# NCEP NOTES

## Performance of the New NCEP Global Ensemble Forecast System in a Parallel Experiment

XIAQIONG ZHOU

*I. M. Systems Group, NOAA/NWS/NCEP/EMC, College Park, Maryland*

YUEJIAN ZHU AND DINGCHEN HOU

*NOAA/NWS/NCEP/EMC, College Park, Maryland*

YAN LUO, JIAYI PENG, AND RICHARD WOBUS

*I. M. Systems Group, NOAA/NWS/NCEP/EMC, College Park, Maryland*

(Manuscript received 21 February 2017, in final form 7 July 2017)

ABSTRACT

A new version of the Global Ensemble Forecast System (GEFS, v11) is tested and compared with the operational version (v10) in a 2-yr parallel run. The breeding-based scheme with ensemble transformation and rescaling (ETR) used in the operational GEFS is replaced by the ensemble Kalman filter (EnKF) to generate initial ensemble perturbations. The global medium-range forecast model and the Global Forecast System (GFS) analysis used as the initial conditions are upgraded to the GFS 2015 implementation version. The horizontal resolution of GEFS increases from Eulerian T254 (~52 km) for the first 8 days of the forecast and T190 (~70 km) for the second 8 days to semi-Lagrangian T574 (~34 km) and T382 (~52 km), respectively. The sigma pressure hybrid vertical layers increase from 42 to 64 levels. The verification of geopotential height, temperature, and wind fields at selected levels shows that the new GEFS significantly outperforms the operational GEFS up to days 8–10 except for an increased warm bias over land in the extratropics. It is also found that the parallel system has better reliability in the short-range probability forecasts of precipitation during warm seasons, but no clear improvement in cold seasons. There is a significant degradation of TC track forecasts at days 6–7 during the 2012–14 TC seasons over the Atlantic and eastern Pacific. This degradation is most likely a sampling issue from a low number of TCs during these three TC seasons. The results for an extended verification period (2011–14) and the recent two hurricane seasons (2015 and 2016) are generally positive. The new GEFS became operational at NCEP on 2 December 2015.

### 1. Introduction

The Global Ensemble Forecast System (GEFS) has been one of the most important components of NOAA's environmental prediction operational systems since its implementation in 1993 (Toth and Kalnay 1993, 1997). The forecast skill of the GEFS has been improved significantly since then, benefiting from upgrades in ensemble initial perturbation generation (Wei et al. 2006, 2008), the inclusion of stochastic model perturbations (Hou et al. 2006), higher model resolution, and larger ensemble size, as well as continuous improvement of the Global Forecast System (GFS) model (Han and Pan 2011; Juang 2011, 2014; Yang et al. 2006, 2008) and the Global Data Assimilation System (GDAS; Wu et al. 2002; Kleist et al. 2009a,b; Wang et al. 2013; Kleist and Ide 2015).

GEFS has a long history of using the breeding scheme to generate ensemble perturbations accounting for the uncertainty of initial conditions for medium-range ensemble prediction (Toth and Kalnay 1993, 1997; Kalnay 2001, chapter 6). The basic idea of the breeding method is to simulate the development of growing errors in the

*Corresponding author*: Xiaqiong Zhou, xiaqiong.zhou@noaa.gov

DOI: 10.1175/WAF-D-17-0023.1

analysis cycle. The ensemble perturbations derived from the difference of the short-range forecasts between the ensemble member and ensemble mean evolve with the time-dependent analysis fields with periodic rescaling as the breeding cycle moving forward. After a few days of cycling, the perturbations are expected to sample fast-growing errors in the analysis that are primarily responsible for forecast error growth.

In the GEFS implementation of 2005, Wei et al. (2006, 2008) introduced an ensemble transformation with rescaling (ETR) technique into the breeding method. Forecast perturbations from the breeding cycles are multiplied by a transformation matrix in order to make analysis perturbations globally orthogonal and consistent. The method is subject to limitations due to the relatively small size of the ensemble and the analysis error covariance needing to be provided by the user. It is expected that the ensemble-based perturbations with ETR span more directions than breeding vectors (BVs). Verification shows that ETR outperforms BV in terms of various probability forecast scores (Wei et al. 2008).

Another major source of model forecast uncertainty is from the model itself. Such uncertainty is rooted in the limitations in the computation representation of the equations of motion of the atmosphere, such as the use of finite-resolution, unresolved subgrid parameterizations. The stochastic total tendency perturbation (STTP) scheme (Hou et al. 2006) was used to represent model uncertainty in GEFS, in which stochastic forcing is added every 6 h to the total tendencies of the model variables [temperature, specific humidity, and winds; Hou et al. (2006, 2008)]. First, the temporal change of the total tendency for each ensemble member and the control is calculated within a 6-h time interval. Then, the differences in the temporal change between each ensemble member and the control are used to perturb the total tendency after a multiplication by a random number and the application of an additional rescaling factor. The scaling factor is a function of location and lead time. Generally, the extratropics have larger perturbations than the tropics, and the perturbations grow with lead time. The inclusion of the model uncertainty increases the ensemble spread, reduces the RMSE, and improves the probability forecast skills (Hou et al. 2006). It was implemented in GEFS in 2012.

One of the major operational upgrades to the NCEP GDAS system at NCEP in May 2012 was the implementation of a hybrid three-dimensional variational (3DVAR)–ensemble data assimilation system. The background error covariance in the variational system is enhanced with the inclusion of an ensemble-derived component (75% weight) in addition to the standard static component (25% weight). Initial perturbations for generating the ensemble used to derive the flow-dependent error covariance information are created by the introduction of a second, independent data assimilation system: the ensemble Kalman filter (EnKF, Whitaker and Hamill 2002; Whitaker et al. 2008; Wang et al. 2013, Kleist and Ide 2015; Wu et al. 2002; Kleist et al. 2009b). A dual-resolution strategy with 3DVAR and EnKF using different horizontal resolutions was applied in the hybrid system to reduce computational costs. The implementation of the hybrid system substantially improved the quality of the analysis (Kleist and Ide 2015). The success of EnKF in the NCEP GDAS provides an alternative source of ensemble initial conditions for the operational GEFS as of the 2015 version.

This study aims at a comprehensive comparison between the operational GEFS and a new GEFS version. The main upgrades of the GEFS include the use of EnKF to generate initial ensemble perturbations, a new version of the NCEP Global Spectral Model (GSM), and increased horizontal and vertical resolutions. Section 2 presents a brief introduction to the upgrades in the new GEFS. Section 3 includes the general verification, the precipitation verification, and the verification of tropical cyclone track forecasts. The last section includes our conclusions and some discussion.

## 2. GEFS upgrade

The 2012 implementation of GEFS (v10, hereafter the operational system) was the operational version when this comparison was performed. GEFS v10 has 20 ensemble members and one control run. The model uses a Eulerian horizontal resolution of T254 (~52 km) for the first 8 days of the forecast and T190 (~70 km) for the second 8 days. The new GEFS version (v11, hereafter the parallel system) has the same ensemble sizes but uses a semi-Lagrangian model with a linear Gaussian grid and a resolution of T574 (~34 km) for the first 8 days and T384 (~52 km) for the second 8 days. The number of sigma pressure hybrid vertical layers is increased from 42 to 64. The increased resolution is chosen to fit the wall-clock window in the NCEP operational environment (about 60 min). With the increase in model resolution, the parallel system provides 0.5° gridded binary (GRIB2) files at 3-h time intervals for the first 8 days. Table 1 summarizes the major updates in the parallel system.

The global forecast model in the new GEFS uses the NCEP GFS/GSM version 12.0.0. (see the detail of the updates online at http://www.emc.ncep.noaa.gov/GFS/impl.php). The model configuration for the GEFS with the low-resolution ensemble forecast (T574) follows the settings of the high-resolution deterministic GFS

TABLE 1. The GEFS configuration in the operational (PROD) and parallel (PARA) runs.

| | GEFS v10 (PROD) | GEFS v11 (PARA) |
|---|---|---|
| GFS model | Eulerian, 2012 | Semi-Lagrangian, 2015 |
| Initial perturbation | ETR | EnKF |
| Resolution 0–192 h | T254 (52 km) L42 (hybrid) | TL574 (34 km) L64 (hybrid) |
| Resolution 192–384 h | T190 (70 km) L42 (hybrid) | TL382 (52 km) L64 (hybrid) |
| Output resolution | 1° × 1° | 0.5° × 0.5° for 0–8 days; 1° × 1° for 8–16 days |
| Output frequency | 6 h | 3 h for 0–8 days; 6 h for 8–16 days |

($T1534_{SL}$ L64) implemented on 14 January 2015 with the exception of some resolution-dependent parameters such as the convective gravity wave drag parameters and the critical relative humidity for the formation of partial cloudiness. These parameters are tuned based on the settings in the previous deterministic GFS operational version because of its comparable horizontal resolution (Eulerian T382). The major upgrade of the GSM is in the replacement of Eulerian dynamics with two time-level semi-implicit semi-Lagrangian dynamics (Sela 2010). The semi-Lagrangian GSM can use a larger time step without losing accuracy (Ritchie et al. 1995; Juang and Hong 2010). The time step applied for the semi-Lagrangian GFS at T574 with an equivalent horizontal resolution of 34 km is 900 s, which is 3 times longer than the time step for the operational Eulerian T254 (~52 km).

The hybrid GFS analysis used as the initial conditions for the GEFS control is also updated with the GFS 2015 implementation version. ETR is replaced by EnKF to generate the initial perturbations for ensemble members. The initial conditions for the ensemble members in the parallel experiment are generated by adding the 6-h EnKF forecast ensemble perturbations to the analysis. Note that the 6-h EnKF ensemble forecast perturbations are used instead of the EnKF analysis perturbations since EnKF is run as part of the late analysis (GDAS) cycle rather than the early analysis (GFS) cycle. Only EnKF forecasts from the previous cycle are available when GEFS starts in the NCEP operational environment.

Zhou et al. (2016) compared the GEFS performance of the initial perturbation generation schemes ETR and EnKF for the GEFS 2012 implementation version. It was found that EnKF is comparable with ETR except for a slight degradation in the Southern Hemisphere as a result of too much spread. The large spread in EnKF is generally favorable during data assimilation to avoid filter divergence but is not favorable for the medium-range weather forecast. Two inflation methods, ensemble covariance inflation (e.g., Whitaker and Hamill 2002, 2012) and additive noise inflation (e.g., Whitaker et al. 2008; Houtekamer et al. 2005, 2009), are applied to the posterior ensemble perturbations to account for errors from other sources (e.g., the GSM).

In the 2015 EnKF implementation, new stochastic physics schemes were employed to represent model error to replace the artificial additive inflation. The upgraded stochastic physics suite has three components, including 1) stochastically perturbed physics tendencies (SPPTs; Buizza et al. 1999; Palmer 1997, 2001), 2) stochastically perturbed planetary boundary layer humidity (SHUM), and 3) stochastic kinetic energy backscatter (SKEB; Berner et al. 2009; Shutts 2005). All three of these schemes use an AR(1) random pattern generator to produce spatially and temporally correlated perturbations with three different horizontal length/time scales 500 km/0.25 days, 1000 km/3 days and 2000 km/30 days.

Figure 1 shows the perturbation amplitude in the 2012 and 2015 EnKF versions and the operational GEFS ETR. Figure 1b shows that the inflated EnKF perturbations dampen quickly in the 6-h forecast in the 2012 implementation. The amplitude of EnKF perturbations remains larger than the perturbations in ETR. Figure 1a indicates that the amplitude of the EnKF perturbations from the 2015 implementation becomes similar to ETR. The amplitude of the EnKF perturbations using SPPT increases during the 6-h forecast, although the growth rate is rather small. The EnKF perturbation amplitude from the 2015 implementation is smaller than that from the 2012 version, especially in the upper troposphere. The smaller perturbations in the 2015 EnKF implementation are expected to result in an improved spread–error relationship in GEFS.

The ensemble size is the same in both implementations (20 ensemble members and one control run). Both EnKF and ETR have 80 members for each cycle. Four groups with 20 members each are selected in turn to initialize four ensemble forecast cycles (0000, 0600, 1200, and 1800 UTC) so that every member is used once per day.

In both the 2012 operational and the 2015 parallel GEFS, tropical cyclones (TCs) are separated from the environment and independently perturbed (Kurihara

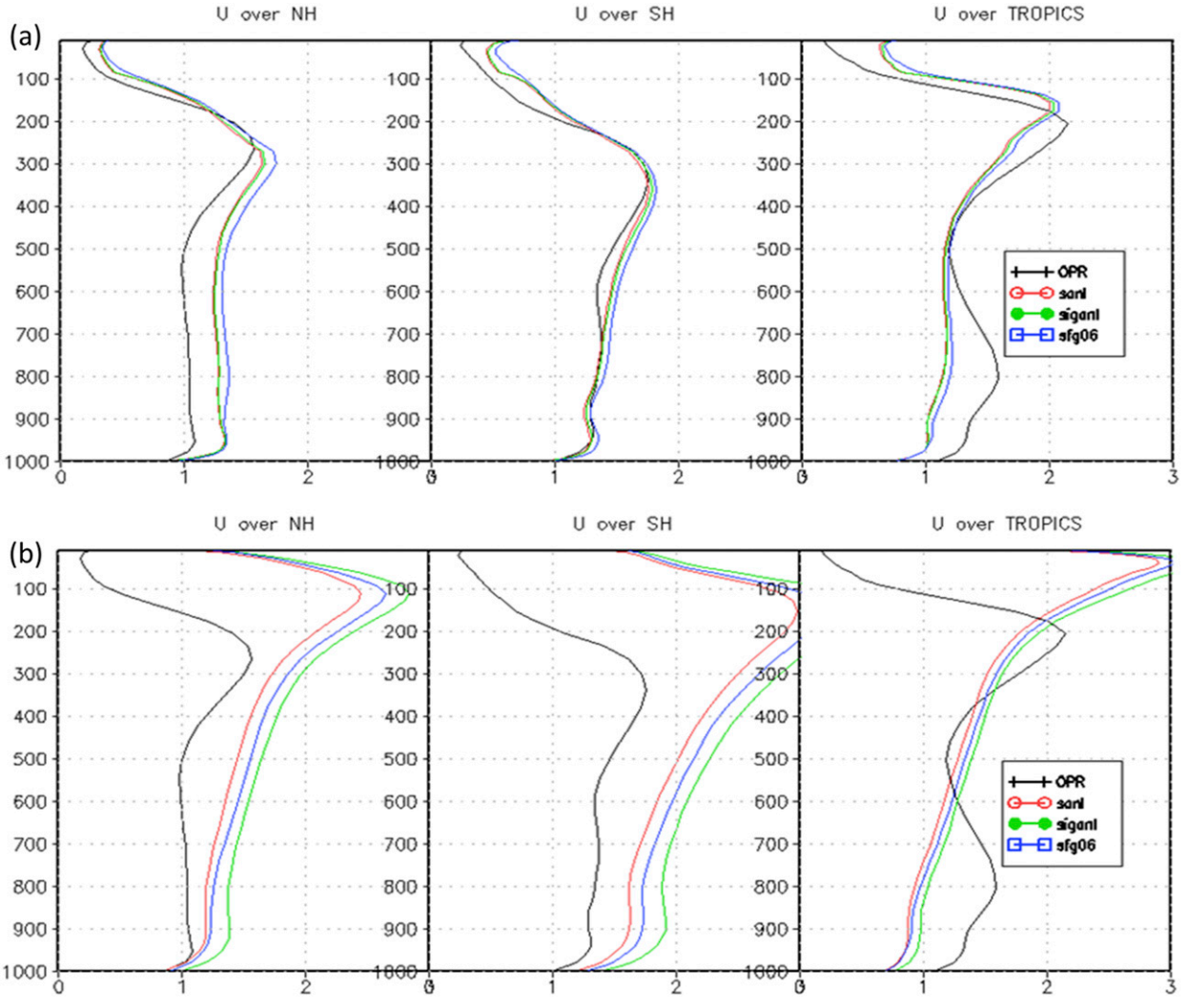


FIG. 1. The vertical profiles of the ensemble perturbation spread for the horizontal wind components averaged over the NH, SH, and tropics in the (a) 2015 and (b) 2012 EnKF implementations. The ensemble spread is calculated based on 20 ensemble members used for GEFS instead of 80 EnKF ensemble members. The black line corresponds to the ETR perturbations from the operational GEFS. Three EnKF profiles represent the spread of EnKF.

et al. 1993, 1995). The TC perturbations are added to the analysis after each ensemble member TC vortex is relocated to the observed location. The TC perturbation adjustments $P$ to the initial state of each ensemble member are calculated by using the following formula (Liu et al. 2000, 2006):

$$P = C \times (X - X_c) \times \|X_c\| / \|X - X_c\|,$$

where $X$ represents the model variables of the TC component of each ensemble member, such as the wind, temperature, mixing ratio, or sea level pressure; $X_c$ corresponds to the same variable for the control; and $||X||$ is the square root of the sum of $X$ over the whole hurricane area. The TC perturbations are calculated from the difference in TC components between the ensemble member and the control forecasts. A scaling factor $C$ is artificially set to 0.05 to reduce the perturbation amplitude, which is about 5% of the magnitude of the TC component in the control forecasts.

The parameters in the 2015 parallel implementation STTP scheme are slightly tuned, which is mainly a result of upgrades to the perturbation scheme and the increased model resolution. The 2015 configuration removes perturbations to the surface pressure tendency to avoid numerical instability. Around the time of model truncation (192 h), the perturbation amplitude of the remaining model state variables is increased to improve the spread–error relationship for the 8–16-day forecast.

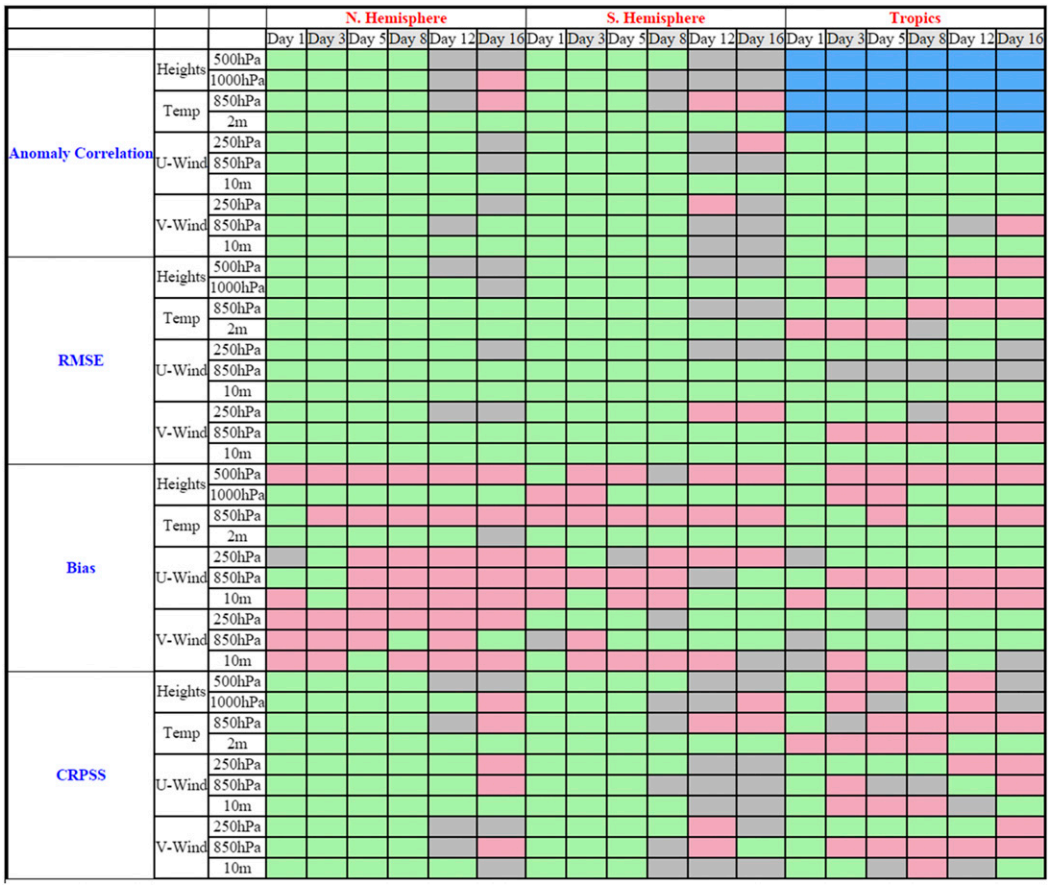


FIG. 2. Scorecard comparing the parallel and operational systems, verified against their respective analyses. Green indicates the parallel system is significantly better and red is significantly worse than the operational system. Gray means no significant difference. Blue means that the corresponding scores are not calculated.

## 3. Verification

### *a. General verification*

A 2-yr parallel run (June 2013–May 2015, 0000 UTC cycle only) was performed and compared with the corresponding operational forecasts. All of the 20-member ensemble forecast data are interpolated to a 2.5° × 2.5° latitude–longitude grid. The ensemble forecast is verified against its own analysis using the NCEP ensemble verification package (Zhu et al. 1996; Toth et al. 2003, 2006; Zhu 2005; Zhu and Toth 2008). One output of the verification package, the scorecard, summarizes the performance of the forecasts of geopotential height at 500 and 1000 hPa; wind fields at 10 m, 850 hPa, and 250 hPa; and temperature at 2 m and 850 hPa in the Northern Hemisphere (NH), the Southern Hemisphere (SH), and the tropics (Fig. 2). The root-mean-square error (RMSE), pattern anomaly correlation (PAC), ensemble mean forecast bias, and the continuous rank probability score skill (CRPSS) are used to compare the two systems. A block bootstrap algorithm (Hamill 1999) is used to test the statistical significance of the differences. Additional verifications can be found online (http://www.emc.ncep.noaa.gov/gc_wmb/xzhou/Para_2013-2015_test.HTML).

The scorecard shows that the parallel system generally outperforms the operational system up to day 12 (8) in the NH (SH) (Fig. 2). The ensemble mean forecasts of all verified variables are more accurate with significantly higher CRPSSs in both the NH and SH. These improvements are generally statistically significant at the 95% confidence level. Degradation is seen in the probabilistic scores at lead times longer than 12 days, but this decrease is considered negligible since the probability forecast skills are already very low at these lead times.

The ensemble mean forecast for wind components over the tropical region is improved at all lead times with respect to PAC. However, there is no clear evidence of systematic improvement in the probability scores of other variables. Generally, the CRPSS is degraded beyond day 3 in the tropics.

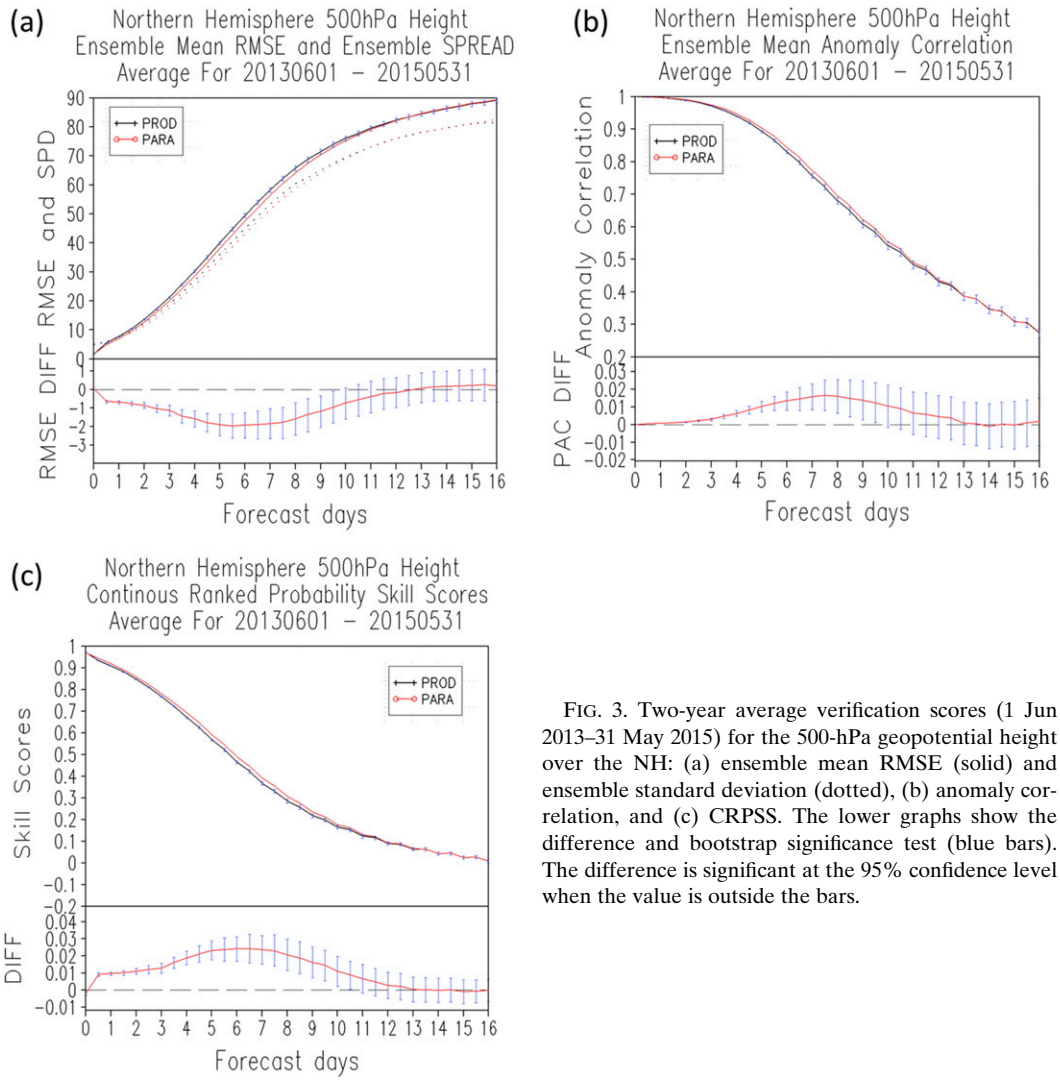


FIG. 3. Two-year average verification scores (1 Jun 2013–31 May 2015) for the 500-hPa geopotential height over the NH: (a) ensemble mean RMSE (solid) and ensemble standard deviation (dotted), (b) anomaly correlation, and (c) CRPSS. The lower graphs show the difference and bootstrap significance test (blue bars). The difference is significant at the 95% confidence level when the value is outside the bars.

Evaluation of the 500-hPa geopotential height field shows that underdispersion is common in both systems, especially during week 2 (Fig. 3a). The updated GEFS has slightly smaller spread than the operational system in the first week, but is similar in the second week. The RMSE of 500-hPa geopotential height is significantly smaller in the NH prior to day 9.

The anomaly correlation of 500-hPa forecasts measures the overall performance of the ensemble mean in capturing large-scale weather patterns. A threshold value of 0.6 is regarded as an indication that the locations of troughs and ridges at 500 hPa are well predicted. The new GEFS extended the skillful forecast from 9 to 9.5 days in terms of PAC (Fig. 3b). Another threshold examined is one of the headline scores in ECMWF's Strategy 2011–2020 (https://www.ecmwf.int/en/forecasts/quality-our-forecasts), which is defined as a 500-hPa geopotential anomaly correlation of 0.8. The forecast lead time for this threshold increases from 8.4 to 8.7 days in the new GEFS. Similar improvement can be found in CRPSS (Fig. 3c).

The new GEFS significantly improves upon the operational version for the first 10 days with respect to the 850-hPa temperatures in the NH. Another ECMWF headline score is the forecast lead time when the CRPSS for ensemble probabilistic forecasts of 850-hPa temperature is greater than 25% for the NH. This score remains at 8.8 days with a slight increase in the new GEFS (Fig. 4).

A major concern is the larger surface temperature bias error in the parallel system. The increase in bias is consistent with the performance of the high-resolution GFS deterministic forecast (Yang 2015), which has a warm and dry bias over some land areas. Figure 5 shows the absolute error and bias of 2-m temperature for the

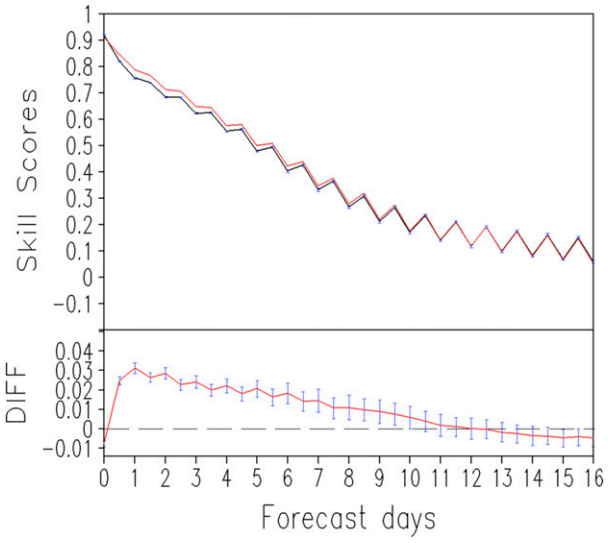


FIG. 4. As in Fig. 3c, but for 850-hPa temperature.

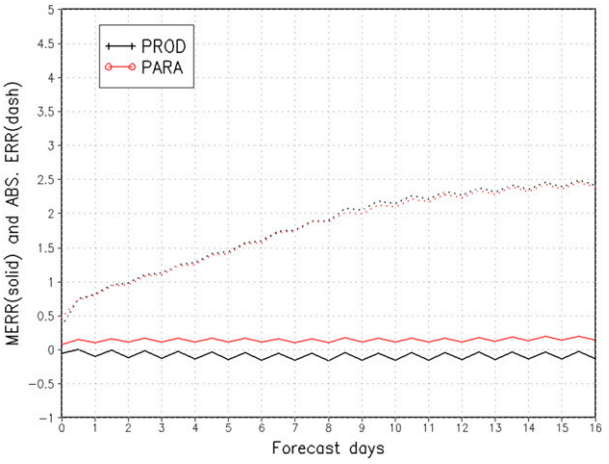


FIG. 5. Ensemble-mean bias (solid line) and absolute error (dashed line) of 2-m temperature over the NH.

two GEFS implementations in the NH. They have the same absolute error, but there is a warmer bias in the new GEFS implementation.

The time series of 2-m temperature bias over North America (NA) shows that the bias varies with season during the 2-yr period (Fig. 6). The large degradation in terms of surface temperature bias is more evident in the summer. The 2012 implementation usually has a warm bias in the summer and a cold bias in the winter. The parallel system reduces the cold bias in the winter, but increases the warm bias in the summer. It is suggested by the GFS development group that the land surface update in the GFS 2015 implementation resulted in lower soil moisture when the Global Land Data Assimilation System (GLDAS)/Coupled Forecasting System (CFS) soil moisture climatology at T574 (~27 km) replaced the 1° bucket soil climatology. Evaporation parameters configured for the drier soil climatology were not retuned as part of the 2015 GFS implementation. This caused increased sensible heat flux and reduced the latent heat flux in hot air masses over cropland. This was corrected in the May 2016 GFS implementation.

This kind of systematic model error can be removed with postprocessing algorithms. Cui et al. (2012) developed a Kalman filter–type algorithm to calculate an online decaying average bias and produce a bias-corrected ensemble. The decaying averaged bias is updated every forecast cycle with the most recent model forecast given a weight of 2%. Thus, the bias contains the accumulated information about the behavior of the ensemble forecasting system in the last 50–60 days. This postprocessing technique is applied operationally at NCEP to both NCEP and Meteorological Service of Canada ensemble forecasts before generating joint products of the North American Ensemble Forecast System (NAEFS). This method is also performed for our parallel experiments to generate bias-corrected ensemble products. Figure 7 shows the bias-corrected ensemble forecasts for NH 2-m temperature averaged over 1 yr. The surface temperature spread is seriously underdispersive in both systems, which is a well-known problem in the field of ensemble forecasting (Buizza et al. 2000). Note that the RMSE of 2-m temperature forecasts decreases significantly in the new GEFS. The horizontal resolution increase is likely responsible for the improved surface temperature prediction as a result of more accurate resolved representations of the surface forcing (i.e., topography, vegetation, land-use fields). Improvements can be seen in both the operational and parallel experiments with bias correction. The RMSE in the bias-corrected parallel forecast is less than in both the bias-corrected operational forecast and the uncorrected forecast.

### *b. Precipitation verification*

Quantitative precipitation forecasts (QPFs) and probabilistic QPFs with uniform 1° horizontal resolution are verified against the climatology-calibrated precipitation analysis (CCPA) over the contiguous United States (CONUS) using both continuous and categorical verification approaches (Hou et al. 2014). Both continuous and categorical verification methods are used to assess the skill of the ensemble-based probability precipitation forecast (http://www.emc.ncep.noaa.gov/gmb/yluo/GEFS_VRFY/GEFS_PARA_SUMMARY.html).

The continuous verification methods include the continuous ranked probability score (CRPS), RMSE/SPREAD, and MERR (mean error)/absolute error.

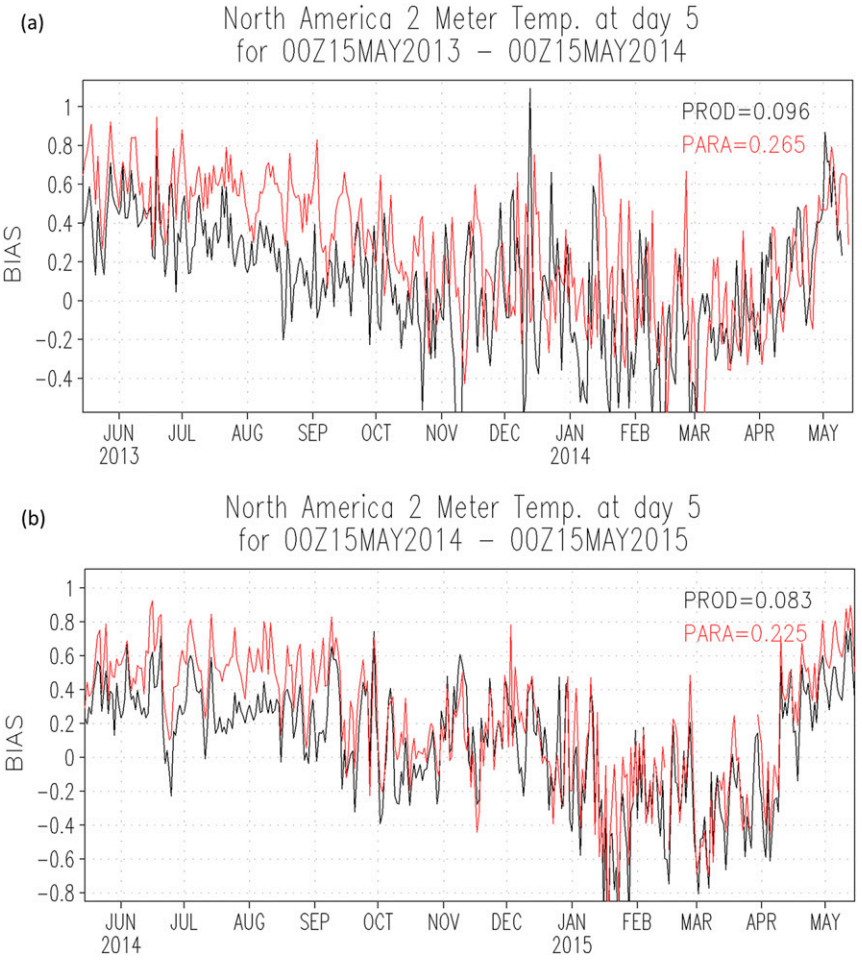


FIG. 6. The time series of 2-m temperature bias over NA (a) from 15 May 2013 to 14 May 2014 and (b) from 15 May 2014 to 14 May 2015.

CRPS measures the area difference between the cumulative distributions of forecasts and observations. For categorical verification, precipitation is categorized by the 24-h accumulated precipitation with threshold amounts greater than 1, 5, 10, and 20 mm. The evaluation methods include the Brier score/Brier skill score (BS/BSS), reliability, bias, the equitable threat score (ETS), and true skill score (TSS). The BS is either 1 or 0 by measuring the mean-square error of a probability forecast from the observed probability depending on whether the categorized event occurred. The BS can be decomposed into three additive components: uncertainty, reliability, and resolution (Murphy 1973). In a perfect forecast, the predicted probabilities should be exactly equal to the observed probabilities. The BSS uses the 10-yr mean of CCPA as the climatology to calibrate the BS and avoids the dependence of BS on the frequency of the event (Fig. 8). A reliability diagram, which displays the observed precipitation probabilities conditioned with the forecast probabilities of all precipitation forecast samples, provides information about probability forecast bias for the GEFS (Fig. 9). ETS, TSS, and bias results are based on the $2 \times 2$ contingency table in which the frequencies of the occurrence of forecasting precipitation greater or less than the thresholds are counted and calibrated with CCPA.

The performance of the precipitation forecast of the parallel system is generally similar to that of the operational system. There is no significant difference between these two systems in terms of CRPS, ETS, and TSS (not shown). The RMSE and ensemble spread of the precipitation are greater in the parallel version. The RMSE results are not presented since they do not objectively measure the complex spatial distribution of precipitation. There are some suggestions that the parallel version has better BS/BSS during the first 3 days (Fig. 8a), less bias during the first week for the light precipitation forecast (not shown), and greater reliability

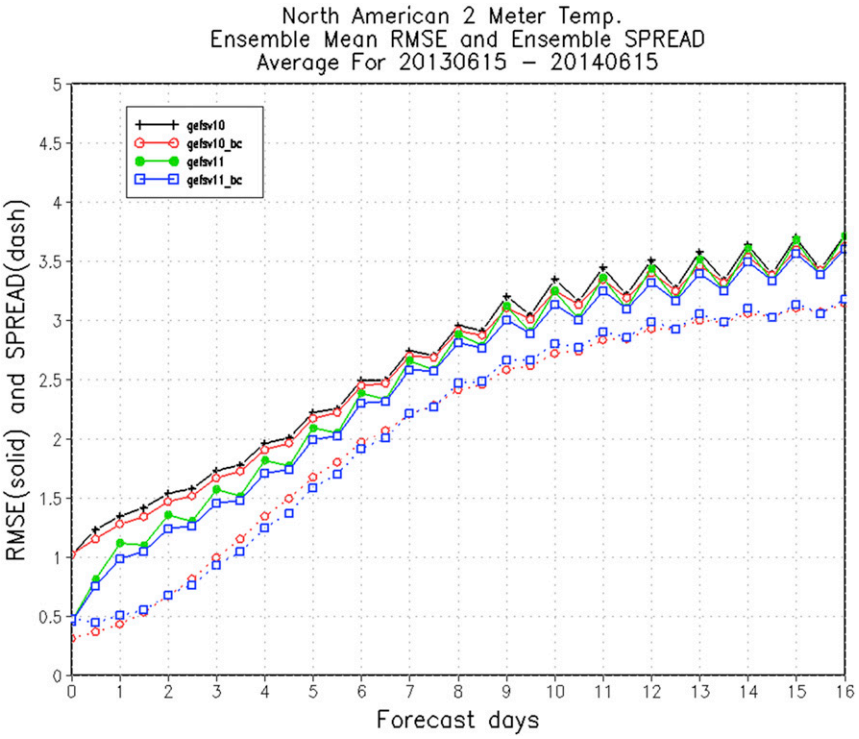


FIG. 7. The RMSE (solid lines) and spread (dashed lines) of raw and bias-corrected 2-m temperature over the NH for the operational (GEFS v10) and parallel runs (GEFS v11).

for short-range forecasts (days 1–3) for precipitation greater than 1 and 5 mm $(24\,\text{h})^{-1}$ (Fig. 9a). The reliability diagram for the parallel version is slightly closer to the diagonal line. Both systems are overconfident for the probability forecast of precipitation greater than 5 mm $(24\,\text{h})^{-1}$. Careful examination shows that the better reliability is only seen during the warm seasons (April–October) for the forecast of precipitation greater than 1 and 5 mm $(24\,\text{h})^{-1}$ at days 1–3, which is also the main contribution to the higher BSS (Figs. 8a,b and 9a,b). The performance of the precipitation ensemble forecast is unchanged during the cold seasons (from November to next March; not shown). The improvement of the precipitation forecast in the early forecast range is probably related to the parallel system resolving smaller-scale features due to an increase of the model horizontal resolution.

### *c. Tropical cyclone track forecasts*

The TC activity in the 2013 and 2014 North Atlantic hurricane seasons was well below average. There were only 8 named TCs in 2014 compared to a climatological mean of 12.1. The total number of TCs in 2013 (14) exceeds the climatology, but only 2 (Hurricanes Humberto and Ingrid) reached hurricane intensity (6.4 in climatology) with no storms reaching category 2 intensity. To increase the sample, the medium-range forecasts for the 2011 and 2012 hurricane seasons (June–October) are included. The cases verified here include tropical depressions and stronger TCs.

Figure 10 shows the ensemble mean forecast errors of TC tracks over the North Atlantic, eastern Pacific, and western North Pacific up to day 7. The mean track errors over the North Atlantic and eastern Pacific are slightly smaller in the parallel version up to day 6 and larger at longer lead times, but the difference is generally insignificant. The ensemble-mean track forecast error is significantly reduced over the western North Pacific; the track error at day 5 is reduced 20% from 250 to 200 nm.

The performance of the TC track forecast varies with season. One concern is the significant degradation of the day-6 and day-7 TC track forecasts during the 2012, 2013, and 2014 hurricane seasons in the North Atlantic (not shown). There are 77 cases for day-6 forecasts, with 61 cases occurring in 2012 and 8 each in 2013 and 2014. Figure 11 shows the displacements of forecasted TC positions for day 6 from their observed locations for the parallel and operational systems. The spread is larger in the forecast TC location deviations from observation in the parallel system with more cases having track error larger than 500 km. The TCs in the parallel system tend

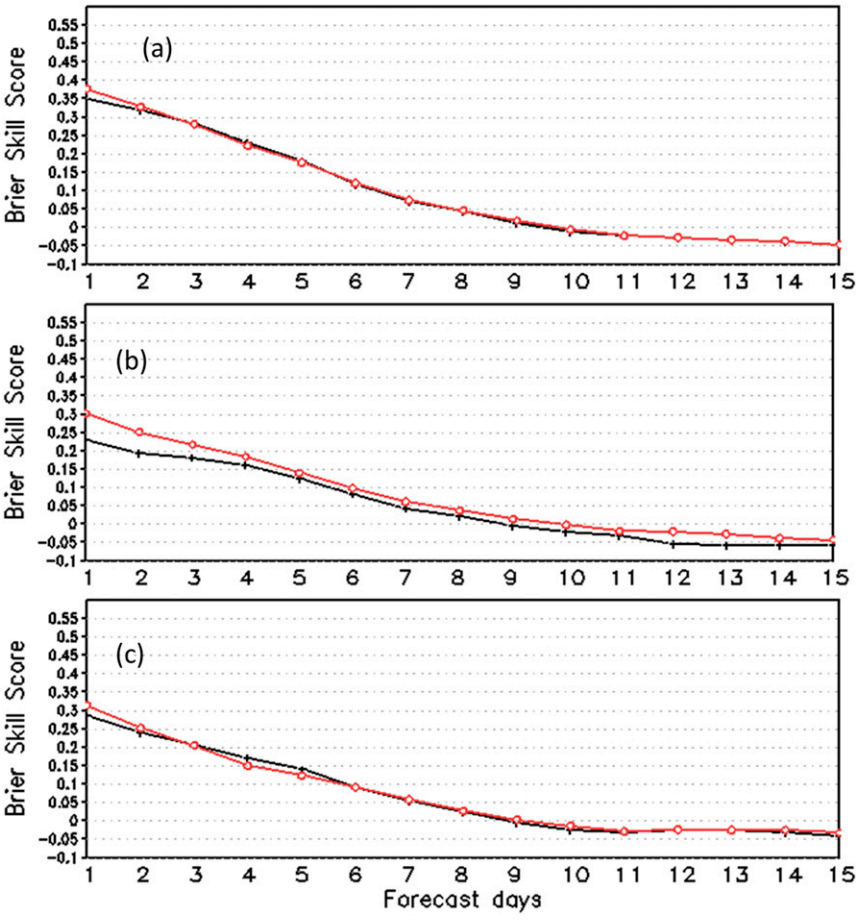


FIG. 8. BSS for the precipitation greater than 5 mm $(24\ h)^{-1}$ averaged over the (a) 2013 warm season (15 May–15 Oct 2013), (b) 2014 warm season (15 May–15 Oct 2014), and (c) 2-yr period (15 May 2013–14 May 2015).

to have larger south and east biases. Note that the degradation of the parallel system came from the forecasts of Hurricanes Nadine (2012), Michael (2012), and Edouard (2014).

As the fourth longest-lived hurricane over the North Atlantic (10 September–4 October 2012), Hurricane Nadine (2012) contributes to 30 of 77 cases of 6-day forecasts. The TC moved northwestward in its early stage and then turned northward and eastward later. Thereafter, Nadine meandered over the ocean, traveling in a clockwise loop and then a counterclockwise loop before transitioning to an extratropical low system. The forecast locations for day-6 forecasts initiated from early in the TCs life cycle were generally located northwest of the observed locations as a result of a slower than observed northeastward turning (Fig. 12a). The eastward deviations resulted from poor forecasts of the TC's unusual movement later in its life cycle. Figure 13a shows that the storm in the parallel system moved eastward instead of looping cyclonically, which contributes to larger track forecast errors. Munsell et al. (2015) suggested that the track forecast during this time period had low predictability, and the track divergence is related to uncertainty in the environmental steering flow associated with the position and strength of a midlatitude trough.

Hurricane Michael (2012) moved northward slowly after it formed from a mid- to upper-level short-wave disturbance. It subsequently zigzagged as a result of a change in the environmental steering flow. The TC in the parallel system moved to the east instead of the north, resulting in large track forecast errors (Fig. 12b).

Hurricane Edouard (2014) was steered by the large-scale flow around a subtropical ridge. The recurvature from northwestward to northeastward was well predicted in both the operational and the parallel systems, except that the parallel system had much slower northward speed resulting in much larger track errors (Fig. 12c).

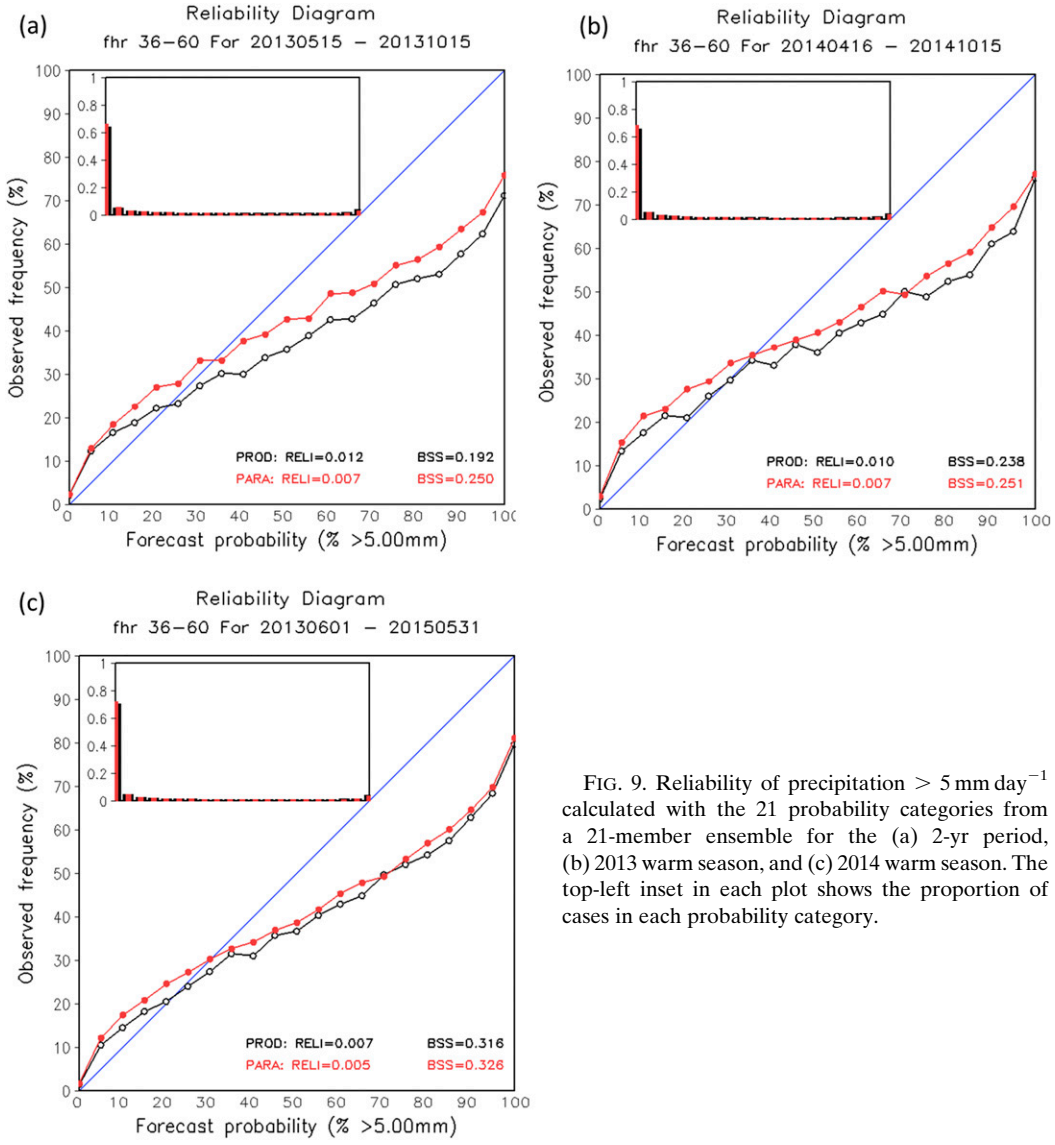


FIG. 9. Reliability of precipitation $> 5\,\text{mm day}^{-1}$ calculated with the 21 probability categories from a 21-member ensemble for the (a) 2-yr period, (b) 2013 warm season, and (c) 2014 warm season. The top-left inset in each plot shows the proportion of cases in each probability category.

The comparison between the operational and parallel systems continued in terms of TC track forecasts when the parallel GEFS was undergoing real-time parallel testing during the 2015 hurricane season and when it became the operational system during the 2016 hurricane season. Figure 13a shows that the ensemble-mean track forecast errors are smaller in the parallel system than in the operational system although the difference is not statistically different. Similar performance is found for the 2016 hurricane season. Figure 13b shows that the track forecasts of Hurricane Matthew (2016), the strongest, costliest, and deadliest storm of the season, are more accurate in the parallel system than the operational system.

## 4. Conclusions and discussion

An upgrade to the Global Ensemble Forecast System implemented at NCEP on 2 December 2015 was introduced and a comprehensive verification study comparing the upgraded system (termed the parallel system in this study) to the older operational version (termed the operational system) was provided. The ETR scheme, an updated breeding scheme implemented in the operational GEFS in 2005, was replaced by the EnKF scheme to generate ensemble initial perturbations in the parallel system. The global forecast model is upgraded to the same version as the high-resolution deterministic GFS implemented on 14 January 2015

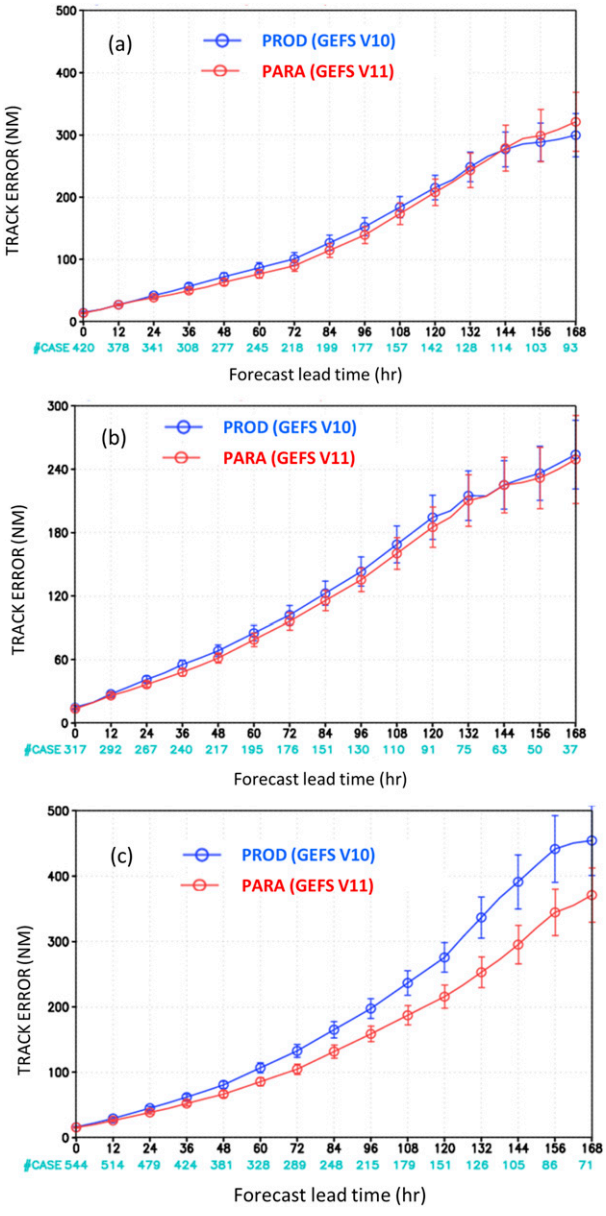


FIG. 10. Ensemble-mean track forecast errors for the 2011–14 hurricane seasons over the (a) North Atlantic, (b) eastern Pacific, and (c) western North Pacific.

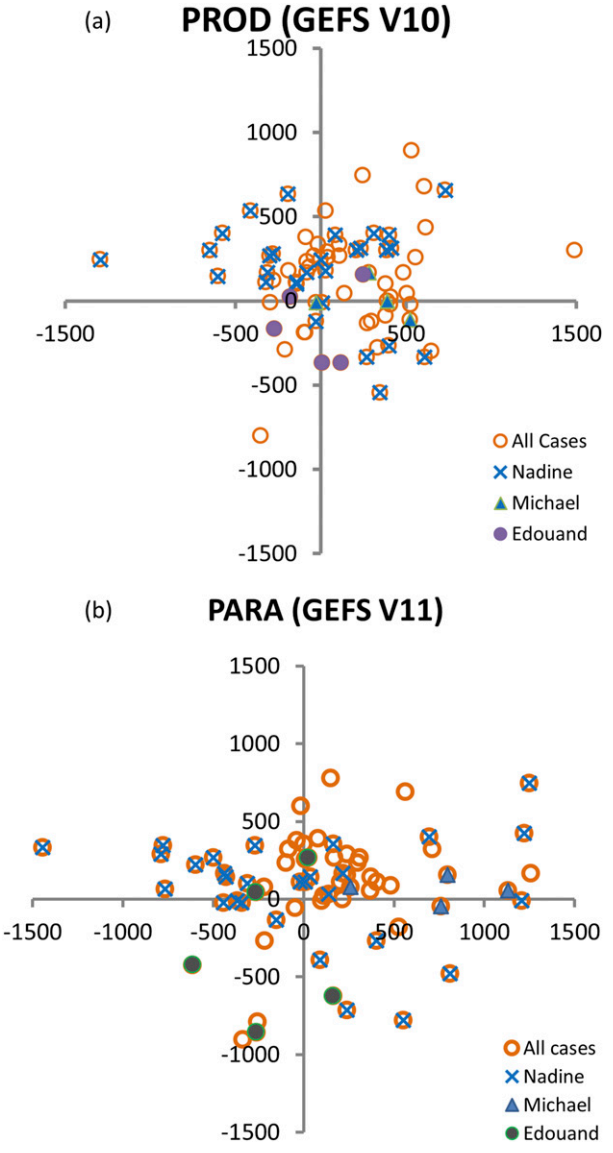


FIG. 11. The deviation of the 6-day forecast TC locations from observation over the North Atlantic (2012–14) for the (a) operational and (b) parallel systems.

with the most notable change from Eulerian to semi-Lagrangian dynamics (Yang 2015). The model resolution in the parallel system increased from Eulerian T254 (~52 km) to semi-Lagrangian T574 (~34 km) and the vertical resolution increased from 42 to 64 levels. Note that the initial conditions in the parallel GEFS use the 2015 implementation version of the GFS analysis.

The parallel system is generally more skillful than the operational system up to days 8–10 over extratropical regions with respect to the ensemble mean and probability forecasts of the model variables including geopotential height, temperature, and wind fields. The improvement is significant at the 95% confidence level as evaluated by a bootstrap test. The parallel system improved on the operational AC of 500-hPa forecasts by extending the skillful forecast from 9 to 9.5 days.

The parallel system has a warm surface temperature bias over the Great Plains in summer. The GFS development group suggested that the land surface update in the 2015 GFS implementation resulted in lower soil moisture as a result of a change in the soil climatology

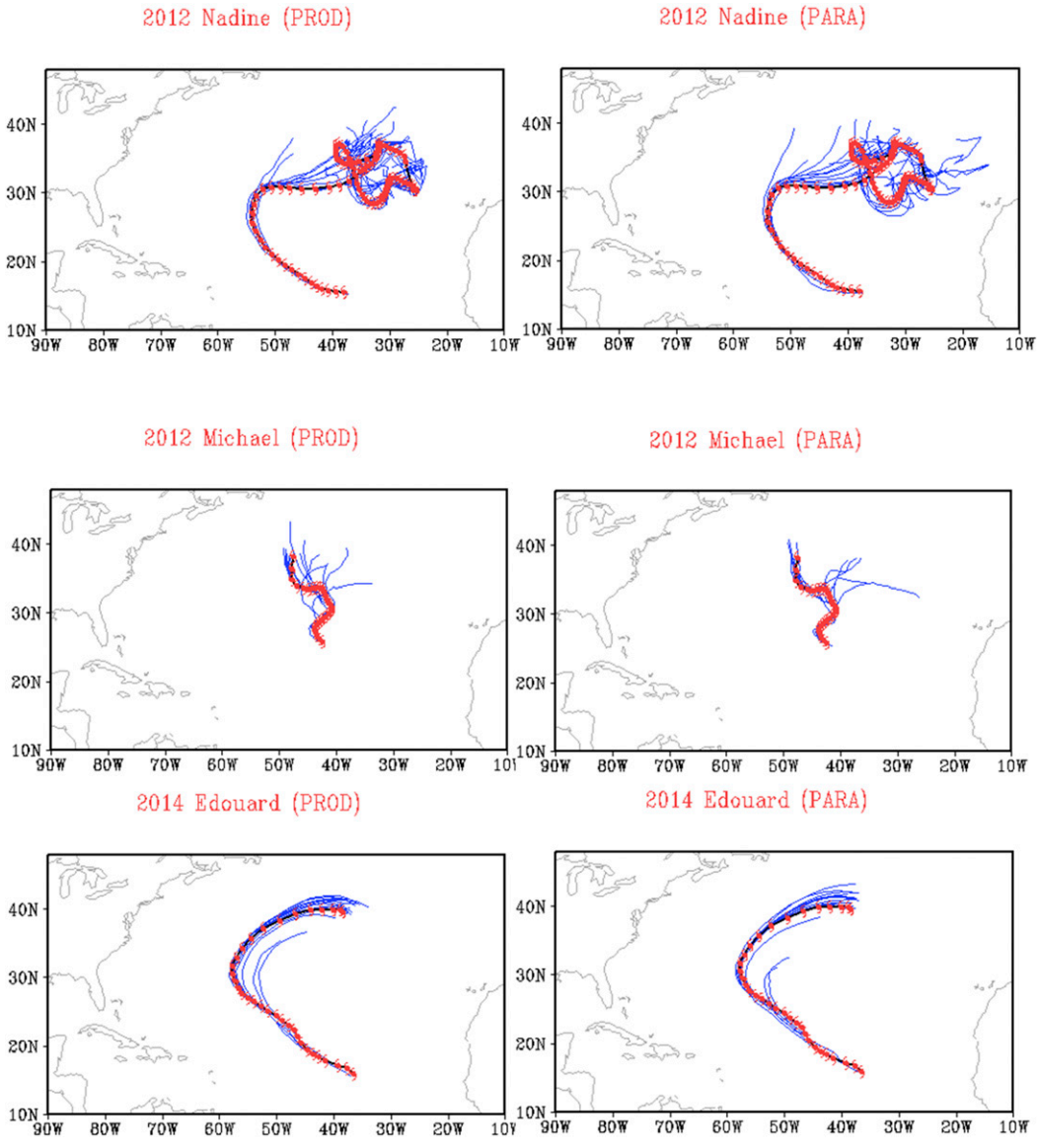


FIG. 12. The observed TC tracks (hurricane signal) and the 168-h ensemble-mean forecast tracks from successive forecasts for Hurricanes (a) Nadine, (b) Michael, and (c) Edouard in the (left) operational and (right) parallel systems.

data. This systematic temperature bias can be corrected by subtracting the decaying averaged bias from the ensemble raw forecasts (Cui et al. 2012).

The parallel system outperformed the operational system in forecasting tropical wind components in the upper and lower layers. The AC scores for the ensemble mean wind fields are significantly better for 16-day forecasts. Nevertheless, there is no clear evidence that the parallel system has a positive impact on the probability forecasts over the tropics.

The probabilistic forecasts of precipitation over the CONUS were evaluated against CCPA using both continuous and categorical verification methods. The performance of the precipitation forecast is similar between the parallel and operational systems, except for higher reliability and BSSs in the short-range forecasts during the summer. The limited improvement in the precipitation forecasts is likely related to the similar physics in these two systems as the major upgrade to the GFS model was in its dynamics with no significant changes to the physical processes. A cumulus convection scheme with scale and aerosol awareness was developed by Han et al. (2017) and will be implemented in May 2017 along with other many updates in GSM (v14). This

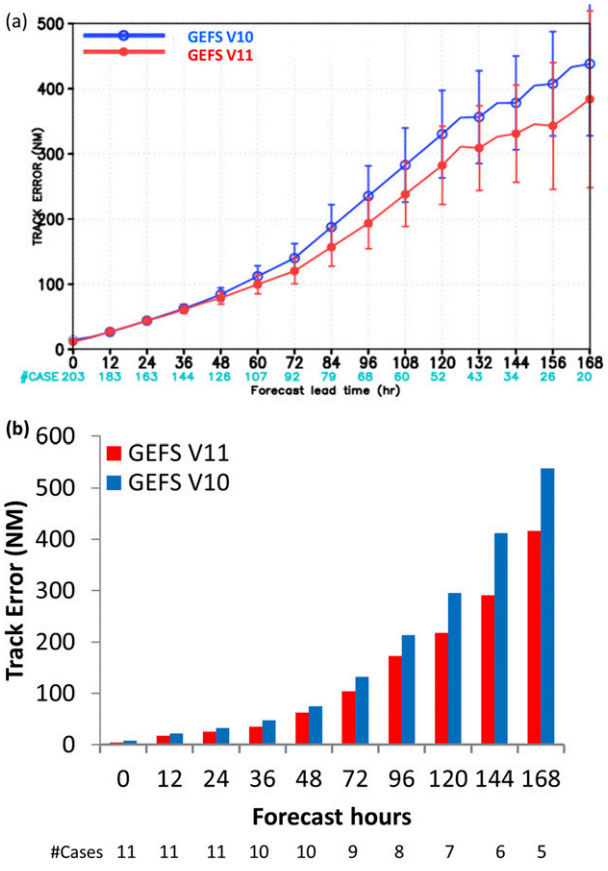


FIG. 13. The ensemble-mean track forecast errors for (a) the 2015 hurricane seasons over the North Atlantic and (b) Hurricane Matthew (2016).

scheme improves precipitation forecasts, especially over the CONUS during the summer.

Based on the statistics of four hurricane seasons (2011–14), it was found that the parallel system improved the TC track forecasts over the western North Pacific significantly, but the improvement over the eastern Pacific and North Atlantic is very limited. In addition, there is a slight degradation in day-6 and -7 track forecasts, especially for 2012, 2013, and 2014. The large track forecast errors are driven by three long-lived hurricanes, including Hurricane Nadine (2012). The degradation of day-6 and -7 track forecasts in the parallel system is consistent with the performance of the deterministic GFS (not shown), although the reason for the degradation is not clear. Further comparison shows that the GEFS TC track forecasts for the 2015 and 2016 Atlantic hurricane seasons are better in the parallel system than in the operational system.

The purpose of the comparison was to evaluate the overall quality of the two ensemble systems rather than the individual system changes. The contribution of the individual upgrades to the main change of GEFS performance is not clear here. The updates of the model and analysis could improve the overall performance of GEFS, but the new GEFS also has the same issues as in the high-resolution GFS deterministic forecast. For example, both GEFS and GFS show a warm surface bias over the CONUS. This systematic bias in the operational GFS was noticed in the surface temperature forecast over the Great Plains, and a fix was proposed and implemented in the 2016 GFS.

The replacement of ETR with EnKF is consistent with the plan to develop a unified NOAA EMC system. However, the performance of operational GEFS may have inconsistency issues as a result of the use of the EnKF ensemble to generate the GEFS initial perturbations. Upgrades to the EnKF could automatically affect the performance of the operational GEFS, especially when the model perturbations generated by the model stochastic perturbation scheme (STTP) are sensitive to the amplitude of the initial perturbations (Zhou et al. 2016). Three alternate stochastic schemes, SKEB, SPPT, and SHUM, were implemented for the short-term EnKF-based forecasts used in the data assimilation cycling. The impact of replacing the operational STTP scheme with a combination of these three alternate stochastic schemes for use in the GEFS medium-range forecasts is currently under assessment.

*Acknowledgments.* This work was completed as part of EMC/NCEP/NWS/NOAA regular work duties. We thank Glenn White and Jason Sippel for their careful internal review and Christopher Melhauser for aiding in an editorial review of the manuscript. We appreciate our colleagues Shirinvas Mooorthi, Fanglin Yang, et al. at NCEP/EMC for setting the GFS model configuration.

## REFERENCES

Berner, J., G. J. Shutts, M. Leutbecher, and T. N. Palmer, 2009: A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ECMWF Ensemble Prediction System. *J. Atmos. Sci.*, **66**, 603–626, doi:10.1175/2008JAS2677.1.

Buizza, R., M. Miller, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System. *Quart. J. Roy. Meteor. Soc.*, **125**, 2887–2908, doi:10.1002/qj.49712556006.

——, J. Barkmeijer, T. N. Palmer, and D. S. Richardson, 2000: Current status and future development of the ECMWF Ensemble Prediction System. *Meteor. Appl.*, **7**, 163–175, doi:10.1017/S1350482700001456.

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

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, doi:10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2.

Han, J., and H.-L. Pan, 2011: Revision of convection and vertical diffusion schemes in the NCEP Global Forecast System. *Wea. Forecasting*, **26**, 520–533, doi:10.1175/WAF-D-10-05038.1.

——, W. Wang, Y. C. Kwon, S.-Y. Hong, V. Tallapragada, and F. Yang, 2017: Updates in the NCEP GFS cumulus convection schemes with scale and aerosol awareness. *Wea. Forecasting*, https://doi.org/10.1175/WAF-D-17-0046.1, in press.

Hou, D., Z. Toth, and Y. Zhu, 2006: A stochastic parameterization scheme within NCEP global ensemble forecast system. *18th Conf. on Probability and Statistics in the Atmospheric Sciences*, Atlanta, GA, Amer. Meteor. Soc., 4.5, https://ams.confex.com/ams/Annual2006/techprogram/paper_101401.htm.

——, ——, ——, and W. Yang, 2008: Impact of a stochastic perturbation scheme on NCEP Global Ensemble Forecast System. *19th Conf. on Probability and Statistics in the Atmospheric Sciences*, New Orleans, LA, Amer. Meteor. Soc., 1.1, https://ams.confex.com/ams/88Annual/techprogram/paper_134165.htm.

——, and Coauthors, 2014: Climatology-calibrated precipitation analysis at fine scales: Statistical adjustment of Stage IV toward CPC gauge-based analysis. *J. Hydrometeor.*, **15**, 2542–2557, doi:10.1175/JHM-D-11-065140.1.

Houtekamer, P. L., H. L. Mitchell, G. Pellerin, M. Buehner, M. Charron, L. Spacek, and B. Hansen, 2005: Atmospheric data assimilation with an ensemble Kalman filter: Results with real observations. *Mon. Wea. Rev.*, **133**, 604–620, doi:10.1175/MWR-2864.1.

——, ——, and X. Deng, 2009: Model error representation in an operational ensemble Kalman filter. *Mon. Wea. Rev.*, **137**, 2126–2143, doi:10.1175/2008MWR2737.1.

Juang, H.-M. H., 2011: A multiconserving discretization with enthalpy as a thermodynamic prognostic variable in generalized hybrid vertical coordinates for the NCEP Global Forecast System. *Mon. Wea. Rev.*, **139**, 1583–1607, doi:10.1175/2010MWR3295.1.

——, 2014: A discretization of deep-atmospheric nonhydrostatic dynamics on generalized hybrid vertical coordinates for NCEP global spectral model. NCEP Office Note 477, 39 pp., http://www.lib.ncep.noaa.gov/ncepofficenotes/files/on477.pdf.

——, and S.-Y. Hong, 2010: Forward semi-Lagrangian advection with mass conservation and positive definiteness for falling hydrometeors. *Mon. Wea. Rev.*, **138**, 1778–1791, doi:10.1175/2009MWR3109.1.

Kalnay, E., 2001: *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 368 pp.

Kleist, D. T., and K. Ide, 2015: An OSSE-based evaluation of hybrid variational–ensemble data assimilation for the NCEP GFS. Part I: System description and 3D-hybrid results. *Mon. Wea. Rev.*, **143**, 433–451, doi:10.1175/MWR-D-13-00351.1.

——, D. F. Parrish, J. C. Derber, R. Treadon, R. M. Errico, and R. Yang, 2009a: Improving incremental balance in the GSI 3DVAR analysis system. *Mon. Wea. Rev.*, **137**, 1046–1060, doi:10.1175/2008MWR2623.1.

——, ——, ——, ——, W. S. Wu, and S. Lord, 2009b: Introduction of the GSI into the NCEP Global Data Assimilation System. *Wea. Forecasting*, **24**, 1691–1705, doi:10.1175/2009WAF2222201.1.

Kurihara, Y., M. A. Bender, and R. J. Ross, 1993: An initialization scheme of hurricane models by vortex specification. *Mon. Wea. Rev.*, **121**, 2030–2045, doi:10.1175/1520-0493(1993)121<2030:AISOHM>2.0.CO;2.

——, ——, R. E. Tuleya, and R. J. Ross, 1995: Improvements in the GFDL hurricane prediction system. *Mon. Wea. Rev.*, **123**, 2791–2801, doi:10.1175/1520-0493(1995)123<2791:IITGHP>2.0.CO;2.

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. NOAA Tech. Procedures Bull. 472, 7 pp., http://www.nws.noaa.gov/om/tpb/472.htm.

——, S. J. Lord, N. Surgi, Y. Zhu, R. Wobus, Z. Toth, and T. Marchok, 2006: Hurricane relocation in global ensemble forecast system. *27th Conf. on Hurricanes and Tropical Meteorology*, Monterey, CA, Amer. Meteor. Soc., P5.13, https://ams.confex.com/ams/pdfpapers/108503.pdf.

Munsell, E. B., J. A. Sippel, S. A. Braun, Y. Weng, and F. Zhang, 2015: Dynamics and predictability of Hurricane Nadine (2012) evaluated through convection-permitting ensemble analysis and forecasts. *Mon. Wea. Rev.*, **143**, 4514–4532, doi:10.1175/MWR-D-14-00358.1.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600, doi:10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

Palmer, T. N., 1997: On parametrizing scales that are only somewhat smaller than the smallest resolved scales, with application to convection and orography. *Workshop on New Insights and Approaches to Convective Parametrization*, Reading, United Kingdom, ECMWF, 328–337.

——, 2001: A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parametrization in weather and climate prediction models. *Quart. J. Roy. Meteor. Soc.*, **127**, 279–304, doi:10.1002/qj.49712757202.

Ritchie, H., C. Temperton, A. Simmons, M. Hortal, T. Davies, D. Dent, and M. Hamrud, 1995: Implementation of the semi-Lagrangian method in a high-resolution version of the ECMWF forecast model. *Mon. Wea. Rev.*, **123**, 489–514, doi:10.1175/1520-0493(1995)123<0489:IOTSLM>2.0.CO;2.

Sela, J., 2010: The derivation of the sigma pressure hybrid coordinates semi-Lagrangian model equations for the GFS. NCEP Office Note 462, 31 pp., http://www.lib.ncep.noaa.gov/ncepofficenotes/files/on462.pdf.

Shutts, G., 2005: A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Quart. J. Roy. Meteor. Soc.*, **131**, 3079–3102, doi:10.1256/qj.04.106.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330, doi:10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2.

——, and ——, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125**, 3297–3318, doi:10.1175/1520-0493(1997)125<3297:EFANAT>2.0.CO;2.

——, O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, I. T. Jolliffe and D. B. Stephenson, Eds., John Wiley and Sons, 137–163.

——, ——, and Y. Zhu, 2006: The attributes of forecast systems. *Predictability of Weather and Climate*, T. N. Palmer and R. Hagedorn, Eds., Cambridge University Press, 584–595.

Wang, X., D. Parrish, D. Kleist, and J. Whitaker, 2013: GSI 3DVar-based ensemble–variational hybrid data assimilation for NCEP Global Forecast System: Single-resolution experiments. *Mon. Wea. Rev.*, **141**, 4098–4117, doi:10.1175/MWR-D-12-00141.1.

Wei, M., Z. Toth, R. Wobus, Y. Zhu, C. H. Bishop, and X. Wang, 2006: Ensemble transform Kalman filter-based ensemble perturbations in an operational global prediction system at NCEP. *Tellus*, **58A**, 28–44, doi:10.1111/j.1600-0870.2006.00159.x.

——, ——, ——, and ——, 2008: Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast systems. *Tellus*, **60A**, 62–79, doi:10.1111/j.1600-0870.2007.00273.x.

Whitaker, J. S., and T. M. Hamill, 2002: Ensemble data assimilation without perturbed observations. *Mon. Wea. Rev.*, **130**, 1913–1924, doi:10.1175/1520-0493(2002)130<1913:EDAWPO>2.0.CO;2.

——, and ——, 2012: Evaluating methods to account for system errors in ensemble data assimilation. *Mon. Wea. Rev.*, **140**, 3078–3089, doi:10.1175/MWR-D-11-00276.1.

——, ——, X. Wei, Y. Song, and Z. Toth, 2008: Ensemble data assimilation with the NCEP Global Forecast System. *Mon. Wea. Rev.*, **136**, 463–482, doi:10.1175/2007MWR2018.1.

Wu, W., R. J. Purser, and D. F. Parrish, 2002: Three-dimensional variational analysis with spatially inhomogeneous covariances. *Mon. Wea. Rev.*, **130**, 2905–2916, doi:10.1175/1520-0493(2002)130<2905:TDVAWS>2.0.CO;2.

Yang, F., 2015: Comparison of forecast skills between NCEP GFS four cycles and on the value of 06Z and 18Z cycles. *27th Conf. on Weather Analysis and Forecasting/23rd Conf. on Numerical Weather Prediction*, Chicago, IL, Amer. Meteor. Soc., 15A.1, https://ams.confex.com/ams/27WAF23NWP/webprogram/Paper273676.html.

——, H. Pan, S. K. Krueger, S. Moorthi, and S. J. Lord, 2006: Evaluation of the NCEP Global Forecast System at the ARM SGP site. *Mon. Wea. Rev.*, **134**, 3668–3690, doi:10.1175/MWR3264.1.

——, K. Mitchell, Y. Hou, Y. Dai, X. Zeng, Z. Wang, and X. Liang, 2008: Dependence of land surface albedo on solar zenith angle: Observations and model parameterizations. *J. Appl. Meteor. Climatol.*, **47**, 2963–2982, doi:10.1175/2008JAMC1843.1.

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.

Zhu, Y., 2005: Ensemble forecast: A new approach to uncertainty and predictability. *Adv. Atmos. Sci.*, **22**, 781–788, doi:10.1007/BF02918678.

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

——, G. Iyengar, Z. Toth, M. S. Tracton, and T. Marchok, 1996: Objective evaluation of the NCEP Global Ensemble Forecasting System. Preprints, *15th Conf. on Weather Analysis and Forecasting*, Norfolk, VA, Amer. Meteor. Soc., J79–J82.