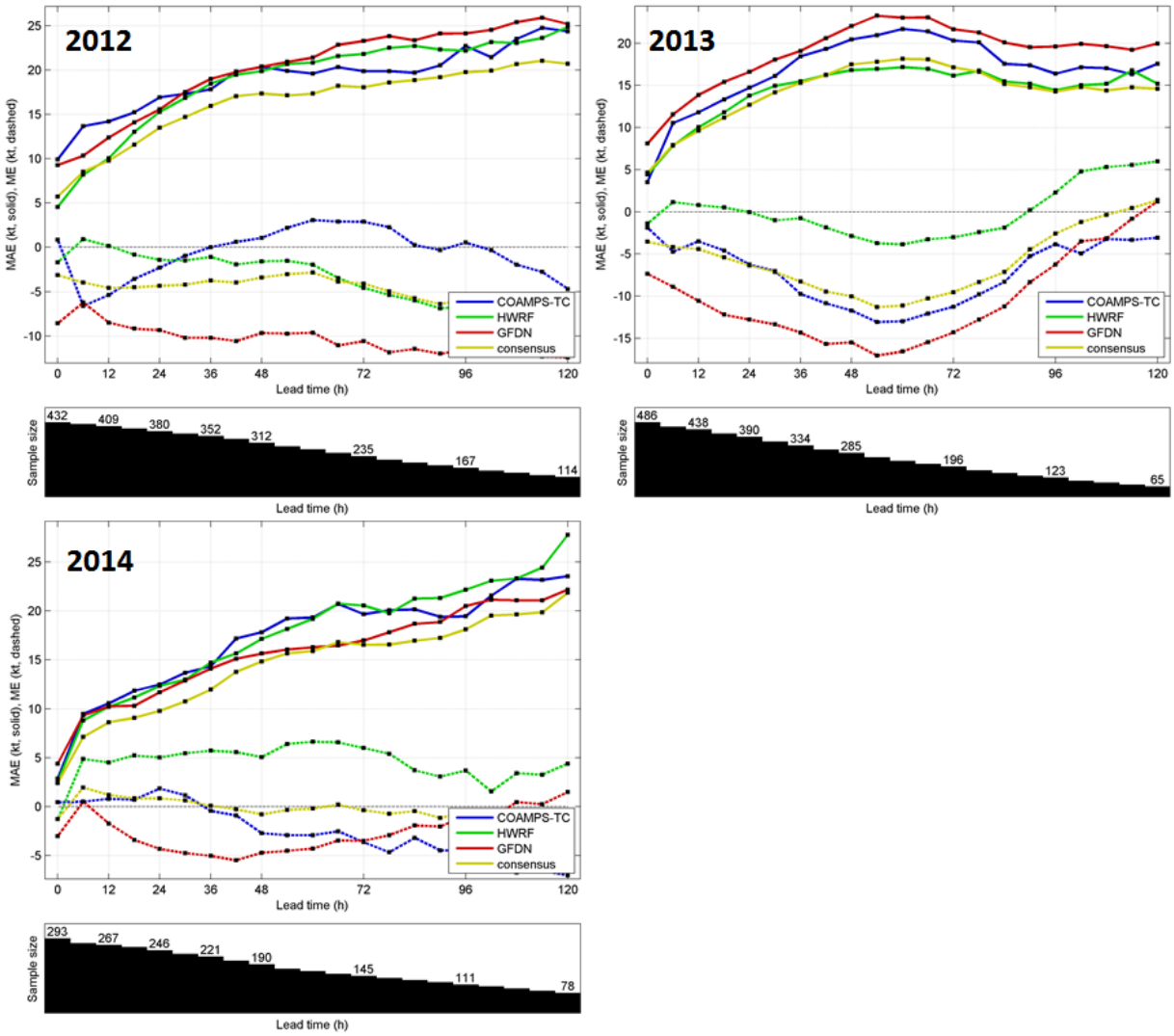

Figure 2: As in Figure 1, except results are presented for 2012, 2013, and 2014 individually rather than accumulating the statistics for the entire 3-year period.

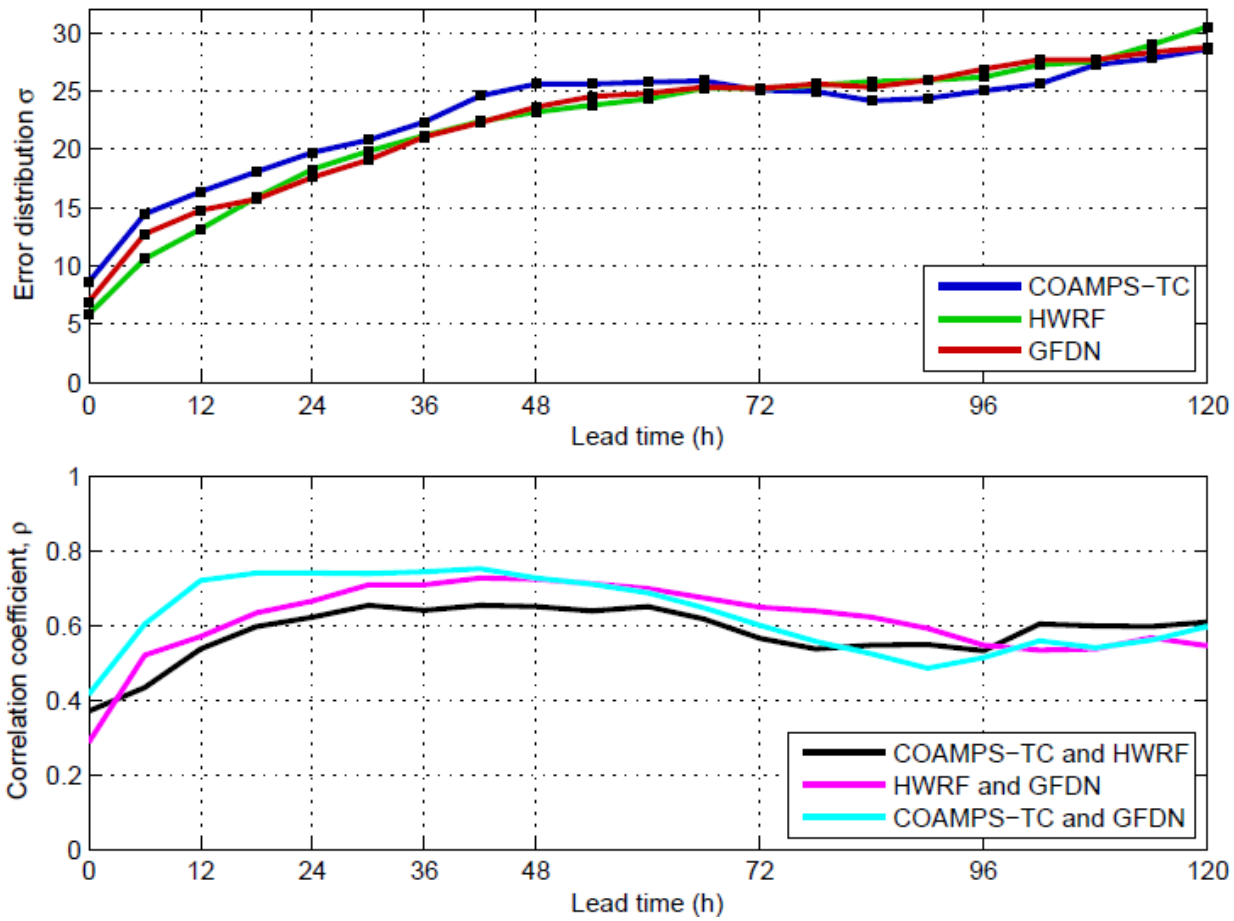

Figure 3: Upper panel shows the square root of the intensity error variance for the COAMPS-TC, HWRF, and GFDN forecasts. The lower panel shows the correlation coefficient between intensity forecasts as a function of lead time, for the three possible model pairs.

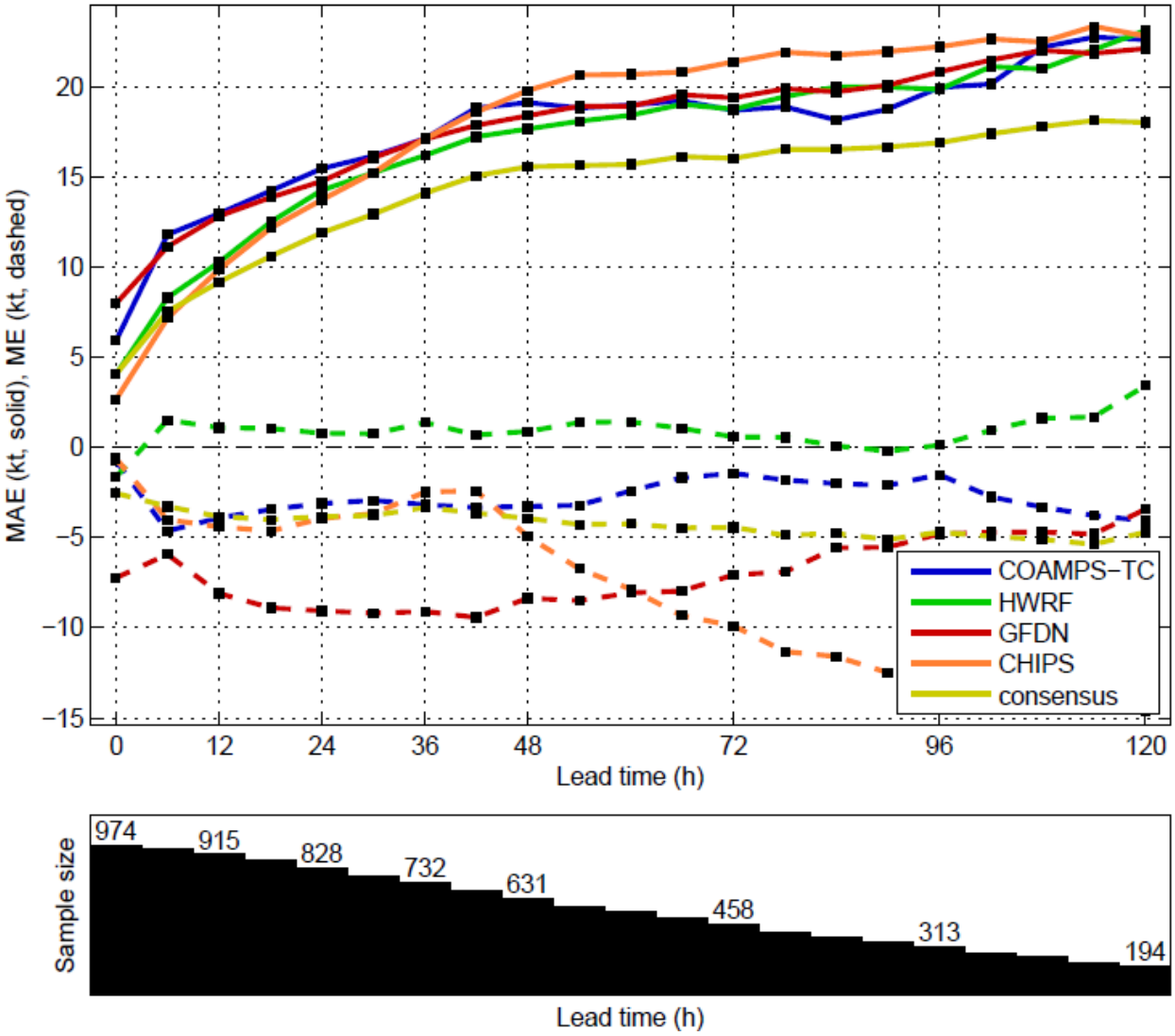

Figure 4: Intensity mean absolute error (MAE, solid) and intensity mean error (ME, dashed) as a function of lead time for 2012-2014 Western Pacific tropical cyclone forecasts from COAMPS-TC (blue), HWRF (green), GFDN (red), CHIPS (orange), and the consensus of the 4 aforementioned models (yellow). The lower panel shows the sample size as a function of lead time for this homogeneous comparison.

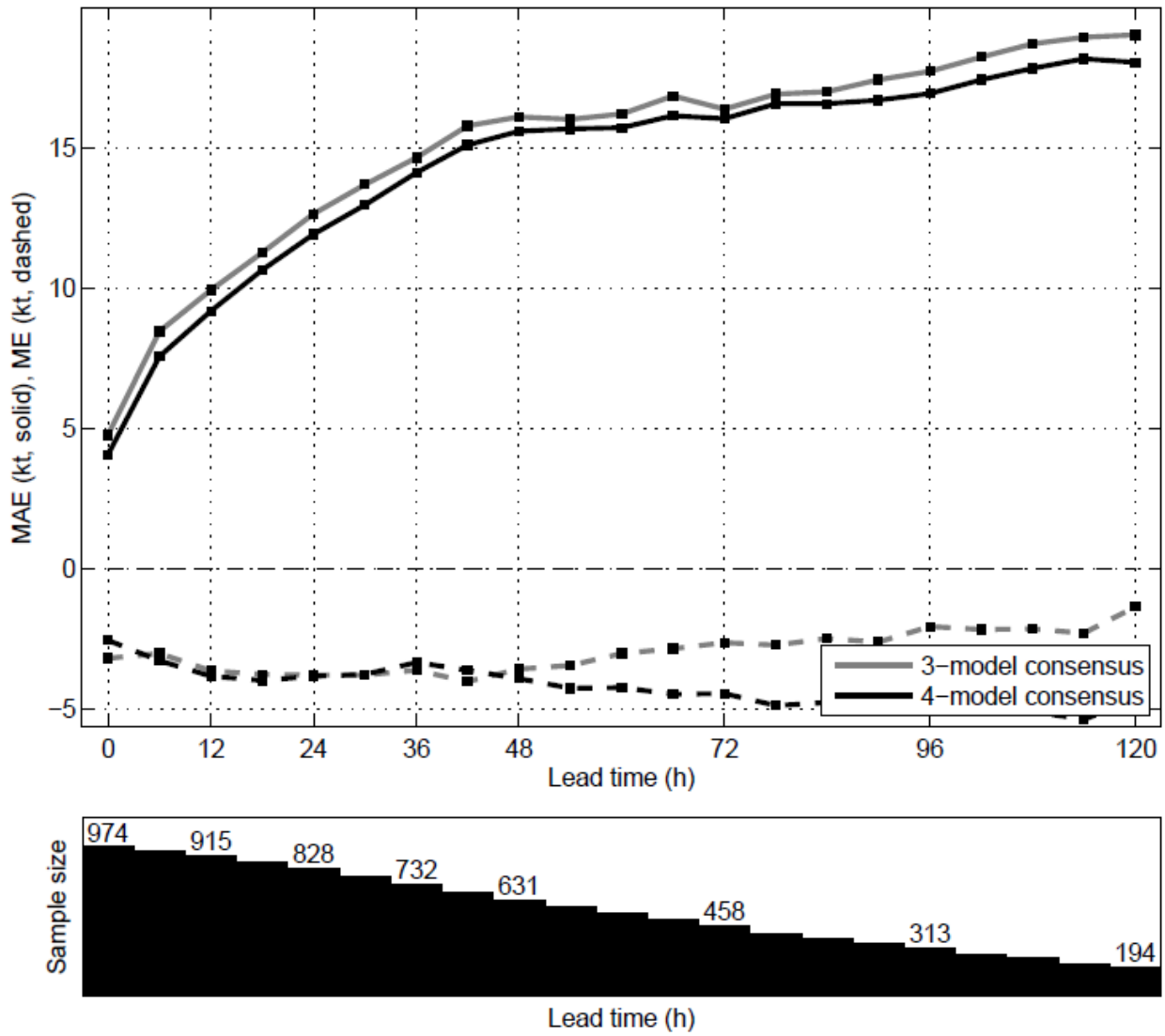

Figure 5: Intensity mean absolute error (MAE, solid) and intensity mean error (ME, dashed) as a function of lead time for 2012-2014 Western Pacific tropical cyclone forecasts from the 3-model COAMPS-TC/HWRF/GFDN consensus (gray), the 4-model COAMPS-TC/HWRF/GFDN/CHIPS consensus (black). The lower panel shows the sample size as a function of lead time for this homogeneous comparison.

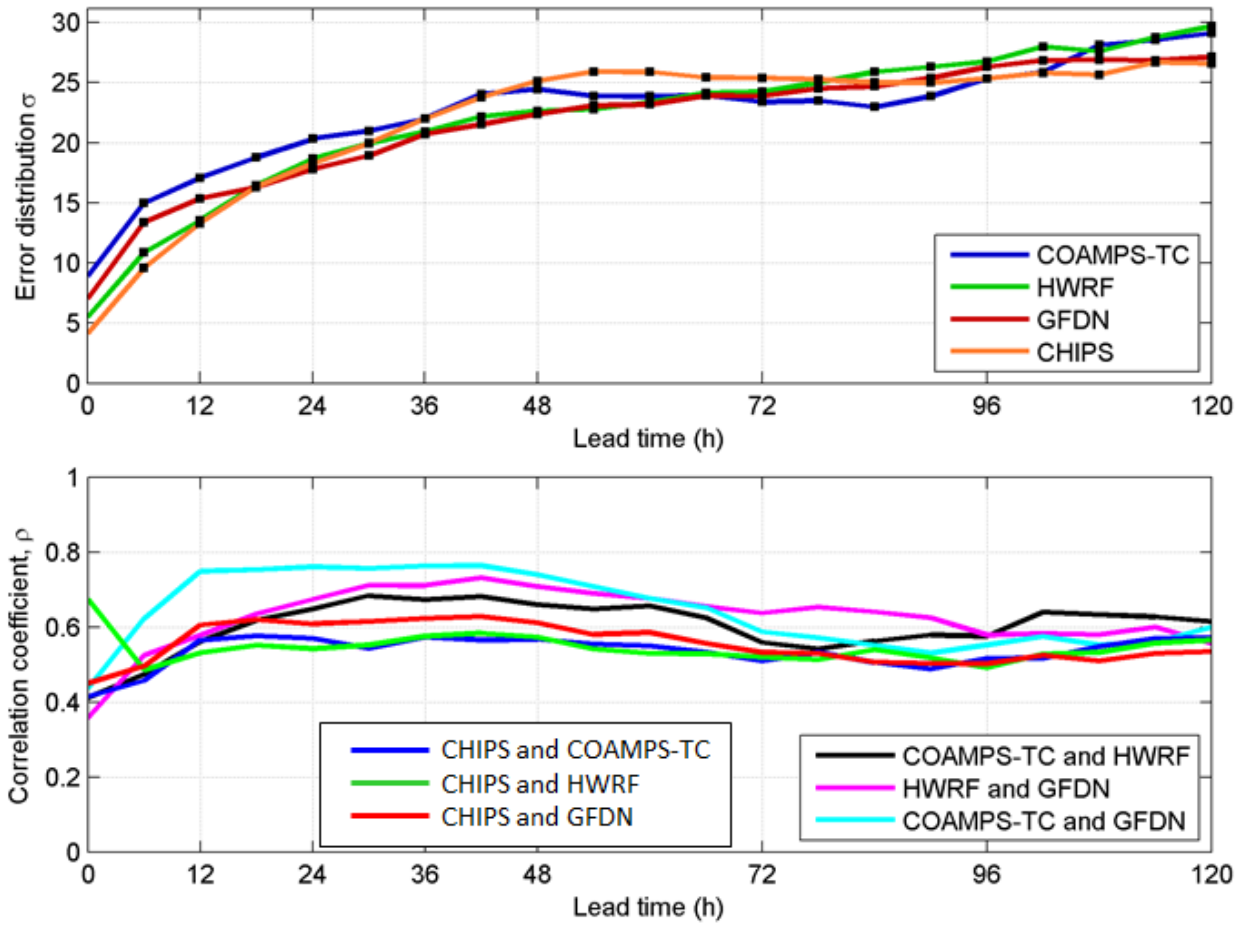

Figure 6: As in Figure 3, but for the 4-model COAMPS-TC/HWRF/GFDN/CHIPS consensus.

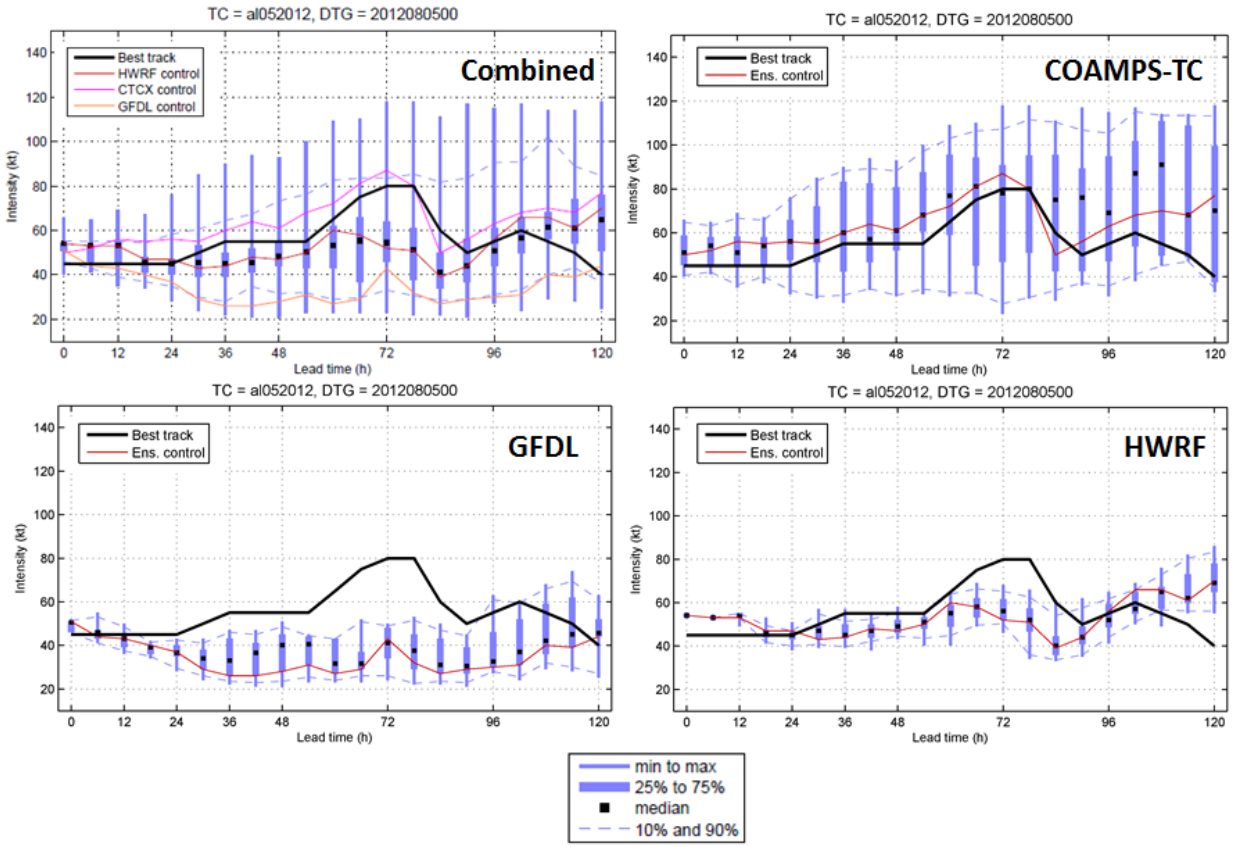

Figure 7: Boxplot diagrams showing ensemble intensity predictions for Ernesto (05L, 2012) for the 2012080500 initial time. The upper-left panel shows the forecast for the 42 member COAMPS-TC/HWRF/GFDL combined ensemble, upper-right the 11-member COAMPS-TC ensemble, lower-left the 10-member GFDL ensemble, lower-right the 21-member HWRF ensemble. The thick black line in each panel shows the observed best-track intensity for reference.

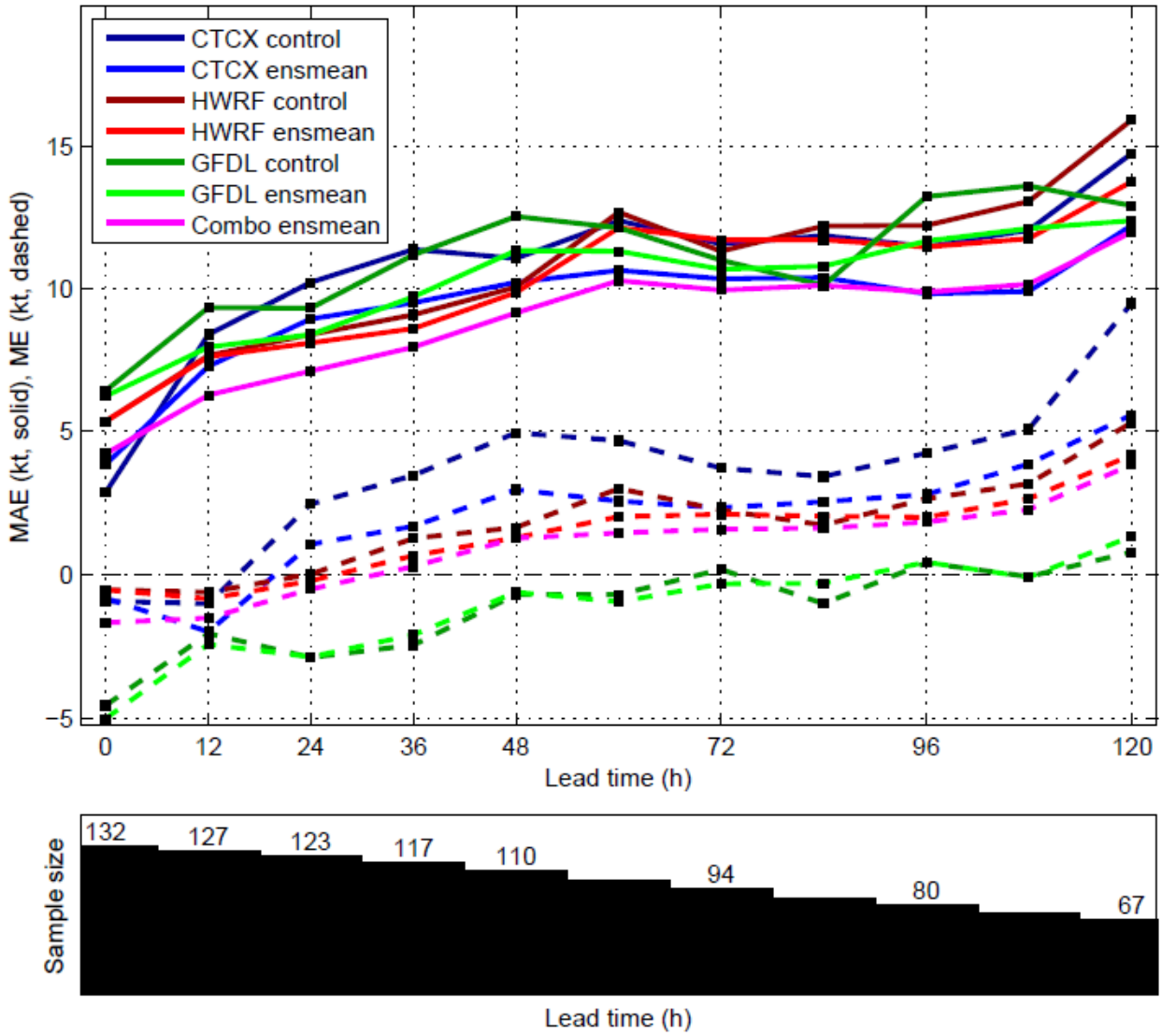

Figure 8: Intensity mean absolute error (MAE, solid) and intensity mean error (ME, dashed) as a function of lead time for a sample of retrospective Atlantic TC forecasts. COAMPS-TC ensemble control and ensemble mean results are shown in blue, HWRF ensemble control and ensemble mean results and shown in red, and GFDL ensemble control and ensemble mean results are shown in green. For each model, the darker color shade pertains to the control and the brighter shade pertains to the ensemble mean. Results for the COAMPS-TC/HWRF/GFDL combined ensemble mean are shown in magenta. The lower panel shows the sample size as a function of lead time for this homogeneous comparison.

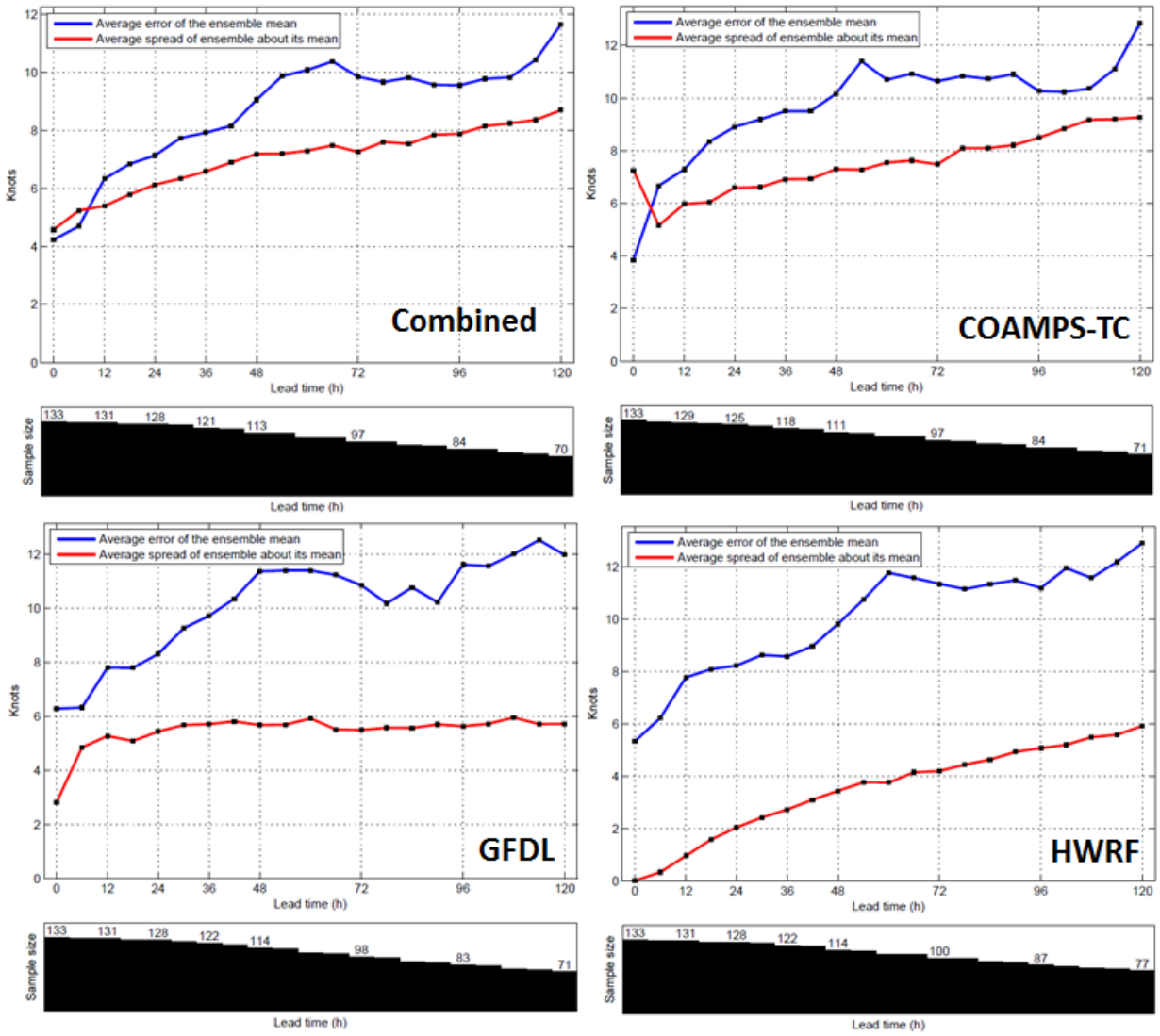

Figure 9: "Spread-skill" diagrams pertaining to the retrospective Atlantic basin intensity predictions for the COAMPS-TC/HWRF/GFDL combined ensemble (upper-left), COAMPS-TC ensemble (upper-right), GFDL ensemble (lower-left) and HWRF ensemble (lower-right). The average error of the ensemble mean ("skill") is the blue line, and the average spread of the ensemble about its mean ("spread") is the red line.

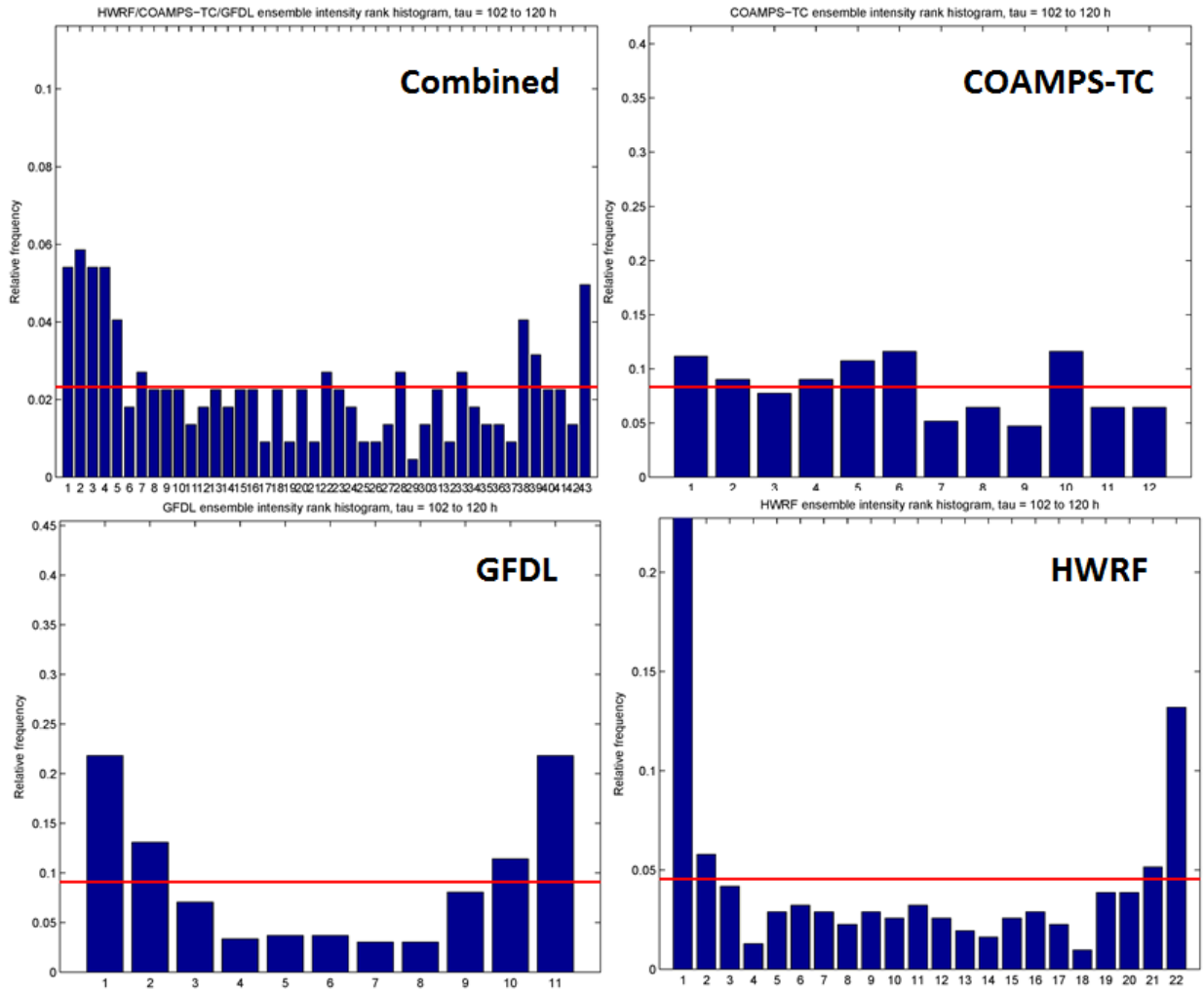

Figure 10: Rank-histogram diagrams pertaining to the retrospective Atlantic basin intensity predictions for the COAMPS-TC/HWRF/GFDL combined ensemble (upper-left), COAMPS-TC ensemble (upper-right), GFDL ensemble (lower-left) and HWRF ensemble (lower-right). Results from 102-120 h lead time are used here. The blue bars show the relative frequency at which the observed intensity falls between the ensemble members ranked from lowest intensity to highest intensity. Note that the left-most bar shows the relative frequency that the observed intensity is lower than that of all the ensemble members, and the right-most bar shows the relative frequency that the observed intensity is higher than that of all the ensemble members. Ideally, if the ensemble forecasts were perfectly reliable, the blue bars would form a flat histogram with relative frequency indicated by the red line.