The Spread of Tropical Storm Tracks in Three Versions of NCEP's Global Ensemble Model: Focus on Hurricane Edouard (2014)

FRANK P. COLBY JR.

University of Massachusetts Lowell, Lowell, Massachusetts

(Manuscript received 7 September 2018, in final form 11 March 2019)

ABSTRACT

Since 2012, the National Centers for Environmental Prediction's Global Ensemble Forecast System (GEFS) has undergone two major upgrades. Version 11 was introduced in December 2015, with a new dynamic scheme, improved physics, increased horizontal and vertical resolution, and a more accurate initialization method. Prior to implementation, retrospective model runs over four years were made, covering multiple hurricane seasons. The second major upgrade was implemented in May 2016, when the data assimilation system for the deterministic Global Forecast System (GFS) was upgraded. Because the GEFS initialization is taken from the deterministic GFS, this upgrade had a direct impact on the GEFS. Unlike the previous upgrade, the model was rerun for only a few tropical cyclones. Hurricane Edouard (2014) was the storm for which the most retrospective runs (4) were made for the new data assimilation system. In this paper, the impact of the GEFS upgrades is examined using seasonal data for the 2014–17 hurricane seasons, and detailed data from the four model runs made for Hurricane Edouard. Both upgrades reduced the spread between ensemble member tracks. The first upgrade reduced the spread but did not reduce the likelihood that the actual track would be included in the family of member tracks. The second upgrade both reduced the spread further and reduced the chance that the real storm track would be within the envelope of member tracks.

1. Introduction

Hurricanes and tropical storms often cause serious damage, leading to injury or death. The average loss of life due to North Atlantic tropical storms is just over 50 per year (Rappaport 2014). The Congressional Budget Office reports annual property damage at \$28 billion (CBO 2016). Year-to-year variations are large. For example, Hurricane Sandy (2012) alone was responsible for 147 deaths and damage near \$50 billion. Timely warnings save lives and property, but warnings depend on accurate forecasts of storm tracks.

The National Hurricane Center (NHC) uses many forecast models to help forecast the track and intensity of tropical cyclones. One of these models is the National Centers for Environmental Prediction's (NCEP) Global Ensemble Forecast System (GEFS). The model is initialized from NCEP's deterministic, operational Global Forecast System model (GFS) and runs at roughly half the horizontal resolution of the deterministic GFS. The initial conditions in the 20 ensemble members are perturbed from the GFS initialization to provide a range

of possible outcomes. One of the intentions of using ensemble models is to give forecasters a sense of the uncertainty in the forecast track. The 20 different member tracks can provide guidance about the range of possible outcomes, with the hope that the actual storm track will fall within the envelope of the ensemble member forecasts.

Version 10 of the GEFS (GEFSv10) was first introduced in 2012 and was operational until December 2015. Following an upgrade to the deterministic GFS in January 2015, the GEFS model core was upgraded in December 2015 to use a new dynamic scheme and improved physics. The horizontal resolution of the model was increased from a nominal 55 km to a nominal 33 km, and the vertical resolution was increased from 42 model levels to 64 levels. In addition, the ensemble initialization method was upgraded to a more accurate method based on the ensemble Kalman filter (Zhou et al. 2016). This version of the model is GEFSv11.

A second major upgrade to the GEFS occurred in May 2016, when the data assimilation for the deterministic GFS was upgraded from a 3D system [hybrid 3D ensemble-variational data assimilation (EnVar)] to a 4D system (hybrid 4DEnVar). Since the GEFS receives its basic initialization from the deterministic GFS model, the

Corresponding author: Frank Colby, frank_colby@uml.edu

DOI: 10.1175/WAF-D-18-0153.1

© 2019 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy](https://www.ametsoc.org/PUBSReuseLicenses) (www.ametsoc.org/PUBSReuseLicenses).

new data assimilation system had significant impacts on the GEFS as well. This model and data assimilation system will be labeled the GEFSv11da.

According to [NCEP \(2018\)](#), the GEFSv11da will continue to be NCEP's operational global ensemble model until at least the end of 2019 (FY2020). Understanding the behavior of this model, especially in comparison with previous operational versions, will be helpful to all forecasters who use this model. Additionally, evaluation of some details of this model should help model developers as they work toward the next version.

The NHC calculates track errors by computing the great circle distance between a forecast location and the NHC best track location at the same time. For the GEFS system as a whole, the ensemble mean forecast, which is the average of the 20 ensemble member forecast locations, is used to measure overall accuracy. [Leonardo and Colle \(2017\)](#) compared the GEFSv10 with other ensemble models to determine the kinds of track errors that were present in forecasts of tropical cyclones in the North Atlantic between 2008 and 2015. Their study provided a comprehensive analysis of the errors of the GEFSv10 by finding track errors, decomposing those errors into along- and cross-track components, and finding what biases may exist in the forecasts. They found that the GEFS was underdispersed in both the along- and cross-track directions. The ECMWF was overdispersed in the cross-track direction. All the ensemble models they examined, including the GEFS, exhibited a slow bias (along track) but little bias in the across track direction.

Prior to the upgrade from version 10 to version 11, NCEP did extensive parallel model runs, to assess the impacts on various forecasting operations. For tropical cyclone forecasting, this provided five years of parallel runs, covering most storms from the hurricane seasons of 2011–15, although the models were generally run from initializations twice a day, rather than the operational four times per day. [Zhou et al. \(2017\)](#) is an overview of the effects of the upgrade from the GEFSv10 to the GEFSv11. They note that the impact on tropical cyclone track errors in the North Atlantic is positive but small at lead times less than 120 h, but negative at longer lead times. There is no discussion of changes to the spread in the ensemble member tracks. [Figure 1](#) shows the results of an independent analysis of the impact the upgrade to the GEFSv11 had on track errors for the hurricane seasons of 2012–15 in the North Atlantic ([Colby 2016](#)). Except for the 72-h forecasts, which showed an improvement of almost 9%, the reductions in the track errors for the ensemble mean are a few percent of NHC's average track error over the previous five years, which is the benchmark NHC compares with model forecast errors.

FIG. 1. Forecast track error of the ensemble mean forecasts as a function of forecast hour for parallel runs of the GEFSv10 and GEFSv11 for the 2014 and 2015 hurricane seasons. The NHC average track error for the past 5 years is shown as a reference.

The new data assimilation system is discussed in [Zhou et al. \(2016\)](#). They compare the old and the new assimilation systems in parallel runs of version 9 of the GFS. For tropical cyclones, they chose certain storms from only two hurricane seasons. They found that the impact on tropical cyclone forecasting was small. The spread in the tracks seemed to depend on whether the initial positions of tropical cyclones were relocated to match observations.

Hurricane Edouard was an official tropical cyclone from 11 to 19 September 2014 ([Stewart 2014](#)), reaching hurricane category 3 in strength on the Saffir–Simpson Hurricane Wind Scale ([Schott et al. 2012](#)). Edouard developed from an easterly wave that left the coast of western Africa. The deep convection near the center of the associated area of low pressure became sufficiently organized by 1200 UTC 11 September to designate the storm as a tropical depression. Edouard eventually intensified, reaching a peak intensity of 105 kt ($1 \text{ kt} \approx 0.51 \text{ m s}^{-1}$) at 1200 UTC 16 September. The track of Edouard traced a smooth, c-shaped curve, as shown in [Fig. 2](#). As Edouard turned eastward, the storm weakened gradually, but was still strong enough to be classified as a tropical storm until 1800 UTC 19 September. Edouard remained over the open Atlantic Ocean for its entire lifetime making it an excellent storm to examine in detail, since there are no interactions with land to consider. It was also the storm for which the most parallel runs were available.

Hurricane Edouard has been the subject of several research projects, some of which involve the use of ensemble modeling. Edouard was one of the storms observed during NASA's Hurricane and Severe Storm Sentinel (HS3; [NASA 2016](#)) field project involving special periods

FIG. 2. Map showing 6-hourly locations and 10-m wind speed (kt) of Hurricane Edouard, starting 0000 UTC 9 Sep 2014 and ending 0000 UTC 19 Sep 2014. Locations taken from National Hurricane Center files. This is also the area used for basinwide calculations of u and v wind component statistics.

of dropsonde deployment. This dropsonde data allowed for detailed analyses of the formation and subsequent dissipation of a secondary eyewall (Abarca et al. 2016). Munsell et al. (2017) discuss the use of a 60-member Weather Research and Forecasting ensemble to understand the dynamics and predictability of Edouard, and in Munsell et al. (2018) the same ensemble system is used to examine the inner-core temperature structure of the storm. Finally, in Melhauser et al. (2017), multimodel and multiphysics ensembles are used to look at the sensitivity of the simulation of Edouard to the use of different models, and in one case, the use of different physical parameterizations in the same model. In both ensemble experiments, the ensemble mean changed, as did the spread in the ensemble members as the model itself changed.

None of this previous work addresses the effects of the changes in the GEFS modeling system. Because the GEFS is the NCEP global ensemble and is used by forecasters at the NHC and elsewhere, the effects of the changes in the GEFS from version 10 to version 11 with the new data assimilation system need investigating. Section 2 details the data and the analysis methodology. Section 3 discusses the results of the analysis, and section 4 has concluding remarks.

2. Methodology

For this study, data from parallel runs of the GEFSv11 and the GEFSv11da were obtained from professional contacts at NCEP. These data were primarily composed of initial and forecast sea level pressure, maximum 10-m

wind, and storm location for the ensemble mean and each ensemble member. Data from GEFSv10 were obtained from the NHC online archives. Full four-dimensional data files from all three model versions were only available for four model runs for Hurricane Edouard (2014).

The data analysis is performed in three groups. First, storm-to-storm statistics for the 2014–15 seasons are compared for the parallel runs of the GEFSv10 and the GEFSv11. Second, the overall statistics for the 2014–15 seasons for the GEFSv10 and the GEFSv11 are compared with the same seasonal statistics for the 2016–17 hurricane seasons when the GEFSv11da was the operational ensemble model. Tables 1–3 are a list of the storms and the number of forecasts from each storm that were used in compiling these statistics. Third, the four

TABLE 1. Storms from the 2014 hurricane season that were used for the seasonal comparisons and the storm-to-storm comparisons. The average track error is the track error for the 120-h ensemble mean forecast if available (1 nm = 1.852 km). For those forecasts that did not reach 120 h, the number of forecasts to the forecast hour in parentheses is shown.

Storm	No. of seasonal forecasts	Avg seasonal track error (nm)	No. of storm-to-storm forecasts
Arthur	25	166	8
Bertha	26	422	8
Cristobal	19	342	9
Edouard	27	115	13
Fay (48 h)	5	187	2
Gonzalo	15	149	6
Hanna	13	147	3

TABLE 2. As in Table 1, except for the 2015 hurricane season.

Storm	No. of seasonal forecasts	Avg track error (nm)	No. of storm-to-storm forecasts
Claudette (36 h)	3	94	2
Danny	11	108	5
Erika (72 h)	4	186	3
Fred	11	227	11
Grace (72 h)	5	87	4
Henri (48 h)	5	92	2
Ida	31	199	19
Joaquin	43	248	41
Kate (96 h)	4	85	1

model runs that were made by all three versions for Hurricane Edouard are compared in detail.

Ensemble modelers often measure the quality of an ensemble system by comparing the spread in the ensemble forecasts to the forecast skill (Zhou et al. 2016). A perfect ensemble system would be initialized with perturbations that fully characterize the uncertainty in the initial data, and the outcome would provide just enough spread to always include the actual evolution of the atmosphere. Too much spread would make it easy to capture the actual atmospheric behavior but could be less useful from a forecasting point of view with too many possible outcomes. For the present study, we choose to quantify the usefulness of an ensemble system from the point of view of a forecaster, who would want the actual track of a tropical storm to fall within the envelope of the ensemble member forecast tracks without an excessive amount of track spread. For each

model run of an ensemble model system, the locations of the storm center in each ensemble member define a latitude/longitude box, based on the largest/smallest latitudes and longitudes among the various members, at each forecast lead time. If we label a given ensemble forecast as successful when the observed storm location falls within this box (hereafter called the track forecast box), the success rate for a series of forecasts will be the percentage of the forecasts in which the actual storm location falls within the track forecast boxes, as a function of forecast lead time. The area of the track forecast box is a measure of the spread of the ensemble member forecasts: larger boxes indicate a larger spread in the tracks. We combine these statistics with the size of the track errors to characterize the seasonal errors. An example of a track forecast box appears in Fig. 3.

As noted above, when the GEFSv11 was tested before implementation, NCEP ran version 11 for many of the storms between 2011 and 2015. The last two years of this period were chosen for this comparison to give sufficient numbers for statistical significance while reducing the influence of improved observations and data assimilation, staying close to the years when only the GEFSv11da was run. As can be seen from Table 4, the storm-to-storm comparison provided many forecasts, even out to 120 h. Since the datasets are for the same model forecasts, Student's *t* test can be used to determine the significance of the results.

The GEFSv11da became operational in 2016 after a very short testing period. Using the last two complete hurricane seasons of 2016 and 2017, track errors, areas of track forecast boxes, and success rates are computed and compared with both the GEFSv10 and the GEFSv11 from

TABLE 3. Storms from the 2016 and 2017 hurricane seasons that were used for the seasonal comparisons. The average track error is the track error for the 120-h ensemble mean forecast averaged over the life of the storm. The number of forecasts is the number of 120-h forecasts. For those forecasts that did not reach 120 h, the number of forecasts for the forecast hour in parentheses is shown.

2016			2017		
Storm	No. of 120-h forecasts	Avg track error (nm)	Storm	No. of 120-h forecasts	Avg track error (nm)
Alex (48 h)	5	36	Arlene (72 h)	4	67
Bonnie	22	103	Bret (72 h)	6	145
Colin (72 h)	4	213	Cindy (96 h)	3	290
Danielle (48 h)	3	48	Don (48 h)	4	144
Earl	12	164	Emily (36 h)	4	42
Fiona	15	385	Franklin (48 h)	3	48
Gaston	34	166	Gert	15	147
Hermine	54	223	Harvey	43	135
Ian	11	387	Irma	39	163
Julia (96 h)	11	270	Jose	56	166
Karl	31	147	Katia (72 h)	7	49
Lisa	12	181	Lee	18	301
Matthew	38	215	Maria	41	120
Nicole	35	233			
Otto	19	127

FIG. 3. Example of forecast track box for the 96-h forecast by the GEFSv11 initialized at 1200 UTC 9 Sep 2014. Filled-in squares are locations of each of the 20-member forecast locations. The black box with the dot in the middle is the actual location of Hurricane Edouard at the same time. This represents a miss for the forecast since the actual location is outside the box.

2014 and 2015. This comparison cannot be done storm to storm since there were very few GEFSv11da model runs for storms in 2014 and 2015, and the GEFSv10 and the GEFSv11 were never rerun for storms in 2016 and 2017. Since the two datasets cannot be considered to have the same size or variance, Welch's t test (Welch 1947) is used to determine the significance of the comparison.

Three-dimensional data for the four forecasts from all three ensemble systems for Hurricane Edouard were obtained from two sources: The Observing System Research and Predictability Experiment (THORPEX) Interactive Grand Global Ensemble (TIGGE) and from personal contacts at NCEP. The operational model in 2014 was the GEFSv10, and these data are archived at TIGGE's website (ECMWF 2014), at 0.5° resolution, at 6-h intervals from 0- to 384-h lead time. This matched the resolution in space and time of the data that were available from the personal contacts at NCEP for the parallel model runs using both the GEFSv11 and the GEFSv11da. The data were analyzed and plotted using Grid Analysis and Display System software (OpenGrADS 2017).

Operational tropical cyclone forecasts are disseminated by the NHC only to 120 h. For this ensemble model evaluation, the 3D data were also analyzed to the 120-h lead times. The lowest 1000-hPa height location for each ensemble member was found using OpenGrADS and a script that uses the gradient of the 1000-hPa height to find the grid point with the lowest value (Fiorino 2009). To evaluate

the winds near Edouard, steering currents were computed as the average wind within 333 km of the storm center (Franklin et al. 1996). Steering currents were computed for four different layer thicknesses, ranging from 1000 to 700 hPa (shallow), from 1000 to 500 hPa (middle), from 1000 to 250 hPa (deep), and from 1000 to 200 hPa (extra-deep). Statistics for the basinwide wind perturbations for the region covering much of the tropical and midlatitude Atlantic Ocean and surrounding land areas (see Fig. 2) were calculated to examine the variability on synoptic scales.

3. Results

Although the official NHC track forecast errors through 120 h shown in figures on the NHC website (NHC 2017a)

TABLE 4. Number of forecasts used for seasonal (2014–17) and storm-to-storm (2014–15) comparisons as a function of forecast hour.

Forecast hour	No. of seasonal forecasts	No. of storm-to-storm forecasts
12	1028	264
24	961	248
36	888	229
48	825	210
72	705	176
96	589	147
120	495	123

FIG. 4. Forecast track error as a function of forecast hour for the GEFS operational ensemble mean forecast. GEFSv11 was run in years 2014–15 while GEFSv11 was run in years 2016–17.

reveal an average trend of improvement over time, especially for the 2016 and 2017 hurricane seasons, Landsea and Cangialosi (2018) suggest that the last four years are characterized by a flattening out of the improvement trends. Model forecasts in general show an ambiguous trend, as seen in figures available at the NHC website (NHC 2017b) especially for the last seven seasons. The data for the recent GEFS operational ensemble mean forecasts appear in Fig. 4. No one year had the smallest average track errors at all forecast hours. For instance, 2017 had the smallest 96–120-h track errors while 2014 had the smallest 72-h track errors. The overall forecast accuracy for the mean GEFS has not changed significantly over this time period.

The areas of the track forecast boxes and the success rates for the last two hurricane seasons, 2014 and 2015, for which data from the GEFSv10 and the GEFSv11 were available for the same set of storms and forecasts are summarized in Fig. 5. The data indicate that the area (Fig. 5a) of the track forecast boxes decreased with the GEFSv11 through lead times of 120 h. The differences are all significant at the 90% or greater level using Student's *t* test. The success rate (Fig. 5b) of the GEFSv11 was up to 8% smaller than that for the GEFSv10 except at the 72-h lead time, and the differences are again significant at the 95% level or higher, except at the 72-h lead time. The conclusion is that the GEFSv11 tracks were less dispersive than those from the GEFSv10, and the envelope of tracks contained the actual storm track less often.

The upgrade from GEFSv11 to GEFSv11da also introduced significant changes to the ensemble spread. The data from the last four hurricane seasons in Fig. 6, comparing the GEFSv11 for seasons 2014–15 with the GEFSv11da for seasons 2016–17, show these changes.

FIG. 5. (a) Area of forecast box for parallel runs of two versions of the GEFS as a function of forecast hour for the 2014 and 2015 hurricane seasons. Numbers in black across the top of the graph show the level of significance of the difference between the areas using Student's *t* test. (b) Success rate for GEFS forecasts as a function of forecast hour. Black numbers along the bottom of the graph show the level of significance of the difference between the two GEFS versions, using Student's *t* test.

While the areas of the GEFSv11da track forecast boxes (Fig. 6a) are almost twice as large as those for GEFSv11 at 12-h lead times and 20% larger at 24-h lead times, the GEFSv11da areas become more than 25% smaller at 48-h lead times. By 96 h and continuing through 120 h, the GEFSv11da areas are 46%–57% the size of those for the GEFSv11, and these differences are significant at the 99% level, using Welch's *t* test (Welch 1947). The spread in the forecast tracks, by this measure, decreased significantly with the addition of the new data assimilation system. This decrease in spread did not help the success rate. As Fig. 6b shows, the percentage of the time the GEFSv11da forecast track envelope included the actual locations decreased to less than 60% for lead times of 96 h and longer. This compares to success rates of 75% or more for the GEFSv11 or the GEFSv10 at those same lead times.

FIG. 6. As in Fig. 5, but for the 2014–15 hurricane seasons for the GEFSv11 and for the 2016–17 hurricane seasons for the GEFSv11da. Significance is measured using Welch's *t* test and shown near the top of the graphs in bold.

The four parallel model runs available for Hurricane Edouard were initialized at 0000 UTC 11, 12, 13, and 14 September 2014. Edouard was a tropical low pressure system at 0000 UTC 11 September, a tropical storm by 0000 UTC 12 September, and first reached hurricane strength by 1200 UTC 14 September. The track shown in Fig. 2 shows the location of the center of Edouard beginning 0000 UTC 11 September and ending 0000 UTC 19 September, 120 h from the final parallel run begun on 14 September. While the four parallel model runs are not enough to provide robust statistics on the effects of the model upgrades, the impact of the upgrades is similar to that found for the seasonal data. The forecast track box areas in Fig. 7a are between 20% and 60% smaller for the first 48 h of the forecasts with the GEFSv11 although up to 20% larger for 72–120-h forecasts. The upgrade to GEFSv11da reduced the forecast track box areas by 35%–70% from those of GEFSv11 and by 50%–80% from GEFSv10.

FIG. 7. (a) Area of track forecast boxes and (b) success rates for the four parallel runs for Hurricane Edouard.

The success rates (Fig. 7b) for GEFSv11 were all smaller than those for GEFSv10 and upgrading to GEFSv11da reduced the success rate even more for these four runs. GEFSv10 was the only model to have a 100% success rate. GEFSv11 had success rates between 50% and 75% through 96-h lead times but dropped to 0% at the 120-h lead time. GEFSv11da had equal or lower success rates than GEFSv11 except at the 120-h lead time, and all were smaller than those for GEFSv10.

There are three mechanisms that can cause variability in the ensemble member forecast tracks: initial location, initial vortex depth, and initial environmental flow. If the vortex is in a different initial location, the resulting steering currents will be different, and the resulting forecast tracks will differ as a result. The depth of the layer that provides the steering current for the vortex depends on the depth of the vortex (Colby 2015; Fovell et al. 2010). If the initialization of the vortex produces circulations with variable depths among the members, the resulting forecast tracks will again differ among the ensemble members. Finally, as intended by the ensemble initialization process, the perturbations in the

TABLE 5. Range and standard deviation (St. Dev.) of initial latitude and longitude of vortices for model runs initialized at 0000 UTC 11 Sep 2014.

Model	Latitude		Longitude	
	Range (°)	St. Dev. (°)	Range (°)	St. Dev. (°)
GEFSv10	1.49	0.53	2.61	0.82
GEFSv11	0.99	0.29	1.99	0.42
GEFSv11da	1.02	0.23	1.76	0.37

initial wind field will produce differing steering currents, even if the initial vortex circulation is the same and located in the same place initially in each of the members.

The initialization of each ensemble member run uses the previous run's 6-h forecast of the storm circulation to initialize the vortex. Once a storm is designated a tropical depression the NHC location and intensity of the vortex is used to relocate the vortex in the global model runs, including the ensemble run. For the model run initialized at 0000 UTC 11 September when Edouard was not yet a depression, this relocation did not take place. Table 5 shows the range and standard deviation of the latitude and longitude of the initial vortices in the three versions. The initial location of the vortex varied in all three ensemble versions. The range in initial latitude was larger in GEFSv10 than that in both GEFSv11 and GEFSv11da, while the range in initial longitude was largest in GEFSv10, more than 20% smaller for GEFSv11 and dropped by almost 12% more for GEFSv11da. Similarly, the standard deviations of the initial latitudes were the largest in GEFSv10, 44% smaller in GEFSv11 and another 20% smaller in GEFSv11da, while the standard deviations of the initial longitudes were largest in GEFSv10, almost 49% smaller in GEFSv11, and 11% smaller still in GEFSv11da. These differences between the versions could account for some of the reduced variability in the forecast tracks for this initialization time.

For the other three parallel model runs, relocation did take place. All of the ensemble members have the vortex in the same initial location, and the depth of the initial circulation does not vary among the members. Any variation in the ensuing forecast tracks is produced by the perturbations in the environmental flow. These perturbations were computed for the basin (the region shown in Fig. 2) at the initial times for all three versions. The mean perturbations themselves (not shown) were small, less than 2.0 m s^{-1} in all three versions. But in each case at multiple levels, as two examples illustrate, Fig. 8a for 850 hPa and Fig. 8b for 250 hPa, the standard deviations of the initial perturbations were largest for GEFSv10, smaller for GEFSv11 and much smaller for GEFSv11da. These differences are all significant at the 99% level using Student's *t* test.

Steering currents, representing the flow in the immediate vicinity of the storm center, for four layers of

FIG. 8. Standard deviation of average basinwide *u* and *v* components as a function of model run initialization at (a) 850 and (b) 250 hPa.

varying thickness (shallow, middle, deep, and extradeep) show the same pattern of variability. The spread in the forecast tracks will be a function of the spread in the steering currents, and the standard deviations of these values quantify the spread. Table 6 shows the percentages of times the standard deviations of the older version of the GEFS were larger than those in the newer version. Of the 12 table entries, three-quarters are above 65%, and all of them are over 50%. Putting all four model initialization times and all model pairs together (four layers with two components for each of the three model pairs at each or

TABLE 6. Percentage of all forecast hours when older version of model had a larger standard deviation of *u* or *v* components of steering currents for layers of four different thicknesses.

Model run	GEFSv11da– GEFSv11	GEFSv11– GEFSv10	GEFSv11da– GEFSv10
11 Sep	59%	73%	77%
12 Sep	79%	67%	84%
13 Sep	51%	88%	90%
14 Sep	82%	55%	89%

FIG. 9. Vector plots of initial ensemble member steering currents valid for the layer from 1000 to 250 hPa for 0000 UTC 12 Sep (a) GEFSv10, (b) GEFSv11, and (c) GEFSv11da model runs. Tick marks are at 1 m s^{-1} intervals. The blue arrow in (a), dark purple arrow in (b), and bold pink arrow in (c) are the mean steering currents.

the four initialization times) gives a total of 96 standard deviation differences. Of these 96 average differences, 85 (89%) showed that the older version of the model had a larger standard deviation than the newer version, and 71 (74%) of these differences were significant at or above the 90% level as measured by Student's t test. Figure 9 shows the initial steering currents for the deep layer (1000–250 hPa) for the 12 September model run, illustrating the difference in variability between the three models. Notice that the initial steering currents for GEFSv10 show significant variability in both zonal and meridional components. This variability becomes smaller in each of the newer model runs. Table 7 shows the standard deviations of the u and v components of the steering currents, averaged over all four layers. Comparing these values with those of the 850 and 250 hPa components over the whole basin, it is clear the basinwide variability is not focused on the region near the tropical cyclone, since the standard deviations of the steering currents are about half the size of the basinwide ones. The pattern of decreasing standard deviations in these initial steering currents shows the same decrease in size with each model upgrade. The vectors for the other three model runs were similar.

4. Conclusions

From the perspective of a forecaster who uses ensemble models to help quantify the uncertainty in a forecast,

one would prefer that the spread in the ensemble member forecasts would include the actual atmospheric evolution. For tropical storm forecasts, a perfect ensemble system would produce an envelope of member forecast tracks that would always contain the actual storm track, while at the same time, providing a measure of the uncertainty in the track forecast. As NCEP's GEFS has evolved since January 2015 when the first of two major upgrades became operational, the spread in the forecast tracks produced by the GEFS has become smaller. As detailed here, parallel model runs showed that the upgrade from the GEFSv10 to the GEFSv11 reduced the area of the track forecast boxes, as well as the success rate.

Analysis of seasonal track errors suggests that model track errors in general have not become smaller over the past four hurricane seasons, and the ensemble mean of the GEFS demonstrates a similar pattern. While a lack of parallel model runs made a direct comparison impossible, differences in track forecast box areas and success rates between the 2014 and 2015 hurricane seasons, and the 2016 and 2017 seasons, showed a further reduction in the track forecast box areas and smaller success rates with the new data assimilation system in GEFSv11da.

Four-dimensional data for the four parallel runs for Hurricane Edouard, one of the few cases of parallel runs using the new data assimilation system, provided the data to examine the details of these differences between

TABLE 7. Standard deviations (m s^{-1}) of u and v components of initial steering currents averaged over four layers of different depth for the four model initializations for each of the three model versions.

Model run	u component			v component		
	GEFSv10	GEFSv11	GEFSv11da	GEFSv10	GEFSv11	GEFSv11da
11 Sep	1.01	0.66	0.57	0.84	0.79	0.67
12 Sep	0.78	0.52	0.47	0.70	0.58	0.43
13 Sep	0.91	0.58	0.55	0.78	0.59	0.47
14 Sep	0.72	0.60	0.59	0.79	0.55	0.52

the three model versions. In every comparison, the same pattern of slightly reduced variability from GEFSv10 to GEFSv11 and larger loss of variability with GEFSv11da is shown. This was true of the average wind perturbations over the Atlantic basin and the steering currents within 333 km of Edouard. These model version differences were statistically significant whenever enough data were available for comparison.

The net result is that with these two model upgrades, the spread in the ensemble members is now significantly smaller than it was prior to the changes. For forecasters using the GEFS to forecast tropical cyclone tracks, the spread in forecast tracks will be less likely to contain the actual track of the storm, and the spread in the forecast tracks may be less representative of the potential uncertainty in the forecast.

Acknowledgments. The staff at NCEP's ensemble modeling section, especially Dr. Yuejian Zhu and Dr. Bing Fu, were tremendously helpful providing datasets for this study. Dr. Andrew Penny from NHC also provided helpful data and information. Scripts written for OpenGrADS by M. Fiorino of the Cooperative Institute for Research in Environmental Sciences made much of the data analysis simpler and faster. The European Center for Medium-Range Weather Forecasts provided operational model output from the GEFS through the TIGGE program. Finally, the three anonymous reviewers deserve much credit for their insightful comments.

REFERENCES

- Abarca, S. F., M. T. Montgomery, S. A. Braun, and J. Dunion, 2016: On the Secondary Eyewall Formation of Hurricane Edouard (2014). *Mon. Wea. Rev.*, **144**, 3321–3331, <https://doi.org/10.1175/MWR-D-15-0421.1>.
- Colby, F. P., Jr., 2015: Global ensemble forecast tracks for tropical storm Debby. *Wea. Forecasting*, **30**, 668–682, <https://doi.org/10.1175/WAF-D-14-00083.1>.
- , 2016: Tropical cyclone track and intensity errors in the 2015 NCEP Global Ensemble Model. *32nd Conf. on Hurricanes and Tropical Meteorology*, San Juan, Puerto Rico, Amer. Meteor. Soc., 18C.3, <https://ams.confex.com/ams/32Hurr/webprogram/Paper293594.html>.
- Congressional Budget Office, 2016: Potential increases in hurricane damage in the United States: Implications for the federal budget. CBO Publ., 33 pp., accessed 22 February 2019, <http://www.cbo.gov/sites/default/files/114th-congress-2015-2016/reports/51518-hurricane-damage.pdf>.
- ECMWF, 2014: TIGGE data retrieval. European Center for Medium-Range Weather Forecasts, accessed 26 May 2017, <http://apps.ecmwf.int/datasets/data/tigge/>.
- Fiorino, M., 2009: Libmf.gex - Mike Fiorino's collection of GrADS extensions. OpenGrADS Project, accessed 11 July 2013, <http://opengrads.org/doc/udxt/libmf/>.
- Fovell, R. G., K. L. Corbosiero, A. Seifert, and K.-N. Liou, 2010: Impact of cloud-radiative processes on hurricane track. *Geophys. Res. Lett.*, **37**, L07808, <https://doi.org/10.1029/2010GL042691>.
- Franklin, J. L., S. E. Feuer, J. Kaplan, and S. D. Aberson, 1996: Tropical cyclone motion and surrounding flow relationships: Searching for beta gyres in Omega dropwindsonde datasets. *Mon. Wea. Rev.*, **124**, 64–84, [https://doi.org/10.1175/1520-0493\(1996\)124<0064:TCMASF>2.0.CO;2](https://doi.org/10.1175/1520-0493(1996)124<0064:TCMASF>2.0.CO;2).
- Landsea, C. W., and J. P. Cangialosi, 2018: Have we reached the limits of predictability for tropical cyclone track forecasting? *Bull. Amer. Meteor. Soc.*, **99**, 2237–2243, <https://doi.org/10.1175/BAMS-D-17-0136.1>.
- Leonardo, N. M., and B. A. Colle, 2017: Verification of multimodel ensemble forecasts of North Atlantic tropical cyclones. *Wea. Forecasting*, **32**, 2083–2101, <https://doi.org/10.1175/WAF-D-17-0058.1>.
- Melhauser, C., F. Zhang, Y. Weng, Y. Jin, H. Jin, and Q. Zhao, 2017: A multiple-model convection-permitting ensemble examination of the probabilistic prediction of tropical cyclones: Hurricanes Sandy (2012) and Edouard (2014). *Wea. Forecasting*, **32**, 665–688, <https://doi.org/10.1175/WAF-D-16-0082.1>.
- Munsell, E. B., F. Zhang, J. A. Sippel, S. A. Braun, and Y. Weng, 2017: Dynamics and predictability of the intensification of Hurricane Edouard (2014). *J. Atmos. Sci.*, **74**, 573–595, <https://doi.org/10.1175/JAS-D-16-0018.1>.
- , —, S. A. Braun, J. A. Sippel, and A. C. Didlake, 2018: The inner-core temperature structure of Hurricane Edouard (2014): Observations and ensemble variability. *Mon. Wea. Rev.*, **146**, 135–155, <https://doi.org/10.1175/MWR-D-17-0095.1>.
- NASA, 2016: NASA Hurricane and Severe Storm Sentinel, HS3. National Aeronautics and Space Administration, accessed 2 March 2019, <https://espo.nasa.gov/hs3>.
- NCEP, 2018: NCEP Environmental Modeling Center (EMC) 3-year implementation plan: FY2018–2020. 81 pp., accessed 22 February 2019, <https://www.weather.gov/media/sti/nggsp/EMC%20Implementation%20Plan%20FY18-20%20v1.docx.pdf>.
- NHC, 2017a: National Hurricane Center forecast verification. Accessed 2 March 2019, https://www.nhc.noaa.gov/verification/figs/ALtkerrtrd_noTD.jpg.
- , 2017b: National Hurricane Center forecast verification. Accessed 2 March 2019, <https://www.nhc.noaa.gov/verification/verify6.shtml?#FIG1>.
- OpenGrADS, 2017: OpenGrADS Project. Accessed 2 March 2019, <http://opengrads.org>.
- Rappaport, E. N., 2014: Fatalities in the United States from Atlantic tropical cyclones: new data and interpretation. *Bull. Amer. Meteor. Soc.*, **95**, 341–346, <https://doi.org/10.1175/BAMS-D-12-00074.1>.
- Schott, T., and Coauthors, 2012: The Saffir-Simpson Hurricane Wind Scale. 4 pp., accessed 22 February 2019, <https://www.nhc.noaa.gov/pdf/sshws.pdf>.
- Stewart, S. R., 2014: Hurricane Edouard (AL062014) 11–19 September 2014. Tropical Cyclone Rep., National Hurricane Center, 19 pp., https://www.nhc.noaa.gov/data/tcr/AL062014_Edouard.pdf.
- Welch, B. L., 1947: The generalization of 'Student's' problem when several different population variances are involved. *Biometrika*, **34**, 28–35, <https://doi.org/10.2307/2332510>.
- Zhou, X., Y. Zhu, D. Hou, and D. Kleist, 2016: A comparison of perturbations from an ensemble transform and an ensemble Kalman filter for the NCEP Global Ensemble Forecast System. *Wea. Forecasting*, **31**, 2057–2074, <https://doi.org/10.1175/WAF-D-16-0109.1>.
- , —, —, Y. Luo, J. Peng, and R. Wobus, 2017: Performance of the new NCEP Global Ensemble Forecast System in a parallel experiment. *Wea. Forecasting*, **32**, 1989–2004, <https://doi.org/10.1175/WAF-D-17-0023.1>.