A Tropical Cyclone Forecast Metric for Operations and Model Development

Michael Fiorino

Cooperative Institute for Research in Environmental Sciences, Univ. of Colorado Boulder and NOAA Global Systems Laboratory, Boulder CO

Submitted on 31 March 2021 as a Manuscript for Consideration in

Weather and Forecasting

Corresponding Author. Email: michael.fiorino@noaa.gov

ABSTRACT

Tropical Cyclone (TC) forecast metrics have two primary applications. For the human forecaster, they suggest which numerical models and aids should be considered in the forecast process, e.g., giving greater weight to a 'good' model. For the modeler, metrics define forecast quality and help diagnose model error.

Position Error (PE; often called 'track' error) and Intensity Error (IE) are the current standard TC forecast metrics but are only indirectly related to the actual TC forecast – the one-minute-average, 10 m (surface) wind field. For both the Joint Typhoon Warning Center (JTWC) and the National Hurricane Center (NHC), the official warnings that drive user response is the onset of 34 kt winds. Thus, PE/IE statistics do not directly measure error in the TC forecast of the surface wind, especially wind ≥ 34 kt.

A metric is proposed (Forecast Error or FE) that is more directly related to the actual TC wind field forecast based on Integrated Kinetic Energy (IKE); FE is defined to be a function of *both* PE and IE. We then demonstrate that the intensity forecast contributes at most 10-12% of total FE, i.e., a perfect intensity forecast can only improve (lower) FE at most by 10-12%. Track is still the most important part of the TC forecast and an18-y model intercomparison shows that the ECMWF global model makes the best PE forecasts in all basins and forecast times, despite having the highest initial PE and IE.

Through a breakdown of PE/FE by storm we find that the not all large PE cases have large FE and *vice versa*, i.e., the new metric can distinguish forecast errors with greater operational impact.

1. Introduction

In 2006, a wise US Navy Captain told me:

"You're only as good as what you measure"[1]

This pithy statement captures the essence of metrics – quantification of quality or putting numbers to 'goodness.' Among the many applications of metrics two stand out: 1) 'guidance on guidance' for the human forecaster – what are the 'good' models and how 'good' is the current 'good' model forecast?; and 2) diagnostics for model development. For mid-latitude weather forecasting, the 5-day 500 hPa northern Hemisphere (NHEM) Anomaly Correlation (5DNAC), is a *single* number found to be highly correlated with *general* numerical weather prediction (NWP) skill and has had a long and strong influence on global model development most notably in the early decades 1980-2000 (Benjamin et al. 2019 and Simmons and Hollingsworth (2002).[2]

Tropical cyclone (TC) forecast metrics have similar history. In the 1960s, the annual typhoon reports from the Joint Typhoon Warning Center (JTWC see JTWC 2020a) show a struggle to define the forecast metric and its summary statistic, but by the late 1970s two TC forecast metrics emerged that is now our current standard: position error (PE; great circle distance between forecast and verifying, post-season 'best track' (BT) position) and intensity error (IE; difference in maximum 1-min, 10-m wind speed between forecast and best track). The summary statistics are mean PE (mPE) and mean absolute IE (maIE) at set forecast times, typically 0, 12, 24, 36, 48, 72, 96 120 h.

[1] Victor Addison, Commanding Officer, Fleet Numerical Meteorology and Oceanography Center, Monterey CA

[2] In 1999 I developed a TC vitals assimilation scheme for ERA-40 while on secondment to ECMWF. The 5DNAC of runs with TCs was about 0.005 lower (worse) than without and the scheme was not used at ECMWF but was used at JMA for the JRA-25 reanalysis. At the time there was extreme sensitivity at ECMWF for any reduction in the 5DNAC of an model update/upgrade.

PE is often called 'track error' but this is both inaccurate in how PE is calculated (has nothing to do with a path or line) and is only indirectly related to track. For example, a TC forecast track can have a 0 n mi PE at 72 h but a very large PE at 24 or 120 h. This 0 n mi 72-h PE would define the 'track' as 'perfect' when in fact it is not. To measure track and speed error, PE is broken into parallel (along track) and orthogonal (cross track) components and while these components are better related to track error they are still point-in-time measures. IE is similarly point-in-time and a perfect intensity forecast at 72 h could be horribly wrong at 24 h and *vice versa*.

A basic problem with the standard PE/IE TC forecast metrics is how they are indirectly related to the actual TC forecast – the surface wind field defined at JTWC/NHC as the 1-min. averaged, 10-m (above the surface) wind. Moreover, the areal extent and time-of-onset of surface winds ≥ 34 kts (gale force) is the basis of NWS TC warnings/watch and the US Navy ship avoidance zone in JTWC warnings.

Given the weak connection between PE/IE and the TC wind forecast, the primary purpose of this paper is to propose a TC forecast error (FE) metric that is more directly related to the operationally critical TC forecast of surface winds ≥ 34 kts. The metric is defined as a function of both PE/IE and the BT mean radius of 34 kt winds (R34). When assuming a perfect intensity forecast (IE=0) the FE metric, on a seasonal basis, is reduced by at most 12%. Thus, approximately 90% of mean FE comes from PE, but not all large PE have large FE…

The record-setting 31-storm 2020 Atlantic (LANT) season is first analyzed to demonstrate how FE compares to PE/IE and how individual storms contribute to the seasonal means. We further show the how the FE v PE relationship depends on R34 and why large PE

90 can have small FE and vice versa. A challenge for model developers, using large-error cases for
91 testing, could be selecting the large-error metric – FE or PE?
92 We then broaden our examination to the three main NHEM basins: 1) the LANT; 2) the
93 central and Eastern north PACific (EPAC); and 3) the Western north PACific (WPAC) for the
94 14-y period 2007-2020. The same models used in the 2020 LANT are considered in the larger
95 intercomparison
96
97 iand the formulation; and additionally, how IE contributes at most 12% to FE. Second,
98 how mid-latitude forecast metrics that measure general NWP model skill are very different than
99 TC metrics and how PE is primary in measuring overall TC model forecast quality. Third, how
100 individual storms contribute to seasonal means and how large PE does not always imply large
101 FE.
102 The TC forecast is first defined in Section 2 and the nature of TC PE is compared to
103 standard Numerical Weather Prediction (NWP) metrics that are used by both forecasters ('what
104 are the good models?') and modeler developers ('is my model getting better?') in Section 3. We
105 then show that TC metrics cannot be used in the same way as NWP scores because they are
106 fundamentally different – an apples vs. oranges problem. More importantly, we will also see in
107 Section 3 that the standard TC metrics are indirectly related to the actual TC forecast of surface
108 wind. In section 4, the new FE metric is proposed based on error in Integrated Kinetic Energy
109 (IKE, (Powell and Reinhold 2007) and we demonstrate the impact of 'perfect' intensity forecasts
110 on FE. Application of TC metrics to model development is reviewed in Section 5 and how 'bad'
111 PE forecasts are often different than big FE errors on an individual storm basis. In Section 6 we

conclude with a discussion how FE gives a different view of TC forecasts and how the metric can be applied in both operations and modeling.

2. A TC Forecast is…

I have conducted informal surveys of both laymen (e.g., my wife) and professionals asking the question 'what is a TC forecast?' The answer almost always include the phrase 'where the storm is going' and certainly track and position are part of a TC forecast. In fact, both the National Hurricane Center (NHC) and JTWC forecast the surface wind field (10-m, 1-min average wind speed). The 2-D surface wind field is *parameterized* by position (latitude, longitude), intensity, maximum surface wind speed (Vmax) at a radius of maximum wind (Rmax), and the radii of gale force (34 kt, "R34"), storm force (50 kt, "R50") and hurricane force (64 kt, "R64") winds by quadrant. Furthermore, the NHC hurricane and tropical storm warnings/watches that call for action by emergency management end users typically depend on the onset of gale forecast winds (NHC). Similarly, JTWC warnings show the areal extent of gale force winds and defines, according to Commander Pacific (PACOM instruction ???), the 'no-sail' zone where US Navy ships are to avoid. Thus, the forecast of 34 kt (and greater) winds is perhaps the single most critical aspect of a TC forecast but is traditionally not verified directly.

Examples of NHC and JTWC TC forecasts are given in Fig. 1. Supertyphoon HAISEN of 2020 was the first supertyphoon (intensity >= 130 kt) of the western North Pacific (WPAC) season and had large impacts on the US military, as well as Japan and the Korean peninsula. The JTWC warning in Fig 1a shows the 'no-sail' zone (hatched) constructed from the R34 forecast plus the mean PE at each forecast time and the R34/R50/R64 radii. Note how the no-sail region considers both the extent of gale forecast winds and the JTWC mean PE. In Fig. 1b we see the NHC

equivalent for hurricane DELTA of 2020 which shows the track forecast and the forecast 'cone' the represents the region with the most probable track (NHC 2020b). The same NHC forecast in Fig. 1b is converted to the JTWC form by Fleet Weather Center Norfolk that shows the no-sail zone in Fig 1c comparable to Fig. 1a. Clearly, the largest contribution to FE comes from PE, especially at the longer lead times.

Suppose the forecaster makes a perfect R34 forecasts (i.e., gets the size of the radius of the 34 kt winds correct), but if the PE > 2*R34 then the forecast winds will be outside the observed extent of gale force winds and the forecast would be a (complete) 'bust.' In this case, IE can be anything and the forecast of the onset of 34 kt winds will be always wrong. This is the key concept behind the proposed FE.

Historically, FE was defined as *only* equal PE until the 1980s when IE was also assessed in the JTWC Annual Typhoon/Tropical Cyclone Reports dating from 1959-present. In the early era of TC forecasting (1960/70s) mean 24-h PE was order 150 n mi and the typical R34 in both WPAC and the Atlantic (LANT) basins was around 100 n mi (see Fig. 7). Thus, forecasts were essentially worthless when PE > 200 n mi. Of course, there is great variability in PE for individual storms and/or seasons, but this scaling argument shows the importance of PE (mathematically positive definite and unbounded) relative to R34 (physically bounded).

3. Nature of the TC PE forecast metric

In this section we compare a standard mid-latitude metric (anomaly correlation coefficient or ACC) to TC PE to show important differences that can confound interpretation of TC forecast quality. The essential problem is that the mean ACC has the same number of forecasts or cases at each lead time over the same area (e.g., northern Hemisphere) and at the

same pressure level (e.g., 500 hPa). Further, the ACC score comes from the model solution (a continuous process) whereas TC PE is calculated from a 'cyclone tracker' as a feature in the model solution. Further, the number of cases at each tau will be different with the possibility of two or more storms verifying at single time. From these considerations alone, statistics of mid-latitude ACC are very different than statistics of TC PE – a classic 'apples' v 'oranges' problem.

a. Mid-latitude die off curve

As introduced in Section 1, the 5DNAC (5-d NHEM 500 hPa ACC) has been used to measure progress in operational global NWP models over the last 40 y (Simmons and Hollingsworth 2002) and is a headline score still used to by NWP centers for monitoring (e.g., detecting 'dropouts') and for model intercomparison. Synoptic meteorologists have also found that when 5DNAC goes below 0.6, the model has little or no skill for deterministic weather forecasts (Hollingsworth et al. 1980).

Fig. 2 gives a recent 30-d time series of 5DNAC, and the mean value, for the 'big 5' global modeling systems. Figure 2 shows that for this period that the United Kingdom Meteorological Office (UKMO) had the highest ACC score (0.919) followed by the European Centre for Medium-Range Weather (ECMWF) (0.917) forecasts and the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS) (0.909). The mean anomaly correlation for lead times 0-192 h are then plotted to show how skill degrades in time. This so-called 'die off' curve is given in Fig. 3 where the 0.6 line is added to indicate the 'no-skill' forecast time. For this 30-d period ECMWF has the longest skill lead time ~ 8.5 d. We also see that the UKMO has a higher mean at day 5 (120 h from the time series in Fig.2) thus showing

how a single mean, while useful as a benchmark/standard, only partially represents model forecast quality and error growth.

b. TC mean PE curve

For TC PE we analyze the historic 2020 LANT season featuring a record number of storms (31), with many of them making landfall over the US (3 in Gulf of Mexico). Three models are considered: 1) HWRF (Hurricane Weather Research Forecast model run at NCEP using GFS initial and lateral boundary conditions see DTC 2018); 2) GFS (EMC 2020) and 3) the deterministic run of ECMWF IFS called HRES (ECMWF 2020). Over the last 15 or so years, ECMWF has been the 'best' model and is the gold standard for TC forecasting as will be discussed below. We also note that we are comparing two versions of the GFS – the global model itself and the GFS solution as dynamically downscaled by the HWRF limited-area model.

Fig. 4 gives two versions of the TC equivalent of the mid-latitude 'die off' curve; i.e., how the mean error grows as a function of forecast lead time. Fig. 4a shows the traditional line plot as would be found in the yearly NHC verification reports e.g., Cangialosi 2020. The forecast tracks for HWRF and the GFS come from NHC and the ECMWF tracks were calculated locally using near-native resolution grids and the GFDL tracker (Marchok 2021). The post-season reanalysis or 'final' best track is not yet available; we use instead the 'working' best track, as analyzed by the NHC Hurricane Specialists during the forecast process. While the final post-season, best track will differ from the operational working best track, there will be little change in the mean PE mostly because of the many 'fixes' (e.g., observational position/intensity estimates from aircraft, satellite, and radar) available in 2020 make the working best track more accurate than in the best tracks for previous years, especially for position.

204 The basic result is that all three models have similar error growth, and that ECMWF has
205 generally lower mean PE, but this standard plot (Fig 4a) is not as informative as Fig. 4b which
206 shows the distribution of PE and the median. Here, the median is similar in the short range of 0-
207 36 h, but at the longer lead times (72-120 h) HWRF has a higher median at 72 and 96 h, but at
208 120 h HWRF has the lowest median. A basic difference in the 24 and 120 h mean PE is that the
209 number of cases is much lower (~60) at 120 h than at 24 h (~260). Another difference is that the
210 comparison in Fig. 4 is heterogenous (all cases) vice homogeneous (only cases in common are
211 included) for two reasons. The model tracker will not always find a TC vortex at all verifying
212 times because of communication and/or tracker failures but more commonly because the model
213 dissipated the storm. The other reason is that the more models in a homogeneous comparison,
214 the lower the number of cases so that a homogeneous mean is conditioned on *all* the models
215 being available simultaneously. The heterogeneous comparison better represents the *individual*
216 model statistics. In contrast to even 10 y ago, the probability of detection (forecasting the
217 observed storm) is around 90% for all the models at 120 h, i.e., the models almost always make a
218 forecast.
219 For statistical significance tests, homogeneous statistics should be used, but for
220 forecasting purposes even small (~10%) differences in mean PE that may not be statistically
221 significant are observable by the forecaster (based on personal experience at both NHC and
222 JTWC). Further, the change in number of cases with forecast time makes assessing general
223 model quality difficult as discussed below.
224
225 *c. R34 distribution as a function of forecast time*

To show how cases in a 72-h forecast are different that at 24 or 120 h consider the histogram of R34 in Fig. 5. The histograms at 24, 72 and 120 h were constructed from best track data in the LANT 2010-2020 by finding the verifying R34 assuming a forecast is made from each time in the best track and following the standard verification rule that the storm must be classified as a TC initially and at the verifying time. Initially there will be 100% cases, but for short-lived storms (say 48 h) there would be no verifying R34 at 72 h, so the percentage of 'forecasts' always decreases. For the LANT basin and in the 10-y period, the percentage of verifying 24, 72, and 120 h is 85%, 51% and 31% respectively and while not shown, the number of storms at these forecast times will also decrease to the point that one or two TCs could dominate the seasonal mean PE at 120 h.

The fundamental problem is that the 120-h mean PE represents a different class of TCs than at 24 h. Thus, the PE error growth curves cannot be interpreted in the same way as ACC die off. From the histogram in Fig. 5 we see that the median R34 at tau0 is 85 n mi (not plotted), at 24 h 90 n mi, at 72 h 112 n mi and 120h 128 n mi. Not only are the number of cases and number of storms different, but the storms are structurally different.

d.. TC mean IE curve

Intensity (Vmax) forecasts have many issues, but the main one for a model is that Vmax is strongly dependent on spatial grid resolution. Lower resolution models will always have lower intensity forecasts (Walsh et al. 2007). A solution used in operations is to start the model intensity forecast from the same point through a bias correction procedure that applies an 'offset' or the initial model-observed intensity difference to subsequent forecasts. The full offset is applied at tau 0 and then is decreases linearly to 0 at a later tau. For global models, the 0 offset

tau is 72 h, for limited-area models is 24 h. The procedure used here is identical to that in NHC/JTWC operations.

The standard IE statistic is the *absolute* mean because unlike PE, IE can be both positive (too strong) or negative (too weak). The mean IE (mIE) is thus a bias and for most models (bias defined as model-observed) is negative. Fig. 6 gives the bias-corrected IE for the same cases as in Fig. 4 except that we plot both the standard absolute mean (amIE; lines) and the distribution as well (mIE; bars). All models have nearly 0 mean abs IE initially because of the bias correction but by 72 h the amIE has nearly the same magnitude as the bias (thick horizontal black lines in the mIE box-and-whiskers) for the global models, except for HWRF which is nearly unbiased. This is truly remarkable. Unlike the global models, the HWRF amIE does *not* come from bias.

We also show in Fig. 6 the ECMWF intensity forecast from their tracker (dark gold) and the one run locally using lower resolution grids (light gold). Somewhat surprisingly, the local tracker, using 0.25 deg grids v the ECMWF tracker using ~0.10 deg grids, has very similar bias although at 120 h the ECMWF tracker bias is somewhat less.

Another feature of the IE curves is that curves plateaus around 36-48 h unlike PE which always grows in time. There must be fundamental difference in this metric as discussed by (Kieu and Moon 2016) but in operational verification IE is treated in the same way as PE and is seemingly given equal weight when assessing overall forecast skill or FE.

e. PE relative to a baseline

Another way to compare model PE is to calculate the percent improvement relative to a baseline as in equation 1:

$$\%IMP_{PE} = \frac{(mPE_{baseline} - mPE_{model})}{mPE_{baseline}} \cdot 100 \quad (1)$$

If the model mPE is less than the baseline (lower PE) then $\%IMP_{PE}$ will be positive and bounded to 100%. Conversely, a negative $\%IMP_{PE}$ means the model has a higher mPE and is unbounded.

NHC uses the 'no-skill' CLImatology and PERsistence model CLIPER (Abserson 1998) as the baseline so that $\%IMP_{PE}$ represents 'skill.' In the 1970s few aids had 'skill' or were unable to make predictions with mPE lower than CLIPER. Another reasons for using a CLIPER baseline is in how the seasonal CLIPER mPE can represent the year-to-year variation in forecast difficulty. For example, a season with mostly 'straight runners' (i.e., TCs that follow an essentially linear path) will be more climatological (when averaging tracks in a 30-y period) and CLIPER mPE will be lower. The problem with using CLIPER circa 2020 is that current the official and model mPE is now almost 30% that of CLIPER whereas in the 1970s forecasters could rarely 'beat' CLIPER. Consequently, $\%IMP_{PE}$ in the annual NHC reports is approximately 50-60% for most models making inter-model comparison more difficult.

A better alternative is to use a global model as the baseline; to measure improvements relative to the global model that drives all forecasts, especially limited-area models such as HWRF. In Fig. 7 we show the $\%IMP_{PE}$ for HWRF and ECMWF relative to the GFS for the same basin/year as in Fig. 4 (the 2020 LANT season). The effect of vortex initialization in the HWRF is clearly seen in the much lower initial PE (HWRF mPE of 8 n mi v GFS mPE of15 n mi and a $\%IMP_{PE}$ > 50% from Fig. 4 table). However, ECMWF is better than the GFS (~10%) at *all* lead times which is remarkable given that ECMWF does no TC vortex initialization in operations (and never has). Further, a mPE of ~100 n mi at 72 h in the LANT is incredible when in the 1980s the mPE for any model was ~360 n mi!

To gain perspective on the operational significance of this massive reduction in mPE, consider the area covered by a radius of mPE in Fig. 8. In the 1980s, a 72-h forecast of a storm

295 in the middle of the Gulf of Mexico could make landfall almost anywhere because its mPE was
296 360 n mi, whereas in 2020 the forecast would be much more accurate due to the reduction in
297 mPE to 100 n mi. In terms of area, this is a 92% reduction in the threat zone and must be
298 considered a singular and perhaps one of the greatest NWP success stories of the last several
299 decades.

300 To further put perspective on the 2020 mPE we plot the $\%IMP_{PE}$ from 2007-2020 and
301 for the main northern Hemisphere basins in Fig. 9a. ECMWF had better mPE than the GFS at all
302 forecast times not only in the LANT, but also in the other basins. Furthermore, ECMWF was
303 better than the GFS in all three basins and at all 5 lead times (i.e., 'ran the board') in 2010, 2011,
304 2015, 2017, 2018, and 2020.

305 Other notable features:

306 • Since 2012, HWRF rarely outperformed the global model providing its lateral
307 boundary conditions (GFS) and in 2020 only at tau 96 and 120 in the WPAC and
308 LANT. This is noteworthy as limited-area models are not expected to provide
309 long-term guidance when the TC forecast track is dominated by the lateral
310 boundary conditions, and thus suggests that the skill of the GFS is actually
311 decreasing at these longer forecast hours.
312 • The big change from 2007 to 2008 in the ECMWF mPE can be attributed to
313 change in physics as discussed in Fiorino 2009.
314 • The GFS improvement (ECMWF less green) from 2011 to 2012 can be mostly
315 attributed to improved data assimilation used in the GFS (hybrid 3DVAR-EnKF).
316 • The exception to ECMWF providing *the best short-term guidance* (tau 12 & 24)
317 was in the WPAC in 2019.

Applying the same analysis to the latest ECMWF reanalysis ERA5 (Hersbach et al. 2020) and the operational run of the model (ECMWF HRES) is show in Fig. 9b. Note the vertical bar to the right of the box plot and horizontal bar below. These give the mean for the 14-y period (vertical) and the yearly mean (horizontal) where we can clearly see how ERA5 forecasts are better than from the real-time run. The ERA5 model and data assimilation are representative of the operational model circa 2016 with a resolution of ~ 31 km and more notably is a coupled reanalysis using the latest 4DVAR data assimilation with a window that looks forward 12 h (not possible in operations). The 2007 problem (Fiorino 2009) is gone in ERA5 and all years and tau-basins are greener, especially before 2016 but even into 2020.

4. TC forecast error metric

We have seen that error growth curves of TC mean PE, while visually similar to AC die off plots, represent very different measure of model forecast error because TC PE is for a unique and discrete event (the TC) in the model solution whereas AC is calculated from the continuous model solution itself. Also, the number of events in mean PE not only decreases with increasing forecast time but come from a different class of storm. But perhaps the larger problem with using PE and IE to measure forecast quality is the indirect relationship to the 'actionable' part of the TC forecast – the onset and areal extent of 34+ kt winds that can range from near 0 n mi for weak tropical storms or greater than 200 n mi for large (and generally intense) hurricanes.

To overcome these deficiencies we propose a metric FE that depends on R34 and is a function of *both* PE and IE. First, we only require the model forecast position and intensity and use the best track R34 as the forecast R34 (a perfect R34 forecast). We also use the best track radius of maximum wind (Rmax) for the model forecast (a prefect Rmax forecast) when

available. When the best track Rmax is not available a statistical relationship between Rmax and Vmax is used (Quiring et al. 2011). To restate, by using the verifying R34 (and Rmax) we only require the model forecast position and intensity and the proposed FE metric is a lower bound.

We then construct a symmetric wind profile used by Fiorino and Elsberry 1989 and shown in Eq. 2 below that only requires two (v, r) points to analytically specify the winds from r=0, R34:

$$v(r) = V_{max}\left(\frac{r}{R_{max}}\right)\left\{\exp\left(\frac{1}{b}\left(1 - \left(\frac{r}{R_{max}}\right)^{b}\right)\right\} \quad (2)$$

For the FE metric we use the ordered pairs (V_{max} , R_{max}) and the (34 kt, R34) to solve for b. Given the wind profile, we integrate the square of $v(r)$ from r=0, R34 and to calculate the Integrated Kinetic Energy (IKE) (Powell and Reinhold 2007) for the model and best track. Whereas IKE in Powell and Reinhold extends past R34, we only consider the kinetic energy of the significant winds (>= 34 kt).

FE is the defined as the error IKE which is the IKE outside the overlap or 'symmetric difference' as shown in Fig. 10. Python is used for the integration of the analytic wind profile and to find the area outside the overlap. When PE = 0, FE = 0 even if IE=0 because there the two R34 circles perfectly overlap – assuming a perfect R34 forecast. When PE = 2R34, FE is maximized and there is no overlap. For PE > 2R34 we assume FE grows linearly from the 2R34 FE with increasing PE. The FE defined here is a lower bound for the best case of perfect R34 and Rmax forecasts.

The FE for the same models/season are shown in Fig. 11. Fortuitously, the magnitude of FE in TJ is the same order as PE in n mi; simplifies intercomparison.

5. Application to Model Development

Recently, advances in TC forecasts have largely been made via improvements to the data assimilation (DA) methods and physics parameterizations; this section describes how TC metrics can be used to aid in these developments. There are many ways to measure DA aspects and it is naturally assumed that TC vortex initialization would be critical for good TC forecasts. Perhaps the easiest way to measure the quality of TC vortex analysis and initialization is to quantify the initial mPE and mIE. In Fig 13a. we find that over the 14-y period 2007-2020, HWRF mPE is 10 n mi, GFS is 14 n mi, ECMWF (HRES) and ERA5 are 19 and 18 n mi, respectively. For initial mIE (Fig 13b), HWRF has a very slight negative bias of -1 kt, GFS is -14 kt, ECMWF HRES is -18 kt and ERA5 is -20 kt.

Clearly, HWRF has a superior vortex initialization – the lowest initial PE and a nearly unbiased initial intensity. However, even 12 h into the forecast both ECMWF(HRES) and ERA5 have lower PE than HWRF/GFS (Fig. 9b). If forecasting is the 'acid test' of an analysis, then ECMWF must have the better vortex initialization, but they *do not assimilate* the TC 'vitals' or TC position, motion and structure as analyzed in the real-time by the human forecasters (Trahan and Sparling 2012). HWRF/GFS *do assimilate* TC vitals, although 'vortex relocation' was recently turned off in the GFS (EMC 2020).

The inescapable conclusion is that for TC track forecasting model physics and DA are critical, especially in the global models. Further, it is not obvious how TC metrics can be used in model development, except possibly for identification of systematic physics errors for 'large' TC forecast errors.

a. 72-h PE and FE

In Section 2 we discussed how the day 5 NHEM AC is a standard metric in global NWP. For TCs, it is argued that a best single number metric is a day 3 or 72 h on two grounds. First, the number of verifiable forecasts at 72 h is greater the 50% (i.e., 61% in the LANT and 67% in the WPAC as shown in Fig. 5) and that by 72 h the forecast is less constrained by the initial conditions and more determined by the model physics. The second reason is human forecast process itself.

Typhoon Duty Officer (TDO equivalent of NHC hurricane specialist) training at JTWC emphasizes that TC track forecast is a 'connect-the-dots' problem. Most of the time in a warning cycle is devoted to analyzing TC location and structure – a process that involves the qualitative assimilation of many observational and disparate data sources from ship reports to satellite imagery that are often not assimilated in the models. Given a solid initial position (and motion), the deterministic forecast problem is where will the storm be in 72 h (if we forecast the TC to exist for 3 or more days)? Once that is set, typically constrained by the 72-h consensus forecast and the previous warning, the issue is how to connect the initial and 72 h points and then to extrapolate from 72 h to 120 h (day 5) in a way the makes both physical and synoptic sense.

As discussed in Section 4, when the PE is $\geq$ 2*R34, IE only changes the magnitude of FE not the location component, i.e., the forecast high-wind area is outside the observed high-wind area. This property of FE is a consequence of the perfect R34 forecast assumption and therefore PE $\geq$ 2*R34 represent a bound on usefulness in terms of wind $\geq$ 34 kt. If a typical R34 is 100 n mi and PE is 100 n mi and if the PE can be either left or right of the TC center then a PE 100 n mi would be 200 n mi and equal to 2*R34. The forecast time when PE is around 100 n mi is around 72 h as seen in Fig. 4a. For FE, note that the error growth in mean FE is linear between 0 and somewhere between 48 and 72 h in Fig. 11a. The similarity between the PE and FE

410 growth curves is not surprising given that the contribution of IE to FE is around 10% as shown in
411 Fig. 12. Thus, the 72 h lead time can be considered the 'linchpin' of the track forecast.
412
413 *b. large 72-h errors – PE v FE*
414 We identify problem storms/cases by plotting the tau72 mPE and mFE for each
415 individual storm in Fig. 14 where mFE by-storm is normalized by the ratio of LANT 2020
416 season 72-h mPE/mFE (1.5) so that the units and scale of FE is approximately the same. First
417 note that the number of TCs contributing to the 72-h mPE/mFE is 21 or a rate of 68% for the
418 total of 31 storms in the 2020 LANT season. This rate is substantially higher than the 10-y rate
419 of 51% (Fig. 5) indicating an active season. However, the number of cases per storm also varies
420 so that high PE/FE may make only a small contribution to the total (season in this paper) mean.
421 In Fig. 15, we first convert the PE/FE to $\%IMP_{PE}$ or contribution to the total mean following Eq.
422 1 and then normalize by the (number of cases per storm / total number of cases). This variable
423 more explicitly shows how each storm contributes to the seasonal mean.
424 Comparing Fig. 15a (PE) to Fig. 15b (FE) we see a different distribution of problem
425 storms. For example, the biggest problem storm (longest bar below the 0 line) for the GFS was
426 **29L**.2020 (the 29th TC in the at**L**antic in 2020 or hurricane **ETA**). For FE, **29L**.2020 was
427 slightly below the 0 line, but TCs **17L**.2020 (hurricane **PAULETTE**) and **20L**.2020 (hurricane
428 **TEDDY**) were the larger problem storms. Table 1 summarizes these problem storms for PE v
429 FE and gives URLs for track plots that highlight the 72-h position. An example from the track
430 plot site is given in Fig. 16 for **14L**.2020 (hurricane **MARCO**) at 2020082112 (12 UTC 21 Aug
431 2020) when the GFS had the largest 72-h PE error as the storm entered the Gulf of Mexico.

The storms with large 72-h FE in Table 1 are cases when the verifying TC was both large (R34 > 100 n mi) and intense (Vmax > 65 kt). Table 2 shows how IKE varies as a function of size (R34) and intensity. For a moderate hurricane (90 kt) the IKE for a large storm is over 10X that of a small storm with an R34 of 50 n mi. Thus, FE will be large even for 'good' PE at 72 h (~seasonal mPE) and furthermore, these storms would be more significant operationally because the high IKE means greater storm surge and wind impacts.

To reiterate, examining error by TC using both the traditional PE and the new FE metric may help to queue the modeler to find possible model problems and, perhaps more importantly, identify TC forecasts errors that have the greatest operational significance.

6. Summary and Conclusions

At both JTWC and NHC, the extent and onset of 34 kt winds drives the warning process and defines 'actionable' intelligence for impacted users. Thus, a TC forecast is a *field* of 1-min (10-min average is the WMO standard) 10-m or *surface wind* and the new metric of Forecast Error (FE) is based on Integrated Kinetic Energy (IKE; Powell and Reinhold 2007) within the TC circulation of wind ≥ 34 kt.

The new metric for TC forecast quality is a function of *both* traditional metrics of PE and IE. Current operational verification treats PE and IE separately (e.g., Cangialosi 2020) even though TC intensity is physically related to track. Furthermore, science-based forecasting demands a holistic consideration of TC structure, not just where the storm is going. Another point of departure is how we compare models against a *model* baseline vice a no-skill CLIPER-type baseline and how we breakdown the seasonal means by TC to identify how each individual and unique storm contributes to the overall mean.

455 We considered two global models for the historic 2020 LANT hurricane season:

456 • NCEP GFS

457 o the operational version of the GFS itself

458 o the GFS dynamically downscaled by the limited-area HWRF model

459 • ECMWF global model

460 o the operational version that makes the 10-d deterministic run (called HRES)

461 o the circa 2016 version of HRES model cycle used in the latest ECMWF

462 reanalysis ERA5.

463 The basic result is that all the models did well in 2020 regarding mean PE (e.g., relative
464 the 2015-2019 mean), but when using the GFS as a baseline, we found that ECMWF had
465 superior PE, expressed as a percentage improvement, at all forecast times except the model
466 initialization time (0 h). We analyzed a 14-y period (2007-2020), including the main NHEM
467 basins, and found that HWRF rarely improves the host model GFS TC track forecast. ECMWF
468 on the other hand was generally superior in its forecasted track relative to both GFS and HWRF.
469 We then compared the ECMWF HRES forecasts to ERA5 reanalysis forecasts, again using the
470 operational GFS as the baseline, and found substantially better and more consistent year-to-year
471 improvements in the ERA5 over the real-time HRES model. The conclusion is that the model
472 very much matters for TC track prediction and more so than low initial position and intensity
473 error (vortex initialization aspects).

474 The new metric of FE was defined as the IKE outside the union of two circles spanned by
475 the forecasted R34 and the observed R34 but offset by the PE as illustrated in Fig. 10. We
476 assumed the forecast R34 was the same as observed to make FE only a function of PE and IE.
477 The mFE error growth curves are similar to mPE growth, except mFE grows faster when the PE

478 is greater than 2R34 after tau48. However, the more important finding is that if the model makes
479 a perfect intensity forecast, the forecast error is only reduced 10-12% and mostly for the early
480 forecasts (0-48 h). In other words, track is still 90% of the TC forecast problem.
481 Application of the standard PE and new FE metric to model diagnosis and verification
482 was then considered. If the quality of the TC vortex analysis is measured by initial PE and initial
483 IE, the HWRF forecast would be considered the best. However, ECMWF had the highest initial
484 PE/IE but the best 12- and 24-h PE which implies that the HWRF vortex initialization degrades
485 the short-term forecasts and hence the initial motion and development of the TC. What is
486 particularly impressive is that the mPE of the ECMWF forecasts at 12 h is about 25 n mi, which
487 is close to observational uncertainty in TC location itself (Landsea and Franklin 2013)!
488 We then considered how each storm contributed to the seasonal mean at 72 h – a critical
489 forecast time for storms predicted to last more the 3 d. In 2020, 68% out of 31 TCs in the LANT
490 had one or more verifying positions at 72 h. This is greater than the previous 10-y mean of 51%
491 as shown in Fig. 5. The most important conclusion regarding FE is that large (R34 > 120 n mi)
492 and intense (Vmax > 90 kt) storms will have much higher IKE error (FE) for a PE that is close to
493 the mPE. All storms are not equal in how they affect the seasonal mean error.
494 Finally, the FE analysis presented here gives a different view of TC forecast error.
495 Application to other basins and years should help identify the big and operationally significant
496 TC forecast error to guide model and forecaster development.
497

Acknowledgements.

I first want to acknowledge and thank Vic Addison, CAPT USN (ret), for his very thought-provoking comment during Captain's call while I was on active duty at FNMOC prior to joining NHC. There are few days when I do not reflect on CAPT Addison's "you're only are as good as what you measure" statement. The late Charlie Neumann, the father of American digital TC data, suggested to me in 1975 that 'forecast error = position error' and that intensity error does not contribute to forecast error. Finally, my long association with JTWC as their models officer and as a qualified Typhoon Duty Officer greatly influenced and contributed to my thinking on TC forecasts and metrics, as did my 3 years as the NHC lead techniques development (Charlie Neuman's billet). The experience of forecasting at both JTWC and NHC has been personally invaluable. The ERA5 forecasts were kindly provided by Hans Hersbach, the ECMWF reanalysis project lead.

Data Availability Statement.

All TC forecast and best track data available upon request.

512 REFERENCES

513 Aberson, S. D., 1998: Five-Day Tropical Cyclone Track Forecasts in the North Atlantic Basin.
514 *Weather Forecast*, **13,** 1005-1015.
515
516 Benjamin, S. G., Brown, J. M., Brunet, G., Lynch, P., Saito, K., & Schlatter, T. W., 2019: 100
517 Years of Progress in Forecasting and NWP Applications, *Meteorological Monographs*, *59*, 13.1-
518 13.67. Retrieved Jan 16, 2021, from
519 https://journals.ametsoc.org/view/journals/amsm/59/1/amsmonographs-d-18-0020.1.xml
520
521 Cangialosi, J.P. 2020: National Hurricane Center Forecast Verification Report. 2019 Hurricane
522 Season. NHC: https://www.nhc.noaa.gov/verification/pdfs/Verification_2019.pdf
523
524 Development Testbed Center, 2018: Hurricane WRF (HWRF) Documentation. Available online
525 at: https://www.dtcenter.org/community-code/hurricane-wrf-hwrf/documentation
526
527 European Centre for Medium-range Weather Forecasts, 2020: IFS documentation. Available
528 online at: https://www.ecmwf.int/en/publications/ifs-documentation.
529
530 Fiorino, M., 2009: Record-setting performance of the ECMWF IFS in medium-range tropical
531 cyclone track prediction. ECMWF Newsletter No. 118, 20-27
532 http://www.ecmwf.int/sites/default/files/elibrary/2008/14607-newsletter-no118-winter-
533 200809.pdf
534

Fiorino, M., and R. L. Elsberry, 1989: Some Aspects of Vortex Structure Related to Tropical Cyclone Motion. *J Atmos Sci*, **46,** 975-990.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Q J Roy Meteor Soc*, **146,** 1999-2049.

Hollingsworth, A., K. Arpe, M. Tiedtke, M. Capaldo, and H. Savijarvi, 1980: The Performance of a Medium-Range Forecast Model in Winter-Impact of Physical Parameterizations. *Mon Weather Rev*, **108,** 1736-1773.

JTWC, 2020a: Annual Typhoon/Tropical Cyclone Reports: https://www.metoc.navy.mil/jtwc/jtwc.html?cyclone

JTWC, 2020b: Indo-Pacific Command Instruction governing JTWC operations: https://www.usno.navy.mil/NOOC/nmfc-ph/RSS/jtwc/pubref/3140.html

Landsea, C. W., and J. L. Franklin, 2013: Atlantic Hurricane Database Uncertainty and Presentation of a New Database Format. *Mon Weather Rev*, **141,** 3576-3592.

Liu, Q., T. Marchok, H.-L. Pan, M. Bender, and S. J. Lord, 2000: Improvements in hurricane initialization and forecasting at NCEP with global and regional (GFDL) models. NWS Tech. Procedures Bull. 472, 7 pp. [Available online at: https://www.researchgate.net/publication/237459506_Improvements_in_Hurricane_Initialization_and_Forecasting_at_NCEP_With_Global_and_Regional_GFDL_Models]

Marchok, T.P., 2021: Important factors in the tracking of tropical cyclones in operational models. *J Appl. Meteo Climo*., submitted for publication.

National Hurricane Center, 2020a: National Hurricane Operations Plan [available on online at: https://www.icams-portal.gov/publications/nhop/nhop2.htm

National Hurricane Center, 2020b: NHC Forecast Cone[available online at: https://www.nhc.noaa.gov/aboutcone.shtml

Environmental Modeling Center, 2020: Global Spectral Model Reference Page available online at: https://www.emc.ncep.noaa.gov/emc/pages/numerical_forecast_systems/gfs/documentation.php

Powell, M. D., and T. A. Reinhold, 2007: Tropical cyclone destructive potential by integrated kinetic energy. *B Am Meteorol Soc*, **88,** 513-+.

Simmons, A. J., and A. Hollingsworth, 2002: Some aspects of the improvement in skill of numerical weather prediction. *Q J Roy Meteor Soc*, **128,** 647-677.

Trahan, S., and L. Sparling, 2012: An analysis of NCEP tropical cyclone vitals and potential effects on forecasting models. *Wea. Fore*., **27**, 744-756, https://doi.org/10.1175/WAF-D-11-00063.1.

Quiring, S., A. Schumacher, C. Labosier, and L. Y. Zhu, 2011: Variations in mean annual tropical cyclone size in the Atlantic. *J Geophys Res-Atmos*, **116**.

Walsh, K. J. E., M. Fiorino, C. W. Landsea, and K. L. McInnes, 2007: Objectively determined resolution-dependent threshold criteria for the detection of tropical cyclones in climate models and reanalyses. *J Climate*, **20,** 2307-2314.

592 TABLES AND FIGURES

593 Table 1. Problem storms for the GFS as defined by 72-h PE and FE for 2020 LANT season

TC	Error Type	Comments	URL for track plots
14L (MARCO)	PE	GFS too far West and South as storm entered the Gulf of Mexico	http://tctrkveri.wxmap2.com/trk-14L-2020-MOD-CUR.htm
17L (PAULETTE)	FE	Large R34 increases makes FE larger than PE	http://tctrkveri.wxmap2.com/trk-17L-2020-MOD-CUR.htm
20L (TEDDY)	FE	Large R34 as with 17L, particularly before extratropical transition	http://tctrkveri.wxmap2.com/trk-20L-2020-MOD-CUR.htm
29L (ETA)	PE FE	Two cases of PE > 700 n mi	http://tctrkveri.wxmap2.com/trk-29L-2020-MOD-CUR.htm
31L (IOTA)	FE	R34 > 100 n mi	http://tctrkveri.wxmap2.com/trk-31L-2020-MOD-CUR.htm

594

595

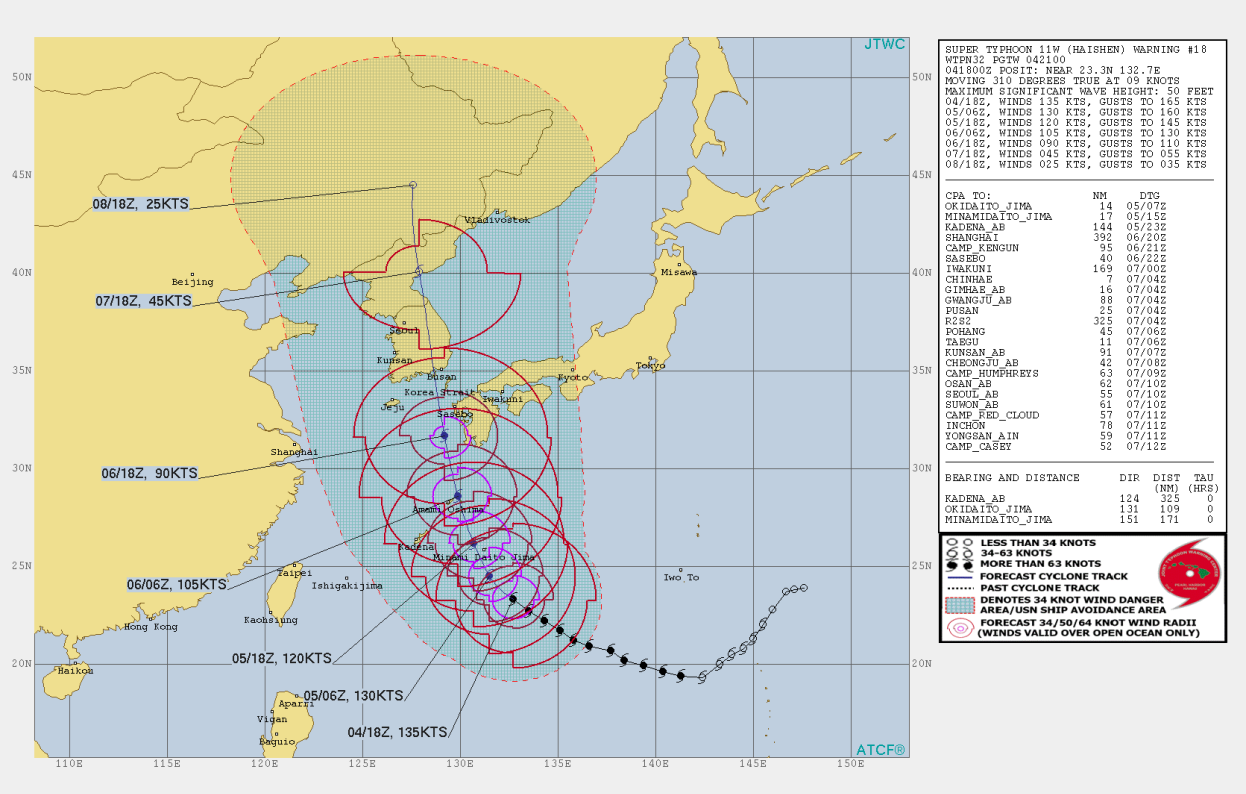
596 Table 2. IKE (TJ) for various Vmax and R34 using the Quirling et al. (2011) Vmax minus Rmax
597 relationship and the wind profile from Fiorino and Elsberry (1989)
598

Vmax (kt)	R34=50 n mi	R34 =100 n mi	R34=150 n mi
65	5	35	76
90	8	56	108
130	15	79	148

599
600

601
602

603 **1A**



SUPER TYPHOON 11W (HAISHEN) WARNING #18
WTPN32 PGTW 042100
041800Z POSIT: NEAR 23.3N 132.7E
MOVING 310 DEGREES TRUE AT 09 KNOTS
MAXIMUM SIGNIFICANT WAVE HEIGHT: 50 FEET
04/18Z, WINDS 135 KTS, GUSTS TO 165 KTS
05/06Z, WINDS 130 KTS, GUSTS TO 160 KTS
05/18Z, WINDS 120 KTS, GUSTS TO 145 KTS
06/06Z, WINDS 105 KTS, GUSTS TO 130 KTS
06/18Z, WINDS 090 KTS, GUSTS TO 110 KTS
07/18Z, WINDS 045 KTS, GUSTS TO 055 KTS
08/18Z, WINDS 025 KTS, GUSTS TO 035 KTS

CPA TO:	NM	DTG
OKIDAITO_JIMA	14	05/07Z
MINAMIDAITO_JIMA	17	05/15Z
KADENA_AB	144	05/23Z
SHANGHAI	392	06/20Z
CAMP_KENGUN	95	06/21Z
SASEBO	40	06/22Z
IWAKUNI	169	07/00Z
CHINHAE	7	07/04Z
GIMHAE_AB	16	07/04Z
GWANGJU_AB	88	07/04Z
PUSAN	25	07/04Z
R282	325	07/04Z
POHANG	45	07/06Z
TAEGU	11	07/06Z
KUNSAN_AB	91	07/07Z
CHEONGJU_AB	42	07/08Z
CAMP_HUMPHREYS	63	07/09Z
OSAN_AB	62	07/10Z
SEOUL_AB	55	07/10Z
SUWON_AB	61	07/10Z
CAMP_RED_CLOUD	57	07/11Z
INCHON	78	07/11Z
YONGSAN_AIN	59	07/11Z
CAMP_CASEY	52	07/12Z

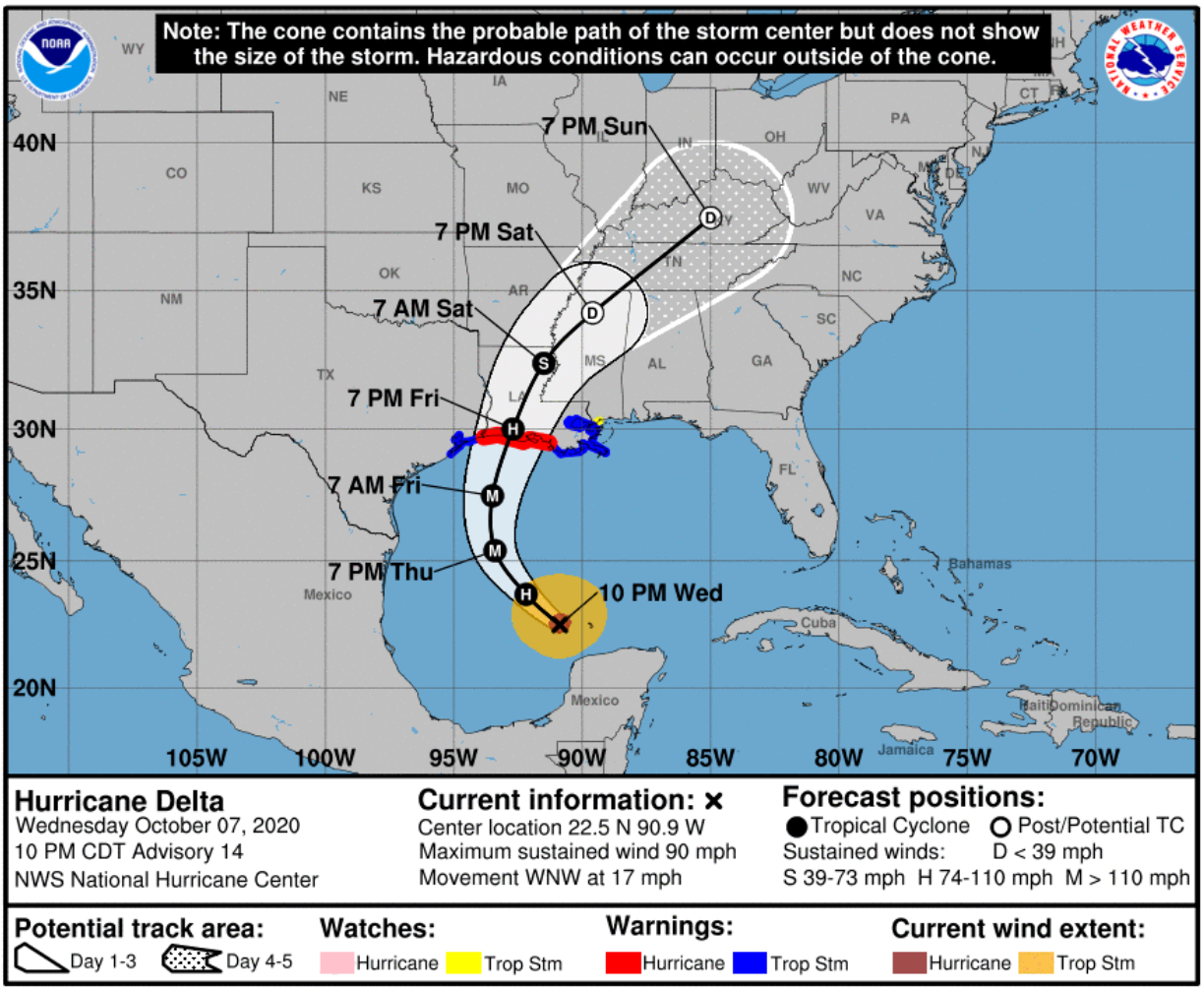
BEARING AND DISTANCE	DIR	DIST (NM)	TAU (HRS)
KADENA_AB	124	325	0
OKIDAITO_JIMA	131	109	0
MINAMIDAITO_JIMA	151	171	0

LESS THAN 34 KNOTS
34-63 KNOTS
MORE THAN 63 KNOTS
FORECAST CYCLONE TRACK
PAST CYCLONE TRACK
DENOTES 34 KNOT WIND DANGER AREA/USN SHIP AVOIDANCE AREA
FORECAST 34/50/64 KNOT WIND RADII (WINDS VALID OVER OPEN OCEAN ONLY)

604

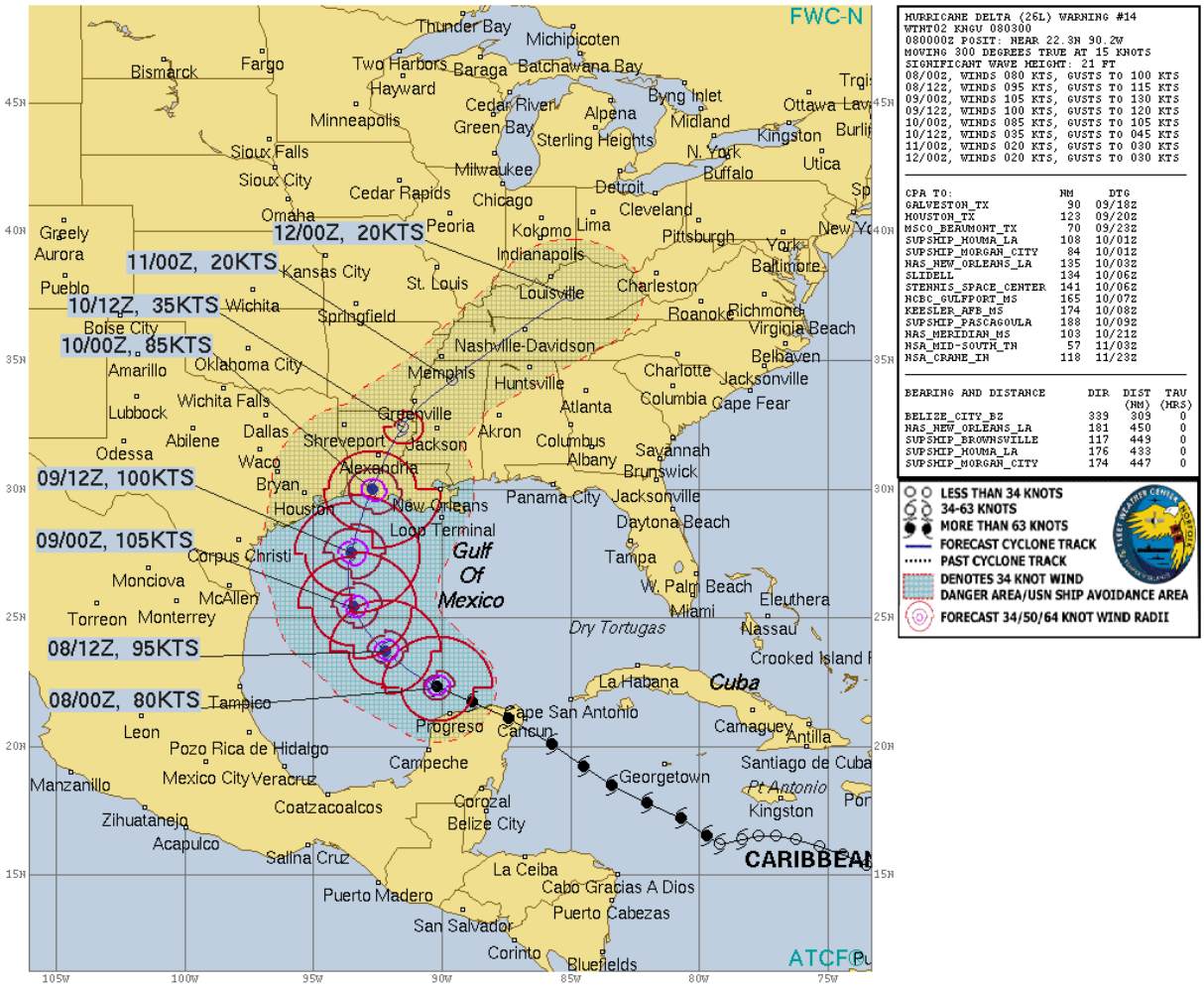
605

606 **1B**


Note: The cone contains the probable path of the storm center but does not show the size of the storm. Hazardous conditions can occur outside of the cone.
7 PM Sun
7 PM Sat
7 AM Sat
7 PM Fri
7 AM Fri
7 PM Thu
10 PM Wed
40N
35N
30N
25N
20N
105W
100W
95W
90W
85W
80W
75W
70W
Bahamas
Cuba
Jamaica
Haiti
Dominican Republic
Mexico
Hurricane Delta
Wednesday October 07, 2020
10 PM CDT Advisory 14
NWS National Hurricane Center
Current information:
Center location 22.5 N 90.9 W
Maximum sustained wind 90 mph
Movement WNW at 17 mph
Forecast positions:
Tropical Cyclone
Post/Potential TC
Sustained winds:
D < 39 mph
S 39-73 mph
H 74-110 mph
M > 110 mph
Potential track area:
Day 1-3
Day 4-5
Watches:
Hurricane
Trop Stm
Warnings:
Hurricane
Trop Stm
Current wind extent:
Hurricane
Trop Stm

607

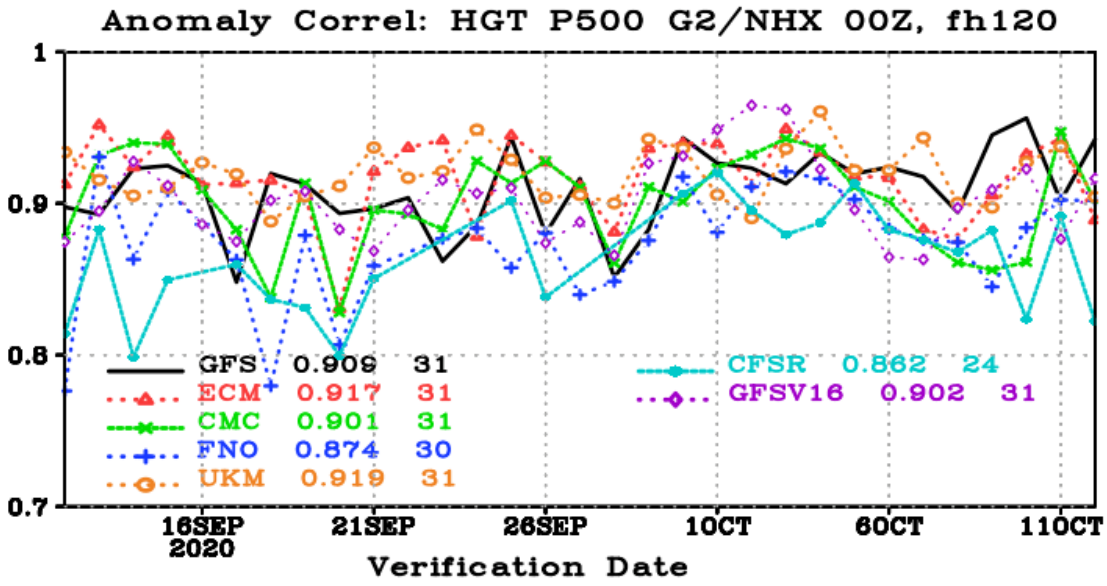
608 **1C**



609

610 Figure 1: a) 18UTC 04 September 2020 (2020090418) JTWC warning (equivalent to NHC
611 advisory) for Typhoon HAISEN; b) 00 UTC 08 October 2020 (2020100800) NHC advisory for
612 hurricane DELTA; and c) Navy equivalent of same DELTA advisory which shows the no-sail
613 zone and the wind radii.

614



615

616 Figure 2: The 30-d time series of 5DNAC for 7 modeling systems over the period 20200912-

617 20201012 courtesy of https://www.emc.ncep.noaa.gov/gmb/SATS_vsdb/

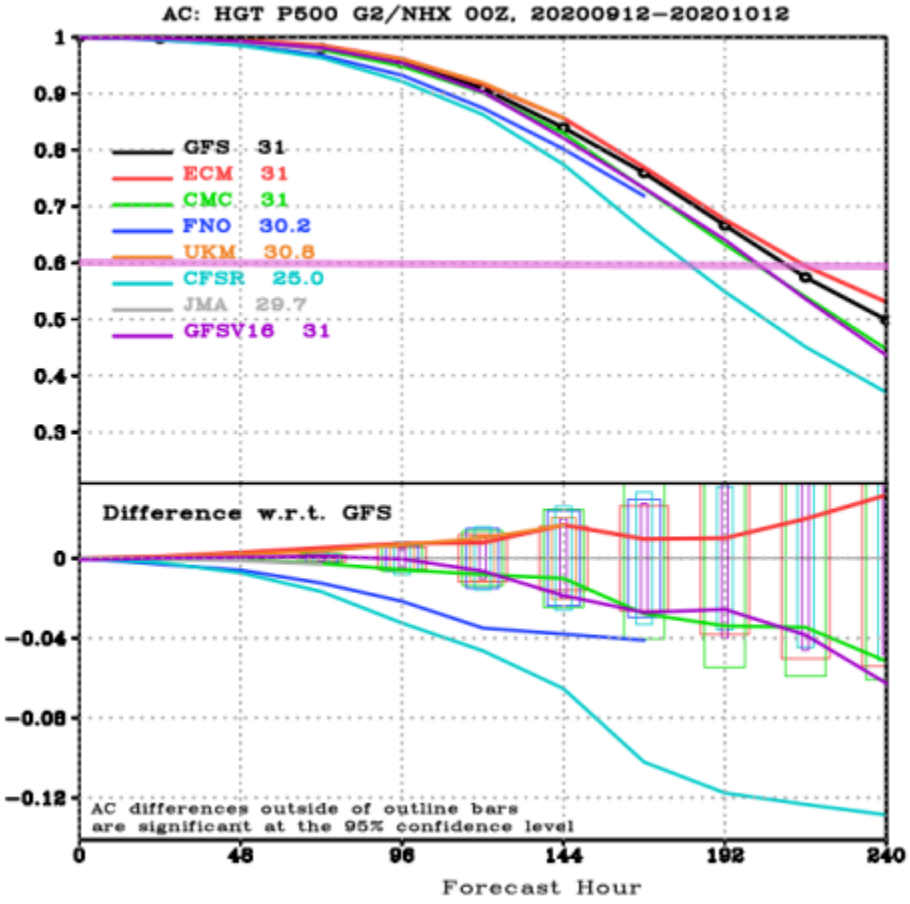
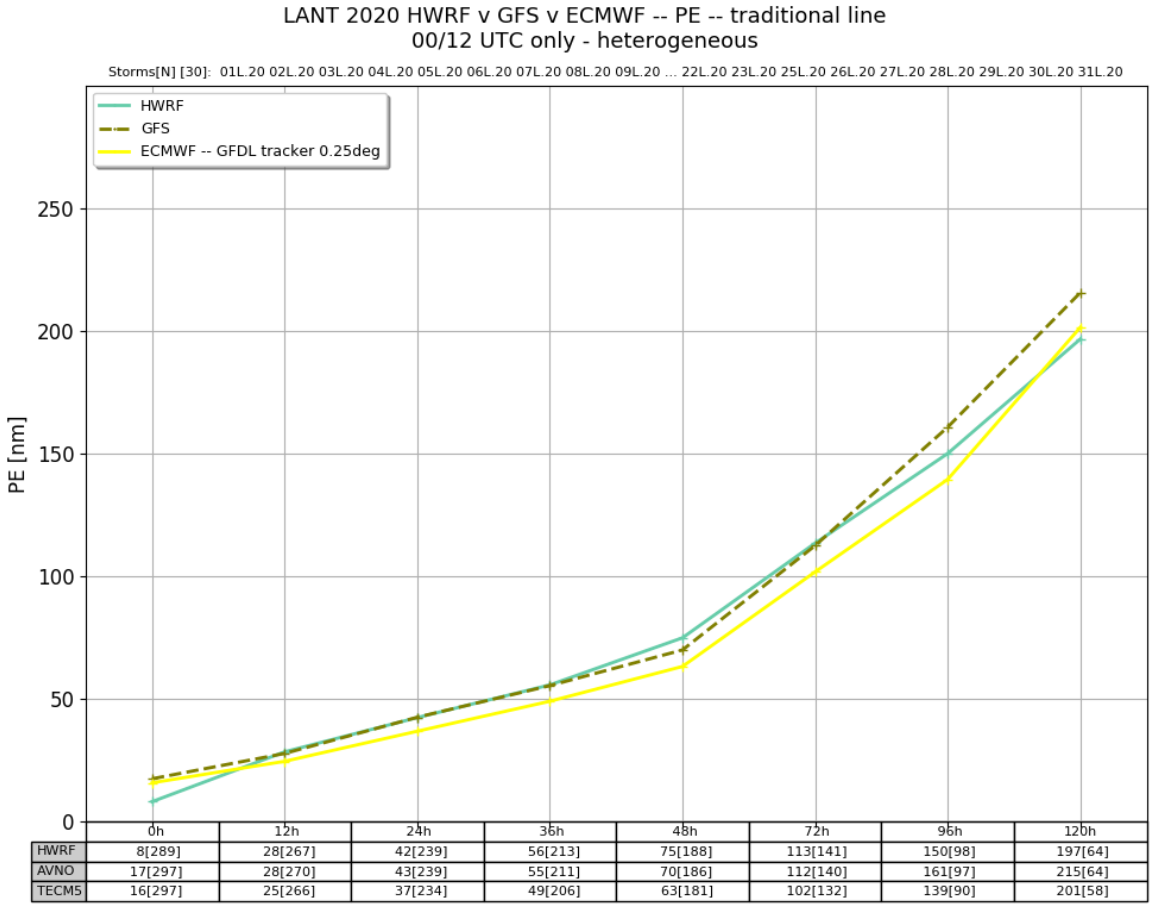


Figure 3: 500 hPa 'die off' curve from the mean of the time series in Fig. 2. The 0.6 line is drawn to show the point the point in time where the model forecast cannot be distinguished from climatology, i.e., no value in deterministic weather forecasting. Available from: https://www.emc.ncep.noaa.gov/gmb/STATS_vsdb/

626

4A



	0h	12h	24h	36h	48h	72h	96h	120h
HWRF	8[289]	28[267]	42[239]	56[213]	75[188]	113[141]	150[98]	197[64]
AVNO	17[297]	28[270]	43[239]	55[211]	70[186]	112[140]	161[97]	215[64]
TECM5	16[297]	25[266]	37[234]	49[206]	63[181]	102[132]	139[90]	201[58]

627

628

4B

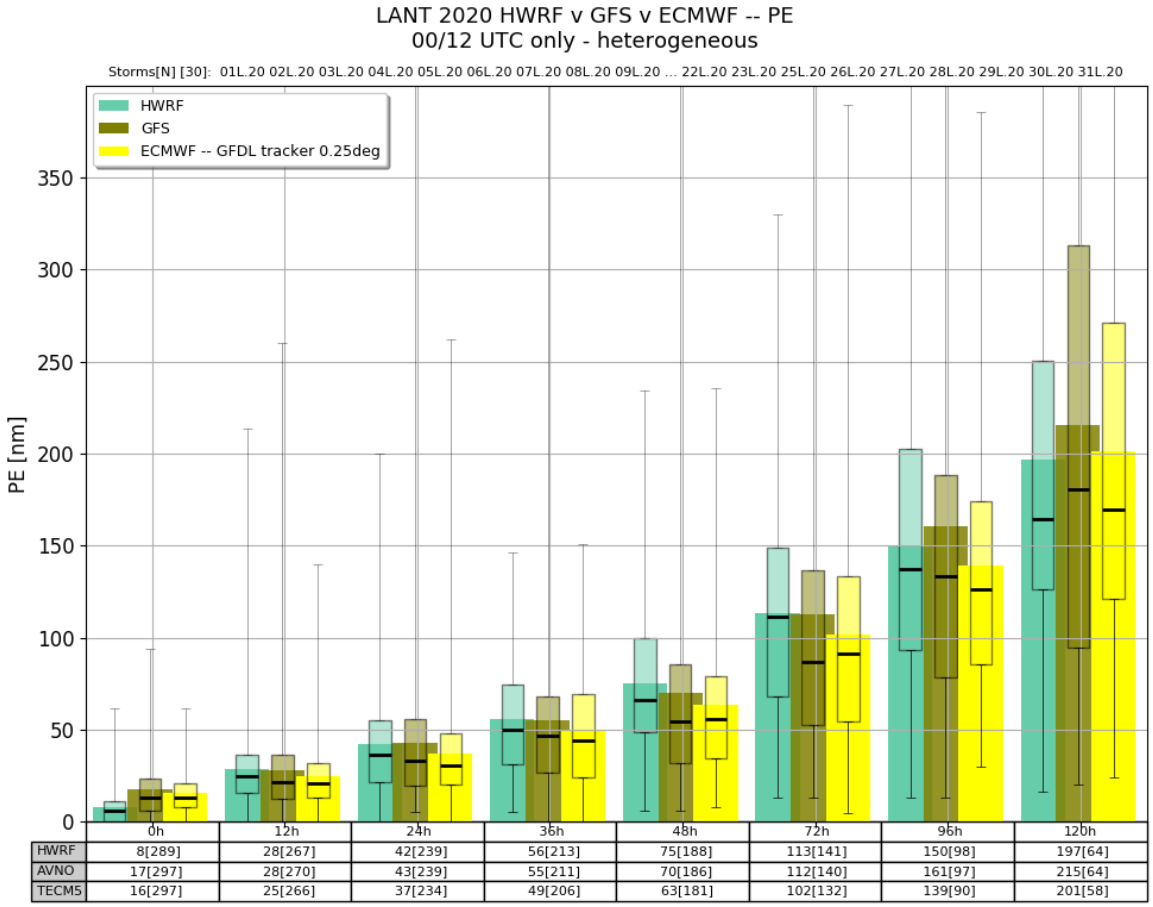
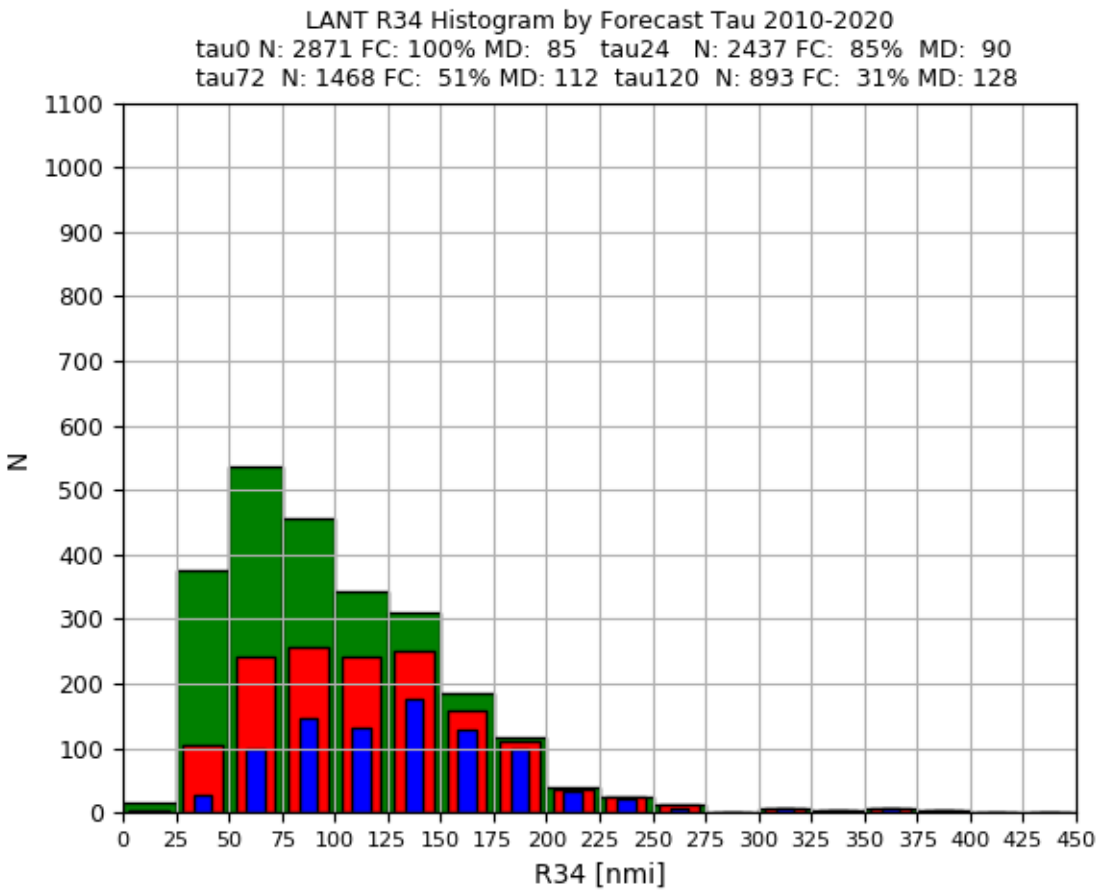


Figure 4. 2020 atANTic mean position error (PE) at standard forecast times for the HWRF, GFS and ECMWF models for 00/12 UTC only. The comparison is heterogeneous. Panel A is the traditional line plot similar to the 500 hPa dieoff curve. Panel B gives a box-whiskers version of the data in 4A showing the distribution of PE. In both panels, the mean value and [number of cases] is shown in the table below the plot. The thick bar in 4B is the mean and the thinner bar extends from the 25th to 75th percentiles with the median indicated by the solid black line.

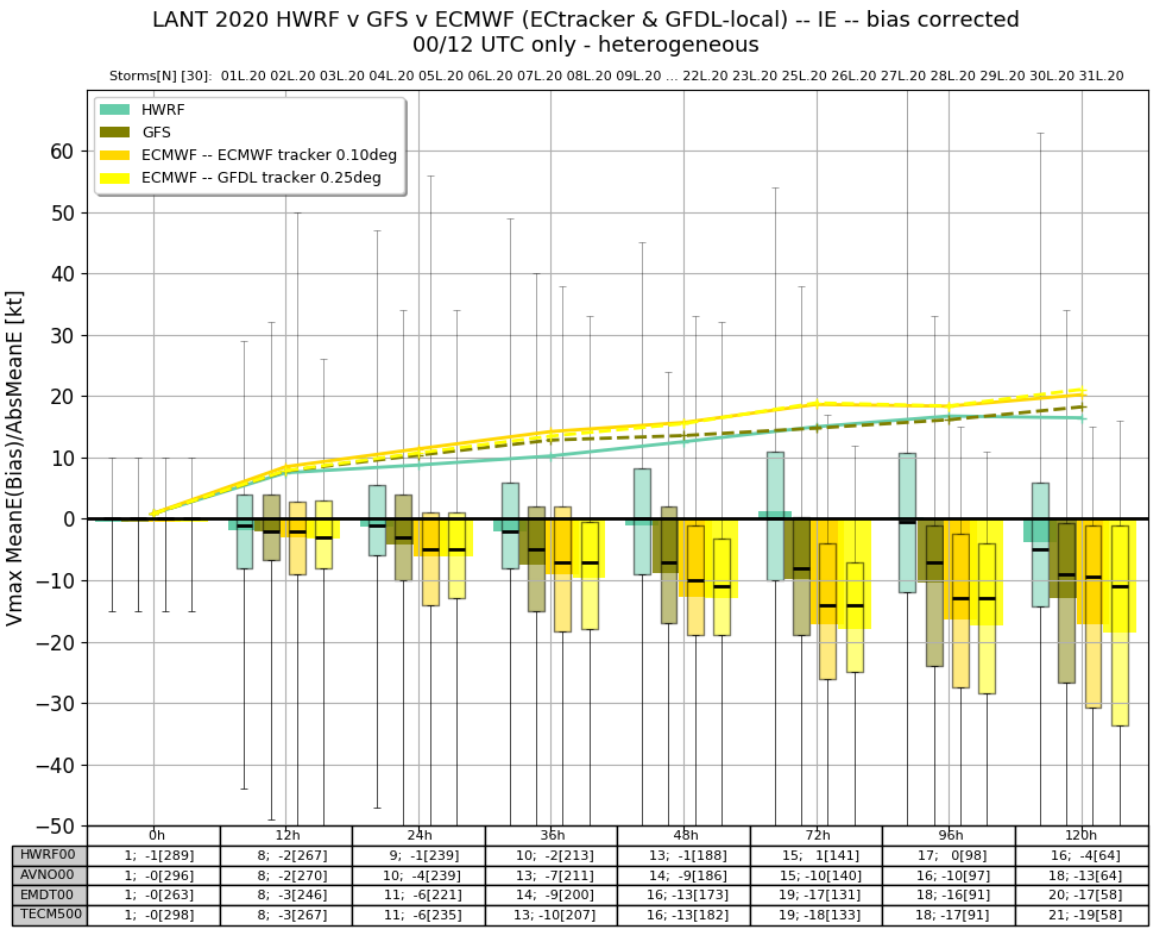


637
638

639

640 Figure 5. Histogram of the radius of 34 kt (R34) for the atLANTic 2010-2020 at three forecast
641 times 24, 72, and 120 h (green, red, and blue bars respectively). The % of forecasts and median
642 R34: 0 h) 100% and 85 n mi; 24 h) is 85% and 90 n mi; 72 h) is 51% and 112 nmi; and 120 h)
643 31% and 128 n mi.

644

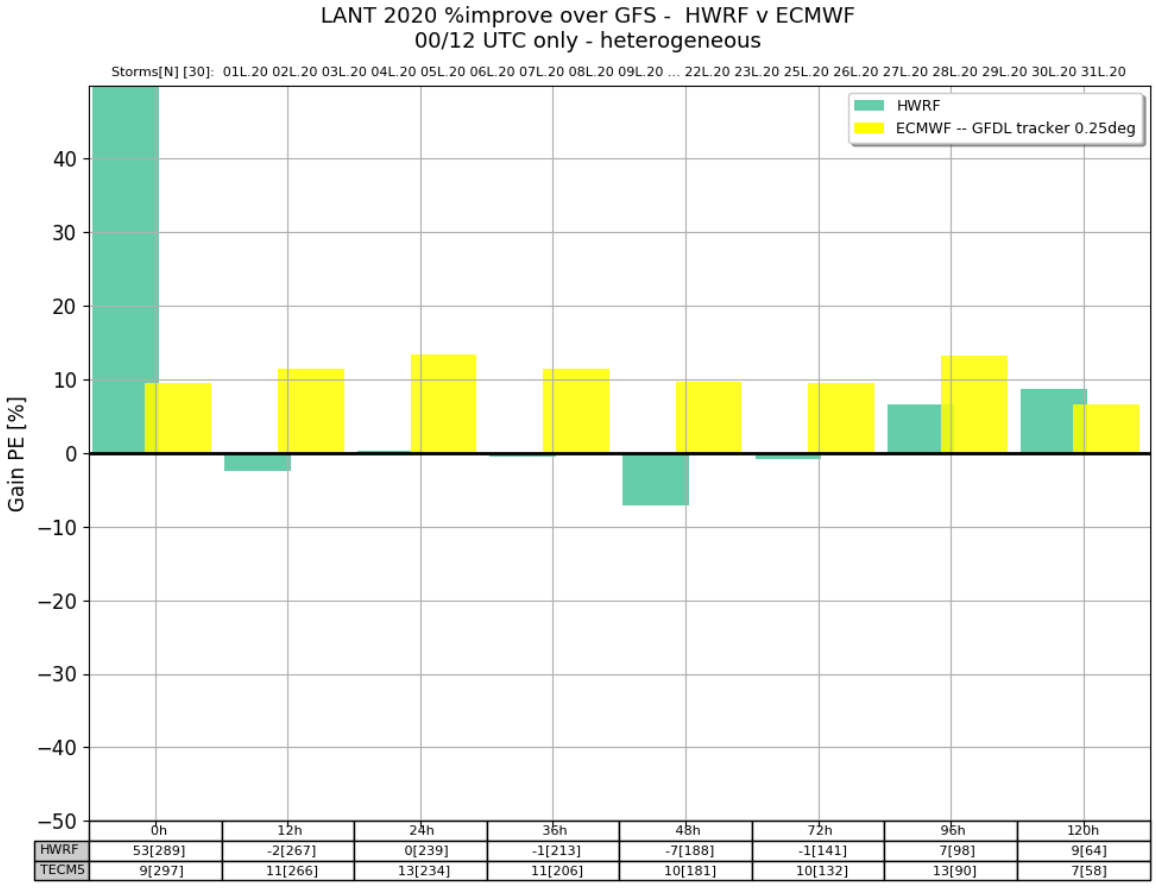


	0h	12h	24h	36h	48h	72h	96h	120h
HWRF00	1; -1[289]	8; -2[267]	9; -1[239]	10; -2[213]	13; -1[188]	15; 1[141]	17; 0[98]	16; -4[64]
AVNO00	1; -0[296]	8; -2[270]	10; -4[239]	13; -7[211]	14; -9[186]	15; -10[140]	16; -10[97]	18; -13[64]
EMDT00	1; -0[263]	8; -3[246]	11; -6[221]	14; -9[200]	16; -13[173]	19; -17[131]	18; -16[91]	20; -17[58]
TECM500	1; -0[298]	8; -3[267]	11; -6[235]	13; -10[207]	16; -13[182]	19; -18[133]	18; -17[91]	21; -19[58]

Figure 6. 2020 atANTic mean absolute IE (lines) at standard forecast times for the HWRF, GFS and ECMWF models for 00/12 UTC only. The comparison is heterogeneous as in Fig. 4. The bars are the mean IE (bias, defined as model-observed). The thick bar is the mean and the thinner bar extends from the 25th to 75th percentiles with the median indicated by the solid black line (box and whisker). The mean value and [number of cases] is shown in the table below the plot.

652 . The bias correction is the same as used in operations. The line is the absolute mean and the

653 bars are the mean or intensity bias.



654

655 Figure 7. Percent improvement in PE (lower PE is a positive improvement see Eq. 1) of HWRF

656 v ECMWF using the GFS as the baseline for the 2020 LANT season (same data as in Fig. 4).

657 This plot is similar to NHC 'skill' where climatology and persistence is the baseline (Cangialosi

658 2020 Figure 5).

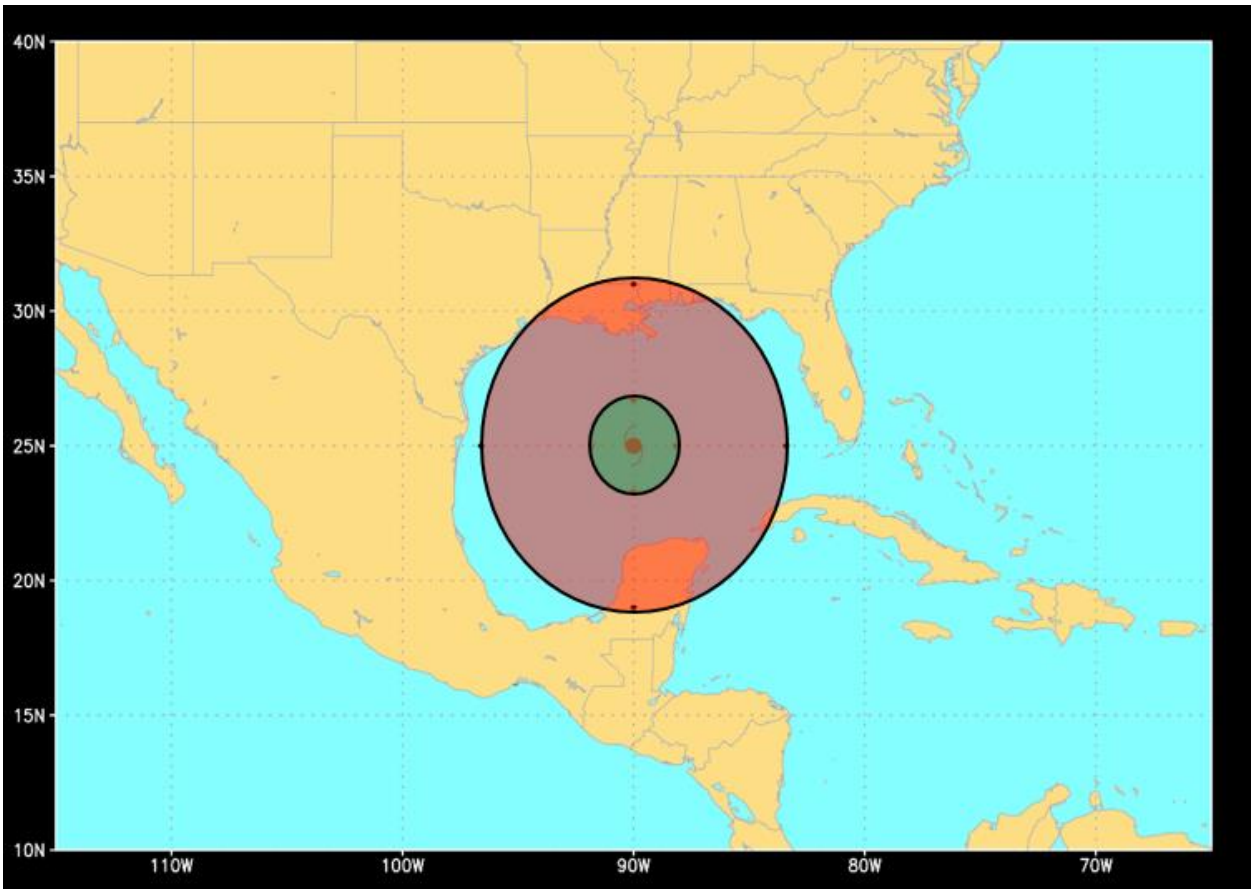
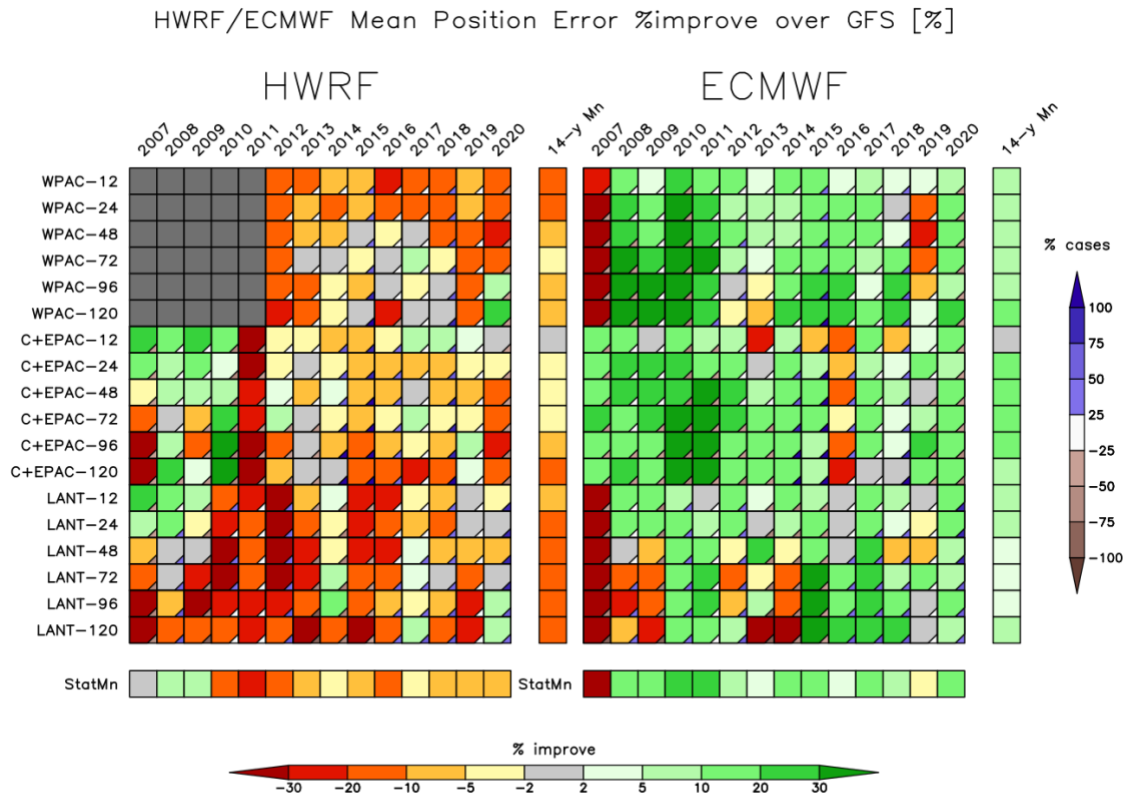


Figure 8. Area of a 360 n mi mPE (red) vs area of 100 n mi mPE (green).

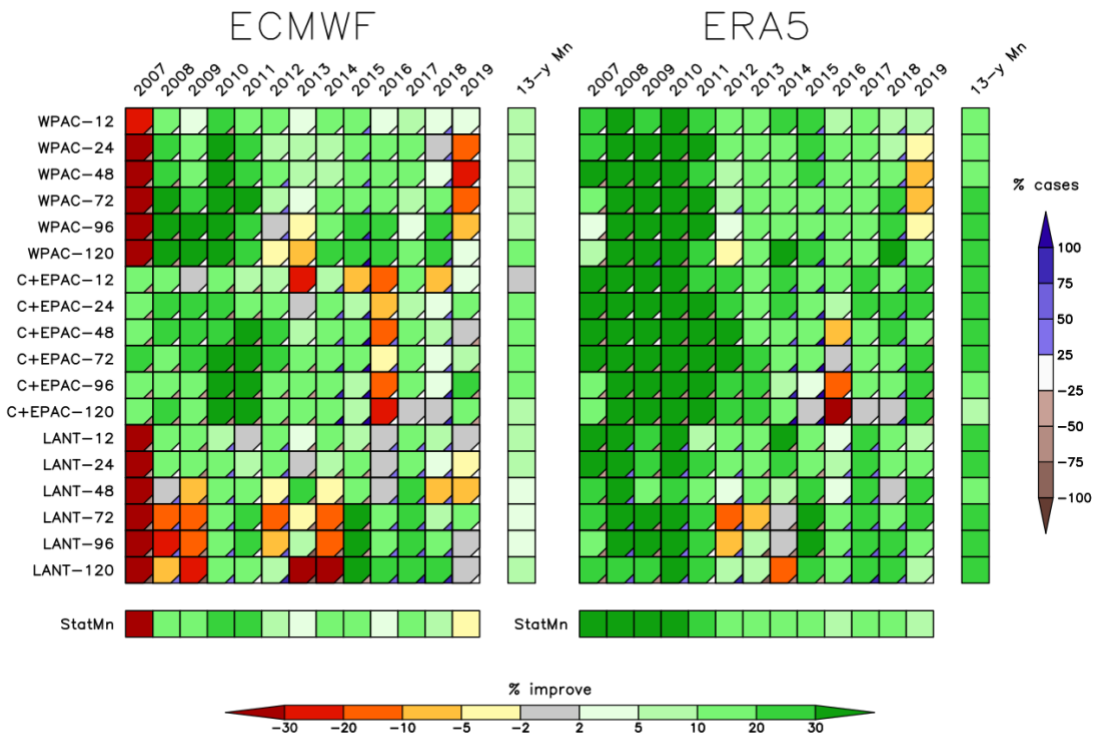
662 **9A**



663
664 Figure 9. $\%IMP_{PE}$ for the NHEM basins 2007-2020. The $\%IMP_{PE}$ is colorized so that green is
665 'good' (lower mPE) than the GFS baseline and red is 'worse' (higher mPE). The % cases
666 contributing to the mean is shown in the triangle in the lower right-hand corner of each box.
667 Panel A compares HWRF to ECMWF operational deterministic run. Panel B compares
668 ECMWF (HRES) to forecasts from the latest ECMWF reanalysis (ERA5).

669

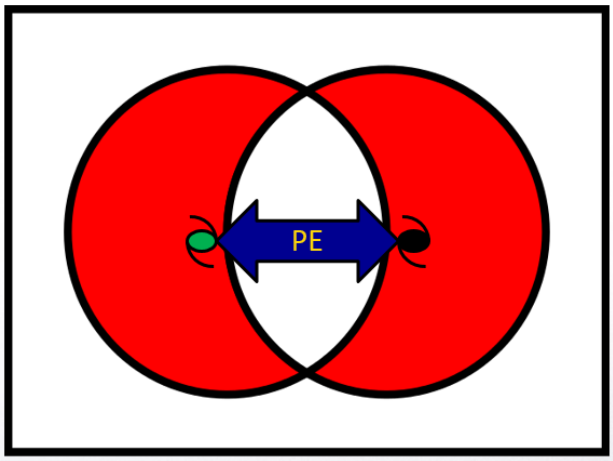
9B



670
671

672

- FE ≡ IKE_{fc} Δ IKE_{bt} assuming:
  - $R_{34}^{fc} = R_{34}^{bt}$
  - $r_{max}^{fc} = r_{max}^{bt}$
- FE = f(PE,IE); IE = 0 @ $2R_{34}$
- FE ranges from 0 to a max at PE=$2R_{34}$ for PE > $2R_{34}$ linear increase



673

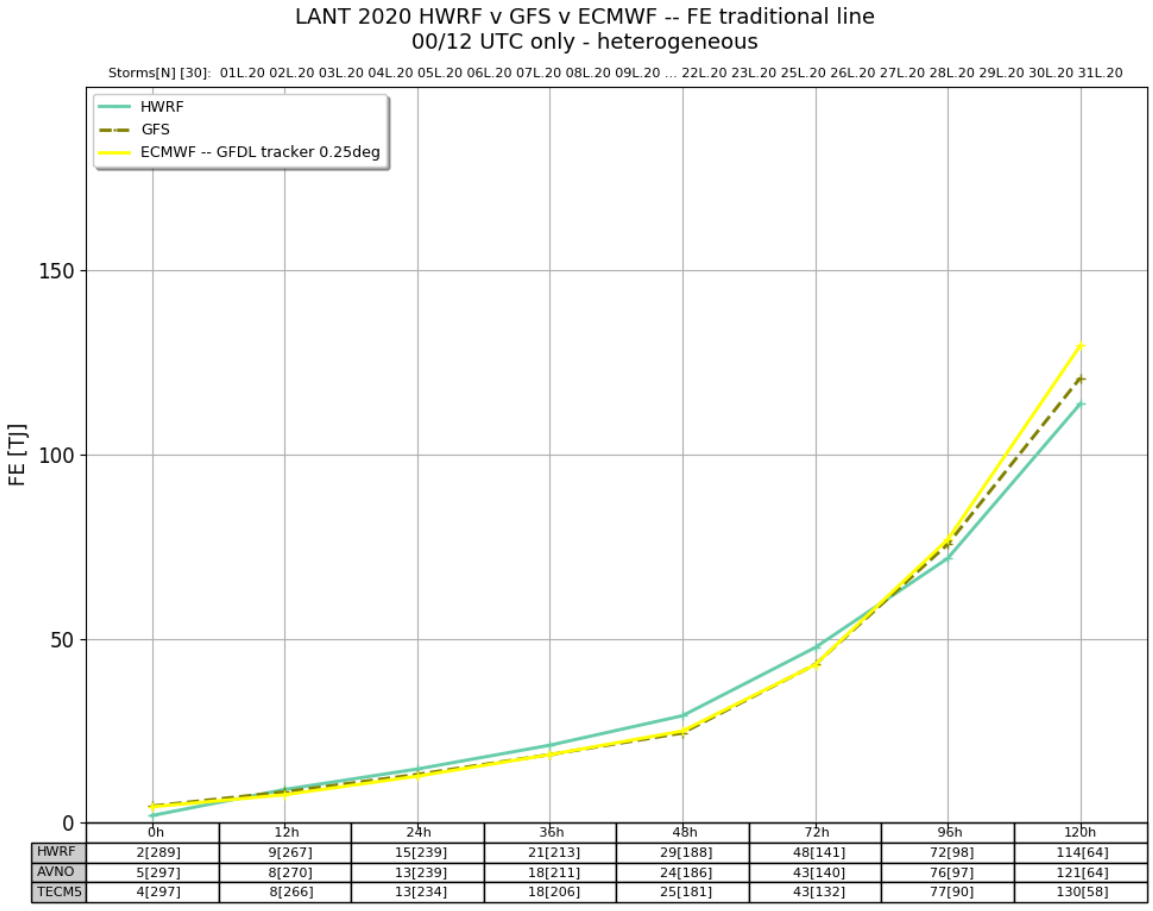
674 Figure 10. FE is the IKE in the 'symmetric difference' (red area) and how FE is related to PE.

675 Details of the calculation are also given. The *fc* superscript indicates a forecast and the *bt*

676 superscript the verifying best track.

677

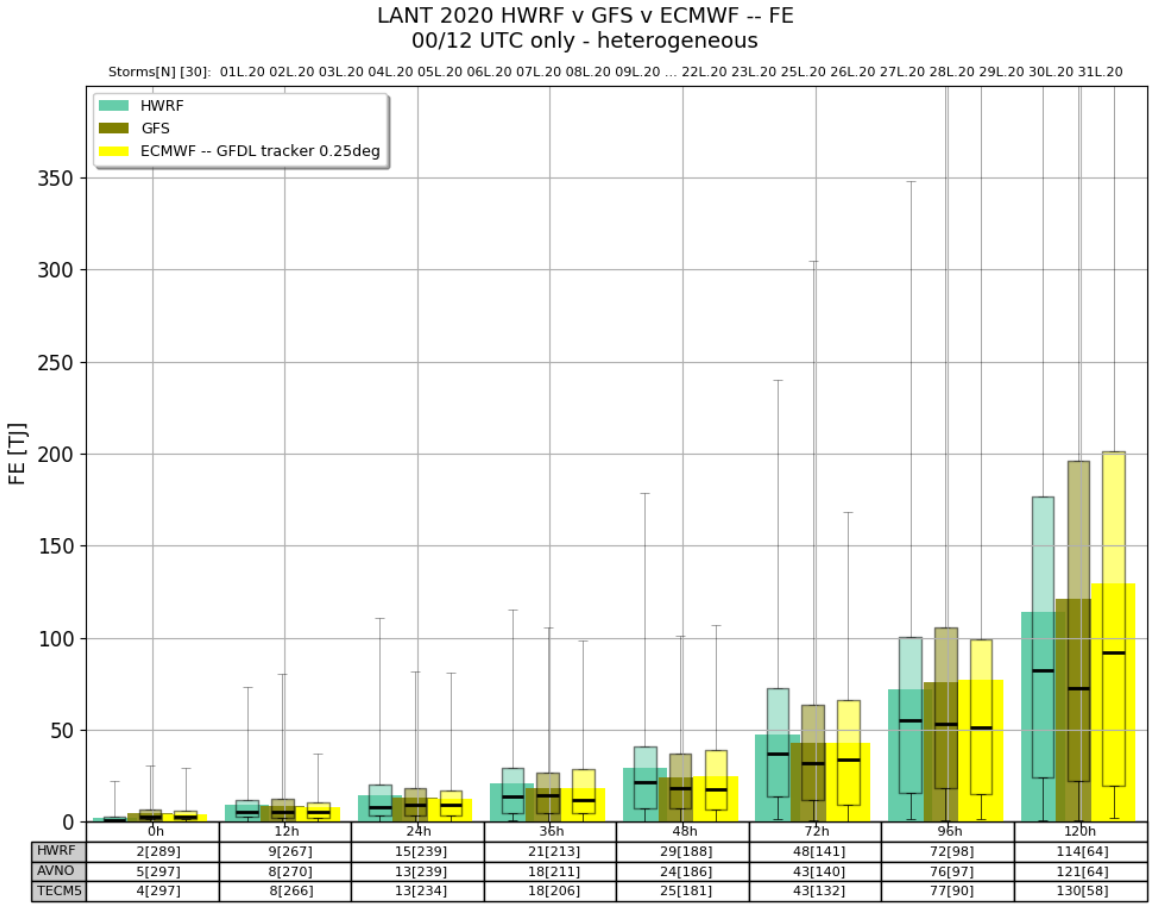
11A



	0h	12h	24h	36h	48h	72h	96h	120h
HWRF	2[289]	9[267]	15[239]	21[213]	29[188]	48[141]	72[98]	114[64]
AVNO	5[297]	8[270]	13[239]	18[211]	24[186]	43[140]	76[97]	121[64]
TECM5	4[297]	8[266]	13[234]	18[206]	25[181]	43[132]	77[90]	130[58]

678

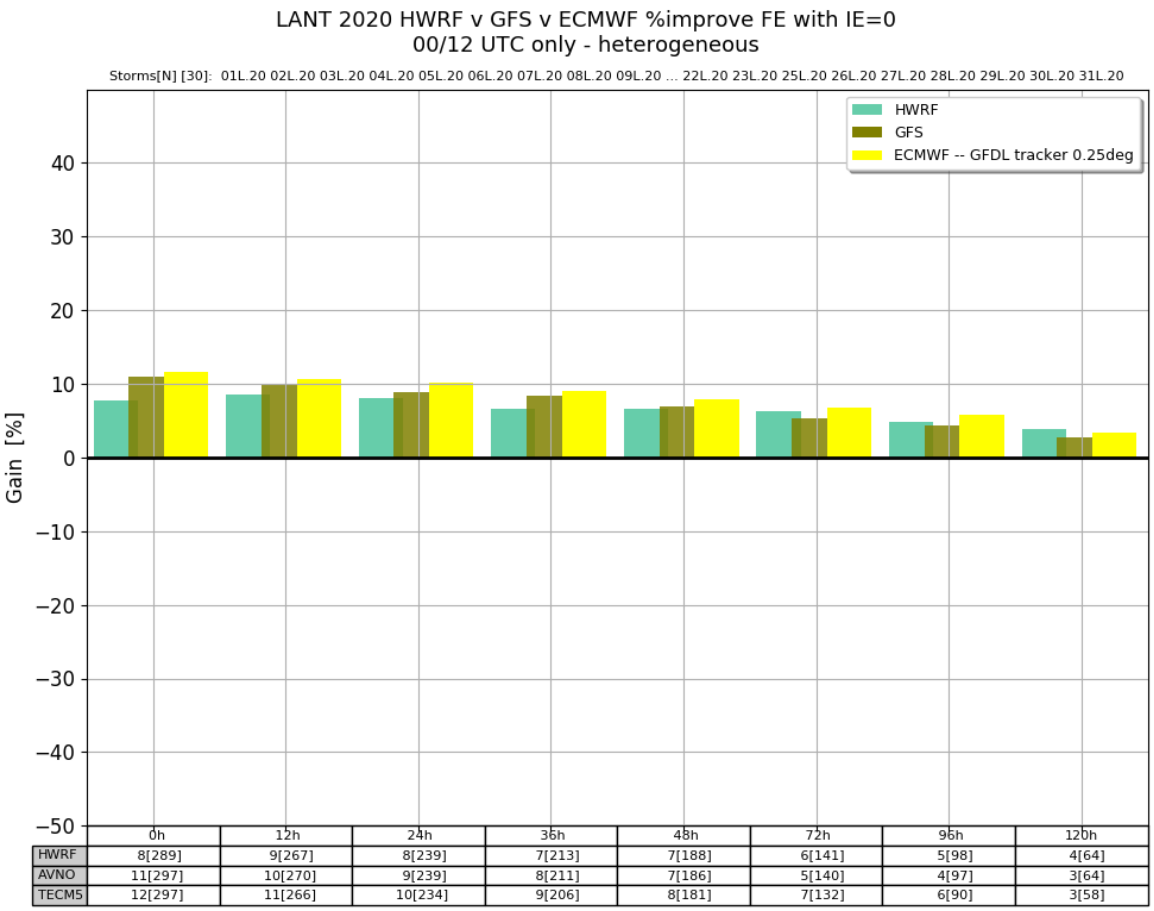
679 **11B**



680

681 Figure 11. as in Fig. 4 except the metric is FE units: TJ. Panel A is the traditional line and B is

682 the box-whiskers version.



	0h	12h	24h	36h	48h	72h	96h	120h
HWRF	8[289]	9[267]	8[239]	7[213]	7[188]	6[141]	5[98]	4[64]
AVNO	11[297]	10[270]	9[239]	8[211]	7[186]	5[140]	4[97]	3[64]
TECM5	12[297]	11[266]	10[234]	9[206]	8[181]	7[132]	6[90]	3[58]

683

684 Figure 12. as in Fig. 4 except metric is percent improve in FE assuming a perfect intensity

685 forecast or IE=0.

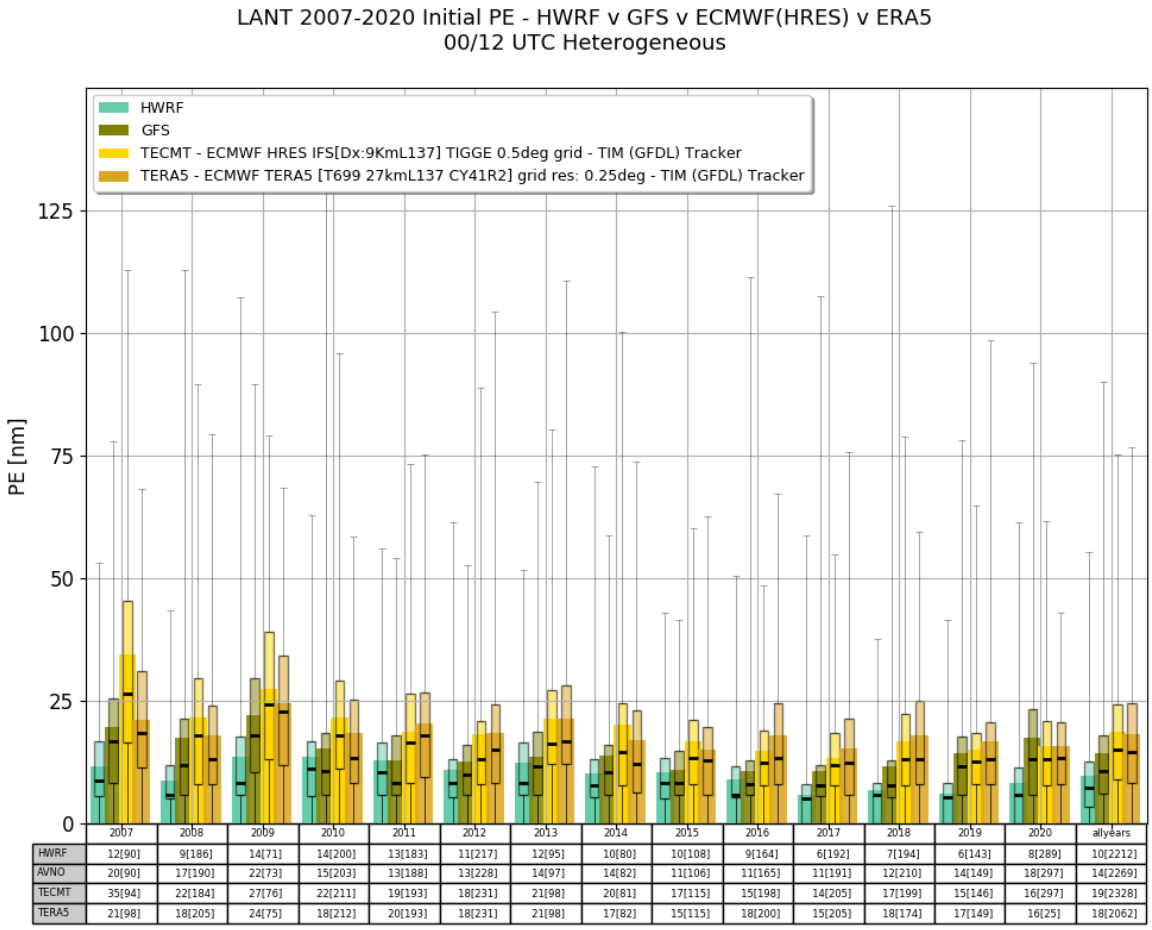
686

687

688

689

13A

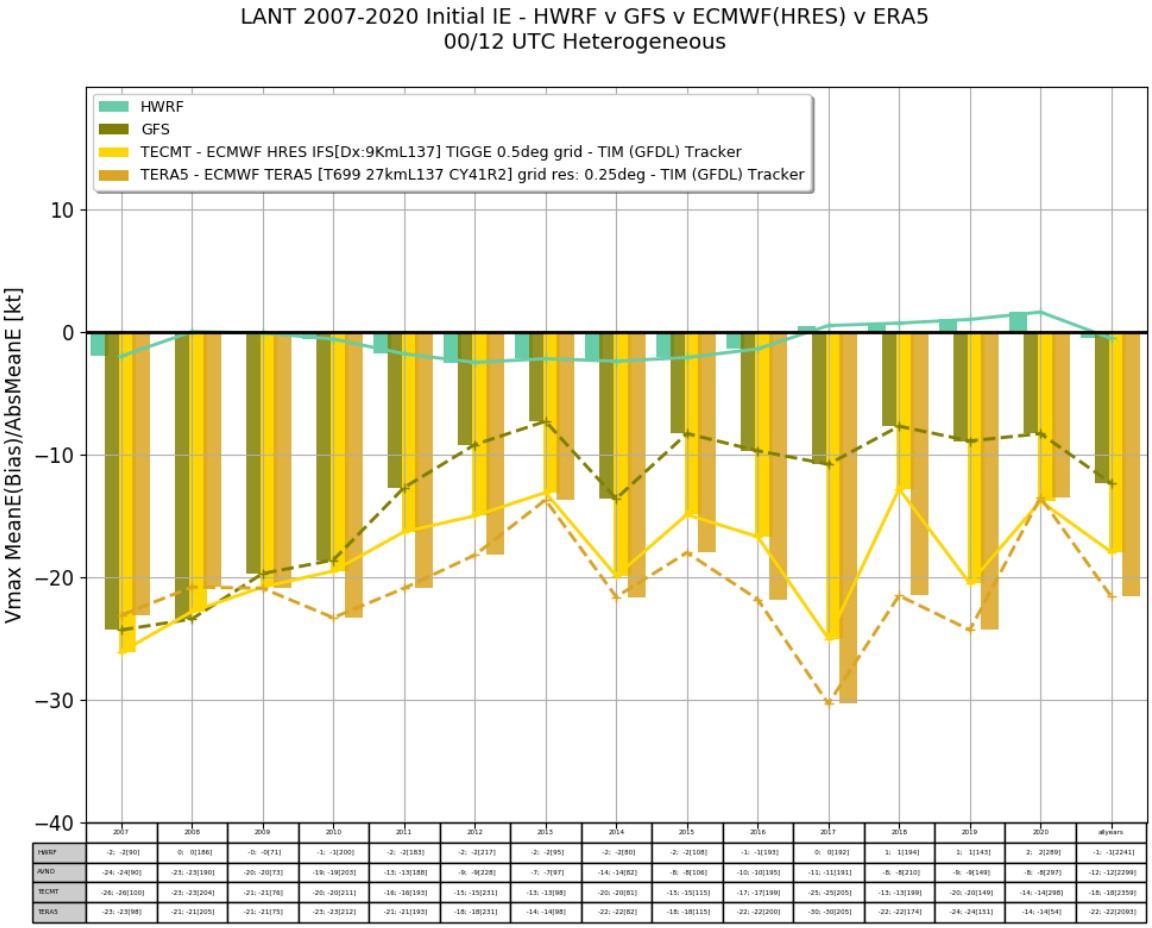


	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	allyears
HWRF	12[90]	9[186]	14[71]	14[200]	13[183]	11[217]	12[95]	10[80]	10[108]	9[164]	6[192]	7[194]	6[143]	8[289]	10[2212]
AVNO	20[90]	17[190]	22[73]	15[203]	13[188]	13[228]	14[97]	14[82]	11[106]	11[165]	11[191]	12[210]	14[149]	18[297]	14[2269]
TECMT	35[94]	22[184]	27[76]	22[211]	19[193]	18[231]	21[98]	20[81]	17[115]	15[198]	14[205]	17[199]	15[146]	16[297]	19[2328]
TERA5	21[98]	18[205]	24[75]	18[212]	20[193]	18[231]	21[98]	17[82]	15[115]	18[200]	15[205]	18[174]	17[149]	16[25]	18[2062]

690

691

692 **13B**



	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	allyears
HWRF	-2; -2[90]	0; 0[186]	-0; -0[71]	-1; -1[200]	-2; -2[183]	-2; -2[217]	-2; -2[95]	-2; -2[80]	-2; -2[108]	-1; -1[193]	0; 0[192]	1; 1[194]	1; 1[145]	2; 2[289]	-1; -1[2241]
AVNO	-24; -24[90]	-23; -23[190]	-20; -20[73]	-19; -19[203]	-13; -13[188]	-9; -9[228]	-7; -7[97]	-14; -14[82]	-8; -8[106]	-10; -10[195]	-11; -11[191]	-8; -8[210]	-9; -9[149]	-8; -8[297]	-12; -12[2299]
TECMT	-26; -26[100]	-23; -23[204]	-21; -21[76]	-20; -20[211]	-16; -16[193]	-15; -15[231]	-13; -13[98]	-20; -20[81]	-15; -15[115]	-17; -17[199]	-25; -25[205]	-13; -13[199]	-20; -20[149]	-14; -14[298]	-18; -18[2359]
TERA5	-23; -23[98]	-21; -21[205]	-21; -21[75]	-23; -23[212]	-21; -21[193]	-18; -18[231]	-14; -14[98]	-22; -22[82]	-18; -18[115]	-22; -22[200]	-30; -30[205]	-22; -22[174]	-24; -24[151]	-14; -14[54]	-22; -22[2093]

693

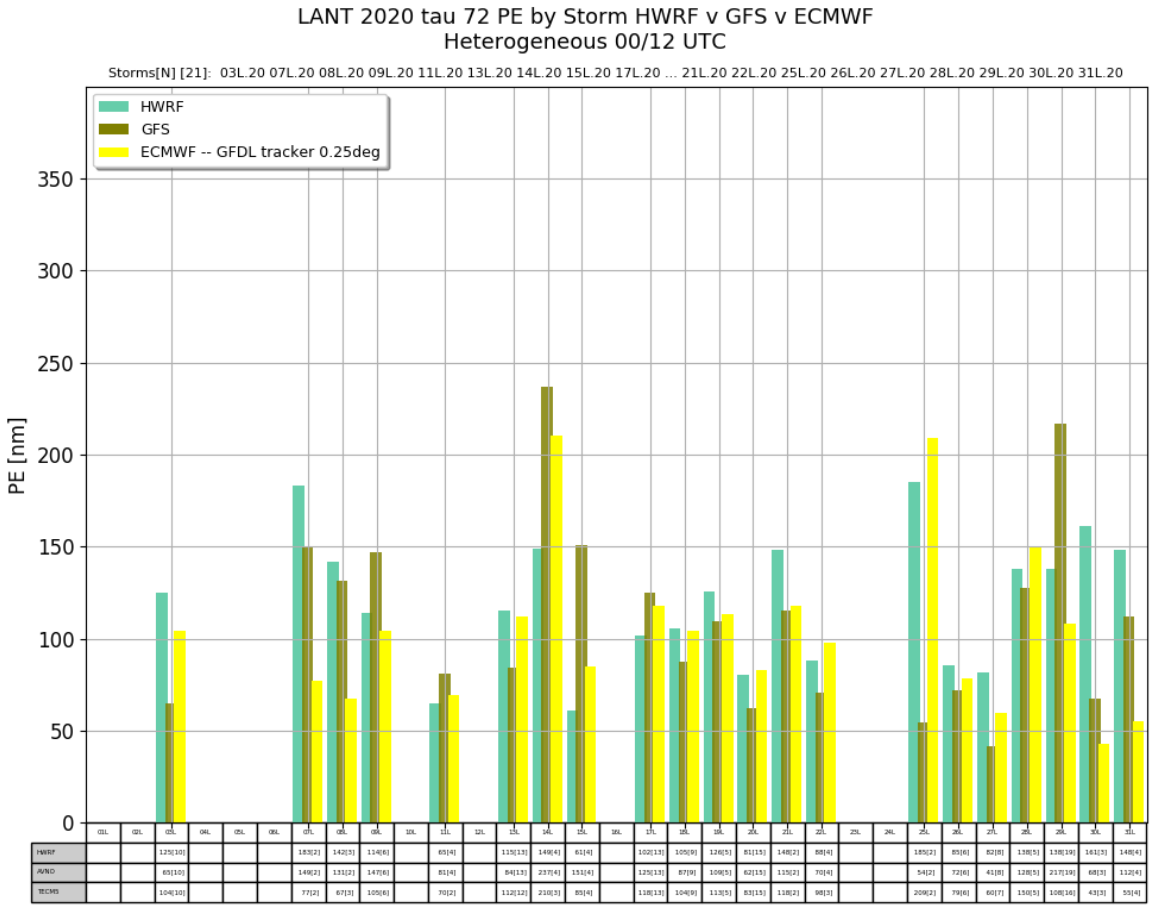
694 Figure 13. initial PE (A) and initial IE (B) for HWRF v GFS v ECMMWF HRES v ERA5 in the

695 LANT 2007-2020.

696

697

14A



	01L	02L	03L	04L	05L	06L	07L	08L	09L	10L	11L	12L	13L	14L	15L	16L	17L	18L	19L	20L	21L	22L	23L	24L	25L	26L	27L	28L	29L	30L	31L
HWRF			125[10]				183[2]	142[3]	114[6]		65[4]		115[13]	149[4]	61[4]		102[13]	105[0]	126[5]	81[15]	148[2]	88[4]			185[2]	85[6]	82[8]	138[5]	138[19]	161[5]	148[4]
AVNO			65[10]				149[2]	131[2]	147[6]		81[4]		84[13]	237[4]	151[4]		125[13]	87[9]	109[5]	62[15]	115[2]	70[4]			54[2]	72[6]	41[8]	128[5]	217[19]	68[3]	112[4]
TECMS			104[10]				77[2]	67[3]	105[6]		70[2]		112[12]	210[3]	85[4]		118[13]	104[9]	113[5]	83[15]	118[2]	98[3]			209[2]	79[6]	60[7]	150[5]	108[16]	43[3]	55[4]

698
699
700

14B

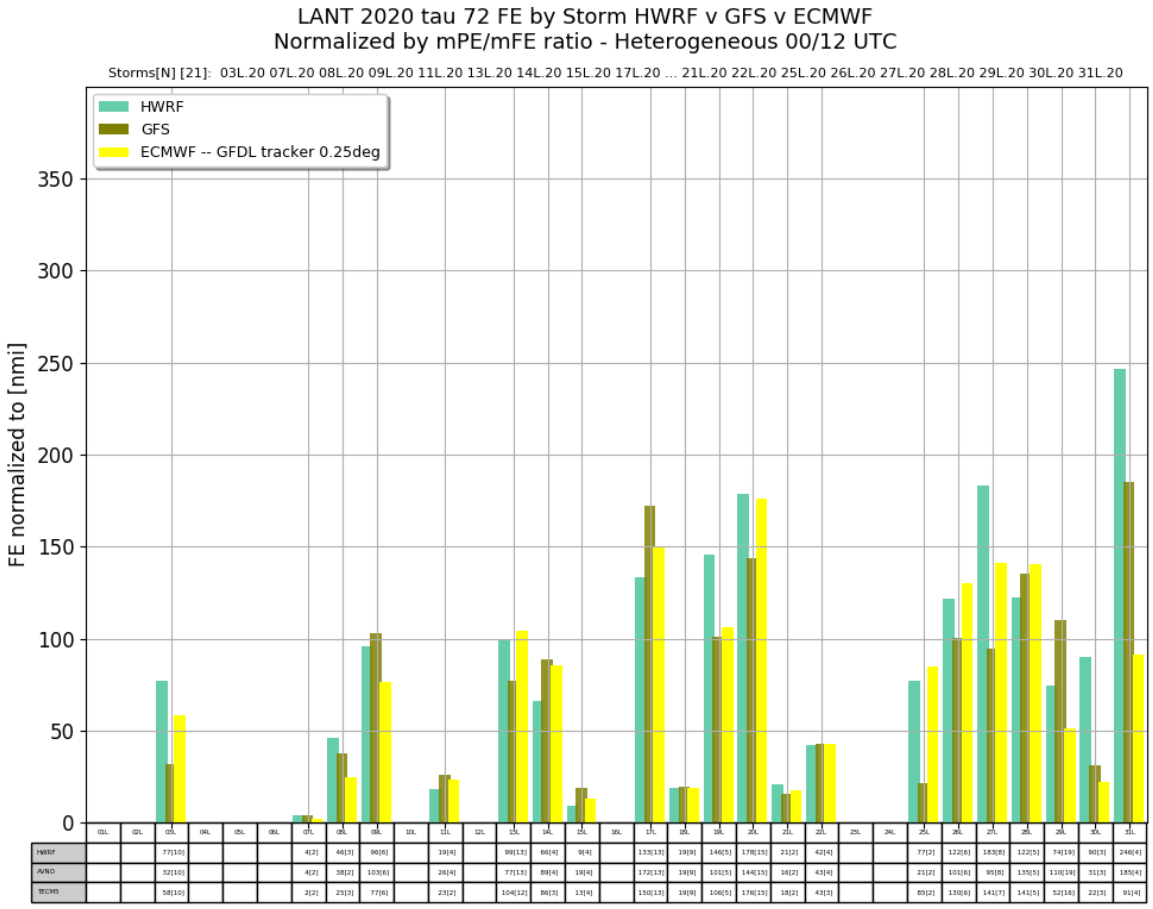
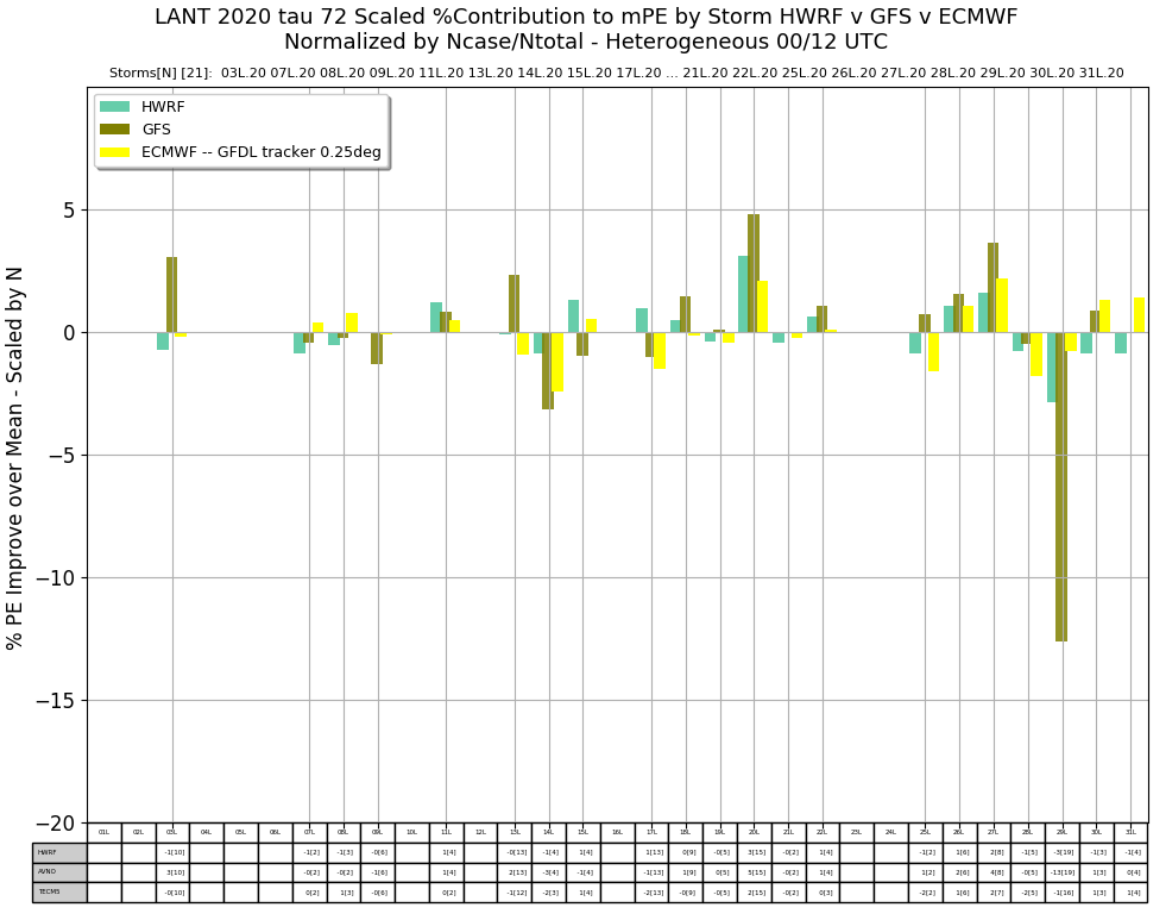


Figure 14. 72-h mPE (A) and 72-h mFE (B) by storm normalized by mean ratio of mPE/mFE at 72 h to put FE in same units and scale as PE. The total number of storms verifying at 72 h is 21 out of 31 for the season. The number of verifying 72-h cases is 68% which is higher than 10-y average of 51% as shown in Fig. 5.

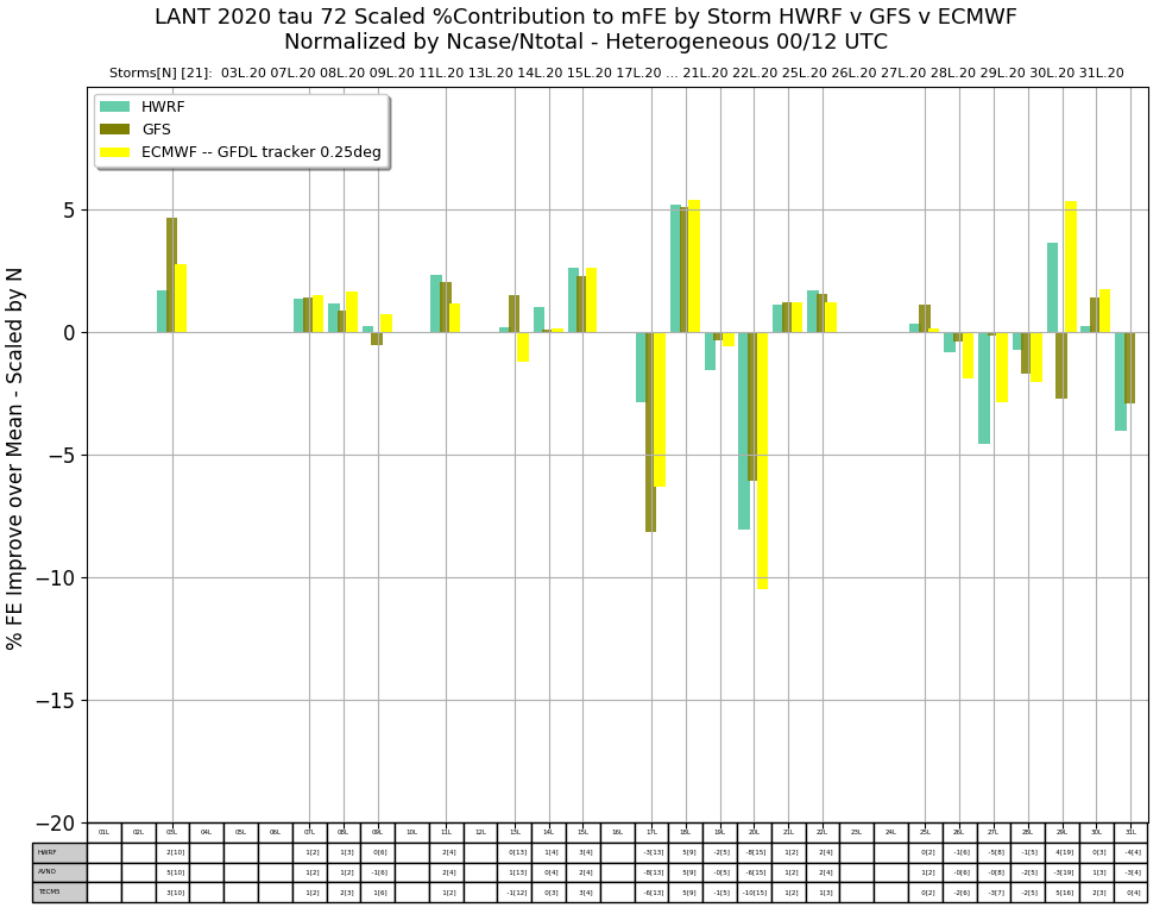
709

15A



710
711

712 **15B**



713
714 Figure 15. Same cases/storms as in Fig. 14 but the metric is $\%IMP_{P/FE}$ of the contribution by
715 each storm to the 2020 season to the 72-h mPE (A) and mFE (B) normalized by ratio of the
716 number of cases by storm and divided by the total number of cases. The mean across all storms
717 for this scaled percent improvement is 0.

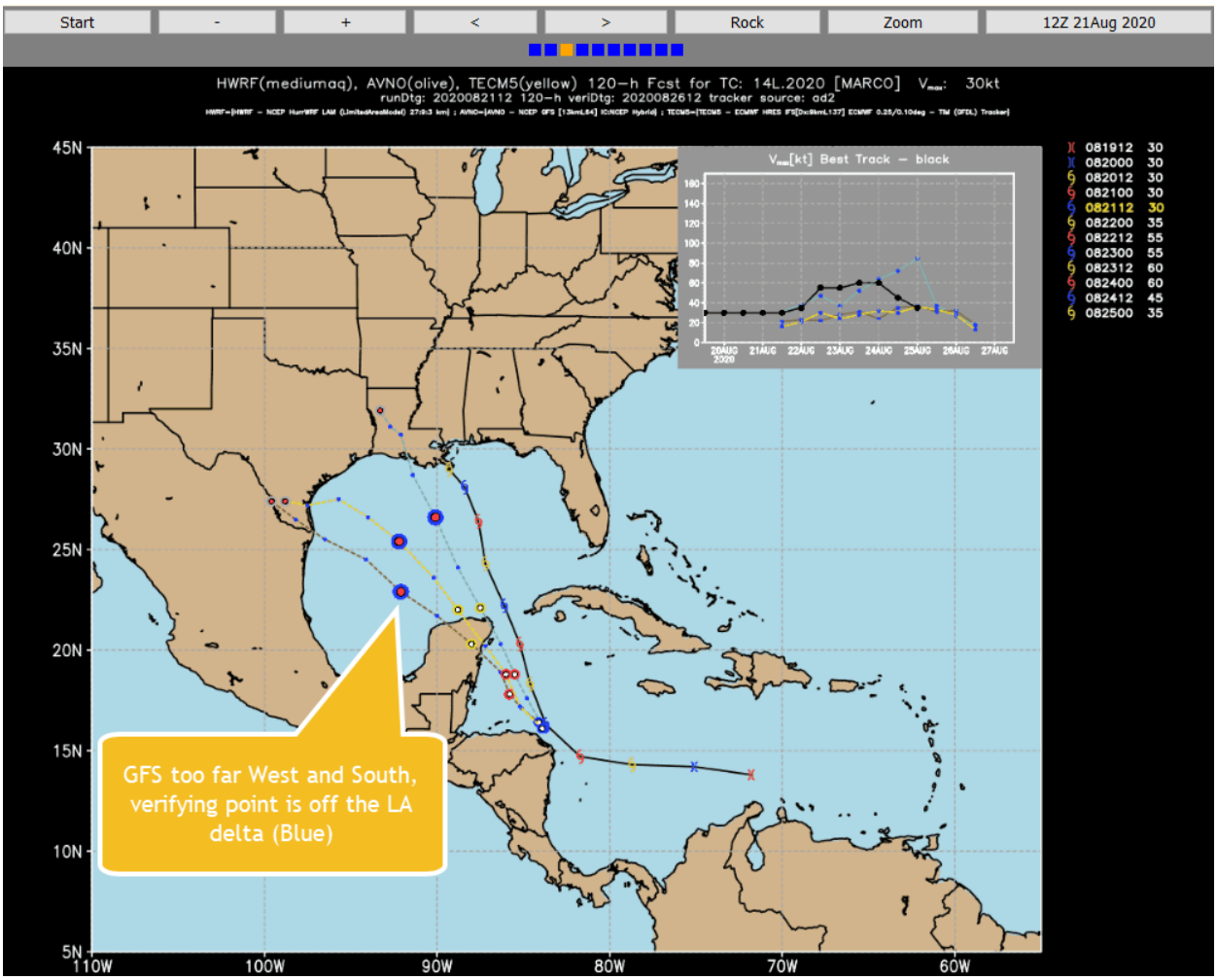


Figure 16. HWRF (light blue) / ECWWF (gold) / GFS (olive green) forecast for 14L (MARCO) on 2020082112 (12UTC 21 Aug 2020). The intensity forecast is show in the upper-right hand insert using the same color code as the tracks with the thick black line the working best track intensity. See: http://tctrkveri.wxmap2.com/trk-14L-2020-MOD-CUR.htm for all forecasts used in the verification.

Kieu, C. Q., and Z. Moon, 2016: Hurricane Intensity Predictability. *B Am Meteorol Soc*, **97,** 1847-1857.

Powell, M. D., and T. A. Reinhold, 2007: Tropical cyclone destructive potential by integrated kinetic energy. *B Am Meteorol Soc*, **88,** 513-+.

730 Simmons, A. J., and A. Hollingsworth, 2002: Some aspects of the improvement in skill of
731 numerical weather prediction. *Q J Roy Meteor Soc*, **128,** 647-677.
732 Walsh, K. J. E., M. Fiorino, C. W. Landsea, and K. L. McInnes, 2007: Objectively determined
733 resolution-dependent threshold criteria for the detection of tropical cyclones in climate models
734 and reanalyses. *J Climate*, **20,** 2307-2314.
735