The NCEP Global Ensemble Forecast System with the EnKF Initialization

Xiaqiong Zhou[1], Yuejian Zhu[2], Dingchen Hou[2], Yan Luo[1], Jiayi Peng[1] and Richard Wobus[1]

1) IMSG at EMC, NCEP, NWS, NOAA, College Park, Maryland,
2) EMC, NCEP, NWS, NOAA, College Park, Maryland

To be submitted to MWR

Corresponding Author: Xiaqiong Zhou, Email: xiaqiong.zhou@noaa.gov

IMSG at Environmental Modeling Center/NCEP/NOAA

5830 University Research Court
College Park, MD 20740

The NCEP Global Ensemble Forecast System with the EnKF Initialization

Abstract

A new version of the Global Ensemble Forecast System (GEFS) with all available upgrades is tested and compared with the operational version in a 2-year parallel run (00z only) in preparation for its implementation in NCEP operations. In the new GEFS, the main upgrades include the initialization scheme, the GFS model, and the horizontal and vertical resolutions. The ensemble initialization scheme, the breeding-based Ensemble Transformation with Rescaling (ETR) in the operational GEFS, is replaced by the EnKF scheme. The global forecast model used the recently implemented Global Forecast System (GFS) model (GSM 12.0.0). The horizontal resolution of GEFS increases from Eulerian T254 (~52 km) for the first eight days of the forecast and T190 (~70 km) for the second eight days to semi-Lagrangian T574 (~34 km) and T384 (~52 km). The sigma pressure hybrid vertical layers increase from 42 to 64 levels.

The verifications of geopotential height, temperature and wind fields at selected levels for the operational and new GEFS are performed against their own analyses. It is found that the new version significantly outperforms the operational up to day 8-10 except for larger warm bias over land in extra tropics. The results for probabilistic forecasts of precipitation show that the new system has similar performance to the operational in a two year period except for better reliability for the short range forecast in the warm seasons. The performance of the ensemble mean tropical cyclone track forecast varies with the basin and hurricane season. Slightly better skill for tropical cyclone tracks is found over Atlantic and Eastern Pacific based on the statics in four hurricane seasons (2011-2014). The improvement of track forecasts over the Western North Pacific is significant.

1. Introduction

The Global Ensemble Forecast System (GEFS) has been the one of the most important components of the NOAA's environmental prediction operation systems since its operation in 1993 (Toth and Kalnay 1993 and 1997). The forecast skill of the GEFS has improved significantly since then, benefiting from the upgrades of the initialization scheme (Wei et al. 2006 and 2008), the inclusion of the model uncertainty (Hou et al. 2006), higher model resolution and larger ensemble size, as well as the continuous improvement of the GFS model (Han and Pan 2011; Juang 2011 and 2014; http://www.emc.ncep.noaa.gov/GFS/impl.php) and the Global Data Assimilation System (Wu et al. 2002; Kleist et al. 2008ab; Wang et al. 2013; Kleist and Ida 2015).

GEFS has a long history of using the breeding scheme to provide an initial ensemble perturbation for medium-range ensemble prediction (Toth and Kalnay 1993 and 1996). The basic idea is to simulate the analysis cycling but without observation data involved. The initial ensemble uses the forecast perturbations from a previous cycle with regional rescaling. The perturbations generated from the breeding cycles represent fast growing forecast errors. It is expected that the breeding vectors (BV) can estimate the fastest forecast error growth, thus improving the forecast (Kalnay 2001).

In 2005, the Ensemble Transformation was introduced into the breeding method (Wei et al. 2006 and 2008). The forecast perturbations from the breeding cycles are multiplied by a transformation matrix to ensure that analysis perturbations would be consistent with the analysis error covariance provided by the user. It expected that the ensemble-based variance with ETR spans more directions than the one without ETR. Wei et al. (2008) compared the probabilistic scores of the BV and ETR. They found that the ETR had better performance than the BV.

Another major source of the uncertainty is from the model itself. The model uncertainty in GEFS is represented by using the Stochastic Total Tendency Perturbation (STTP) scheme (Hou et al. 2006). It was implemented in the operational GEFS in 2012. In the STTP scheme, the stochastic forcing is added every 6 hours to the total tendencies of ensemble perturbations for the model variables (temperature, specific humidity and winds, Hou et al. 2006 and 2008). The total time tendency for each variable is perturbed randomly after multiplying by a rescaling factor. The rescaling factor is a function of the location and lead time, which is used to control the perturbation's amplitude based on previous experience. Generally, the extratropics have larger perturbations than the tropics and the perturbations grow with lead time. The inclusion of the model uncertainty increases ensemble spread, thus reducing the RMSE and gaining higher probability skill (Hou et al. 2006).

There was one major operational upgrade to the Global Data Assimilation (GDAS) system at NCEP in May 2012: an Ensemble Kalman Filter (EnKF) was implemented in the NCEP GDAS (Whitaker and Hamill 2002; Whitaker et al. 2008; Wang 2010; Wang et al. 2013, Kleist and Ide 2015). The analysis was performed using a hybrid variational-ensemble data assimilation system in which the flow-dependent background error covariance, based on the short-range ensemble forecasts from EnKF, is incorporated with the static background error covariance of the Gridpoint Statistical Interpolation (GSI), a 3DVAR algorithm (Wu et al. 2002; Kleist et al. 2009). A dual resolution strategy was applied in the hybrid system to reduce computational cost. The analysis of GSI was performed at the model resolution T574 while the 80-member EnKF ensemble ran on T254 resolution. In the hybrid system, the error covariance estimate combined the static and ensemble solution with 25% and 75% weighting respectively. The implementation of the hybrid system improved the quality of analysis substantially (Kleist and Ide 2015). The

success of EnKF in the NCEP GDAS provides an alternative source of ensemble initial conditions for the operational GEFS. A direct benefit of the use of EnKF perturbations is the computational savings since EnKF had already been a part of the NCEP operation system.

Zhou et al. (2015) compared the GEFS performance with the GEFS operational (ETR) and EnKF initialization schemes. It is found that EnKF is comparable with ETR except for slight degradation in Southern Hemisphere due to overspread. Large spread in EnKF to some degree is favorable for the data assimilation to avoid filter divergence but not favorable for the medium range weather forecast especially when STTP is included as in the operation. The ensemble covariance inflation (e.g, Whitaker and Hamill 2002, 2012) and additive noise inflation (e.g., Whitaker et al. 2008, Houtekamer et al. 2005) is carried out on the posterior ensemble to parameterize the other error sourcse including the model itself in the 2012 EnKF implementation. It is a challenge for EnKF to satisfy the data assimilation and the medium range forecast at the same time.

In the recent EnKF implementation (Q1FY2015), a new stochastic physics suite was employed to represent the model error in order to replace the artificial additive inflation. The stochastic physics suite has three components: stochastically perturbed physics tendencies (SPPT, Buizza et al., 1999; Palmer, 1997 and 2001), stochastically perturbed PBL humidity (SHUM) and stochastic kinetic energy backscatter (SKEB, Shutts, 2005). All use auto-regressive process of first order (AR(1)) random pattern generators to produce spatially and temporally correlated perturbations (Li et al. 2008; Berner et al. 2009). Both GSI and EnKF analysis were performed at same model resolution T574.

The main goal of this paper is to perform a comprehensive comparison between the current operational GEFS and the new version with all available updates. In addition to the change of the ensemble initialization scheme, we also upgraded the GFS model to correspond to the Jan. 2015 GFS upgrade and increased the ensemble model resolution.

Section 2 presents a brief introduction of proposed changes in GEFS. Section 3 mainly focuses on traditional verification. The performance of precipitation forecasts and tropical cyclone track forecasts are demonstrated in section 4 and 5. The last section includes the conclusion and discussion.

2. GEFS upgrade

Table 1 lists the major updates in the new system. The present operational GEFS was implemented on Feb. 2012 with the horizontal resolution Eulerian T254 (~52 km) for the first eight days of the forecast and T190 (~70 km) for the second eight days. The new version increases resolution to the semi-Lagrangian T574 (~34 km) and T384 (~52 km). The number of sigma pressure hybrid vertical layers increases from 42 to 64. The increased resolution is chosen to fit the wall time window in the operational environment (about 60 mins). With the increase of the model resolution, the new GEFS version provides 0.5 degree GRIB2 files at 3 hr time intervals up to day 8.

The GFS model is updated to GSM version 12.0.0. The details of the updates to the GFS model can be found at http://www.emc.ncep.noaa.gov/GFS/impl.php. The model configuration follows the settings of the high-resolution deterministic operational GFS T1534 except for some resolution-depended parameters such as convective gravity wave drag and the critical relative humidity which allows the formation of the partial cloudiness. The major change in the GFS

model is the replacement of Eulerian dynamics with the Semi-Lagrangian with the use of Hermite interpolation in a hybrid vertical coordinate and a reduced Gaussian grid in the horizontal. The advantage of the Semi-Lagrangian GFS is to allow the use of a larger time step thus integrating the high-resolution model more efficiently (Harold et al 1995; Juang and Hong 2010). A time step of 900 seconds is applied for the Semi-Lagrangian GFS at T574 with equivalent horizontal resolution 34 km, whereas the operational Eulerian T254 (~ 52 km) uses a 300 seconds time step.

As mentioned in the previous section, the major upgrade in the recent EnKF implementation is the use of the stochastic physics in EnKF ensemble forecasts. The short-range forecast in the EnKF cycling uses the same horizontal and vertical resolution as the new GEFS version ($T574_{SL}L64$). ENKF 6hr ensemble forecast perturbations are used instead of the analysis ensemble since only the EnKF from the previous cycle is available at the time when GEFS starts in the NCEP operational environment. This is because the EnKF is only run as part of the late cycle within the global data assimilation system for the purposes of prescribing background error covariance in the cycle that follows. Figure 1a shows that the artificially inflated perturbations damp quickly in 6 h forecast. The amplitude of EnKF perturbations remains much larger than the perturbations in ETR. Figure 1 shows that the amplitude of the EnKF perturbations from the recent implementation becomes similar to that of the ETR. In addition, the amplitude of the EnKF perturbations with the SPPT suite increases in 6-hr forecast although the growth rate is rather small. The updated EnKF has become suitable to provide the initial ensemble for the GEFS.

The ensemble size remains the same as the operation (20 members and one control run) due to the limitation of computer resources. The initial perturbations are from EnKF 6-hr forecast

ensemble instead of the breeding cycle. The EnKF has 80 members every 6 hours. 20 of 80 members are selected to initialize the ensemble forecast for each cycle so that every member is used for a GEFS medium range forecast once per day.

As in the operation, the tropical cyclone perturbations are separated and updated separately. The ensemble forecast fields are separated into the environment and tropical cyclone (TC) components when TCs are active over the ocean (Kurihara et al. 1993 and 1995). The TC perturbations are calculated and added to the analysis after the vortex of each ensemble member is relocated to the observed location. The TC perturbation adjustments (P) to the initial state of each ensemble member are calculated by the following formula (Liu et al. 2000 and 2006):

$$P = C \cdot (X - Xc) \cdot \|Xc\| / \|X - Xc\|$$

X represents the model variables of the TC component such as the wind, temperature, mixing ratio,or sea level pressure for the ensemble member. Xc corresponds to the same variable for the control. ||X|| is the square root of the sum of X over the whole hurricane area. The TC perturbations are calculated from the difference of TC components between the ensemble member and the control forecasts. A scaling factor C is artificially set to 0.05 to scale down the perturbation amplitude, which is about5% of the magnitude of the TC component in the control forecasts.

The parameters in the STTP scheme are tuned slightly due to the new perturbation scheme, the increased model resolution and other model changes. The surface pressure tendency is no long perturbed to avoid numerical instability. The perturbation amplitude of other model state variables around the time of model truncation (192 hours) is increased to obtain better error-skill relationship in the day 8-16 forecast.

3. Verification of model variables

A two-year parallel run (June 2013 – May 2015, 00Z cycle only) was performed and compared with operational forecasts. All of the 20-member ensemble forecast data are interpolated to a 2.5 X 2.5 degree horizontal resolution. The ensemble forecast is verified against its own analysis using the NCEP ensemble verification package (Zhu et al. 1996; Zhu and Toth 2008; Zhu. 2005). The scorecard summarizes the performance of the forecasts of geopotential height at 500 hPa and 1000 hPa levels, wind fields at 10 m, 850 hPa and 250 hPa levels, temperature at 2 m and 850 hPa level over Northern America (NA), NH, SH and tropics (Fig. 2). The comparison of these two systems is quantified by the root mean square error (RMSE), pattern anomaly correlation (PAC), ensemble mean forecast bias, the Continuous Rank Probability Score Skill (CRPSS), and Brier Skill Score (BSS). Verifications by additional methodologies can be found at the website http://www.emc.ncep.noaa.gov/gc_wmb/xzhou/Para_2013-2015_test.HTML.

The scorecard shows that the new system generally outperforms the operation up to day 8 over extratropical regions. The ensemble mean forecasts of all verified variables are more accurate and the probability scores (CRPSS and BSS) are significantly higher over NH, SH, as well as NA. The improvement is generally significant as shown by the bootstrap test at the 95 % confidence level up to day 8. Degradation is seen in the probabilistic scores at lead times longer than 12 days, but this negative influence is considered negligible since the probability forecast skills are already very low at these lead times.

The ensemble mean forecast for wind fields over the tropical region is much better at all lead times with respect to AC and RMSE, but there is no clear evidence of improvement with respect to other variables and probability scores over the tropics. CRPSS and BSS are generally degraded beyond day 3.

The evaluation of the 500 hPa geopotential height field shows that underspread is common in both systems especially in week 2. The updated GEFS has slightly smaller spread than the operational in week 1 but becomes the same in week 2 (Fig. 3a). The RMSE of 500 hPa geopotential height is significantly smaller over NH to day 9.

The anomaly correlation of 500 hPa forecasts is a measure of the models' overall performance for large-scale weather patterns. A threshold value of 0.6 is regarded as an indication of a useful forecast in which the locations of trough and ridges at 500 hPa are well predicted. The new GEFS extended the skillful forecast from 9 days to 9.5 days in terms of AC (Fig. 3b). Another threshold which is defined as a headline score in ECMWF's Strategy 2011-2020 is a 500 hPa geopotential anomaly correlation of 80%. The forecast lead time for this threshold increases from 8.4 to 8.7 days in the new GEFS.

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

A major concern is the larger bias error in the new system than in the operational. The degradation in bias is consistent with the performance of the high resolution GFS deterministic

forecast (ref). The new GFS has warm/dry bias over some land areas. Fig. 5 shows the absolute error and bias error of 2-meter temperature over NA averaged for the two-year period. The two systems have same absolute error but the warmer bias in the new system. The time series of 2-m temperature bias over NA shows that bias varies with season during the two-year period (Fig. 6). The large degradation in terms of surface temperature bias is more evident in summer. The operational usually has warm bias in the summer and changes to cold bias in winter. The new system increases the warm bias in summer but reduces the cold bias in winter. The distribution of temperature bias (Fig.7) shows the summer warm bias over NA mainly located in Central America. It is suggested by the GFS development group that the land surface update in GFS 2015 implementation resulted in lower soil moisture when the Global Land Data Assimilation System (GLDAS) /the Coupled Forecasting System (CFS) soil moisture climatology at T574 (~27 km) replaced 1 degree bucket soil climatology. Evaporation parameters were not retuned as part of the 2015 implementation which used drier soil climatology. Dry soil allows too much sensible heat flux and too little latent heat flux in hot air masses over cropland. Further testing had been proposed and will be upgraded in the next implementation.

This kind of the systematic model errors can be removed with postprocessing algorithms. Bo et al. (2012) developed a Kalman filter type algorithm to accumulate the decaying averaging bias and produce the bias-corrected ensemble, in which the decaying averaged bias is updated with a weight of 2% given to the most recent data, thus the bias contains the accumulated information about the behavior of the ensemble forecasting system in the recent 50-60 days. This post processing technique, as one of the operational components in NCEP, is applied to the ensemble forecasts of both NCEP and Meteorological Service of Canada 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. Fig. 8 shows the bias corrected ensemble forecasts for the NH 2-meter temperature averaged over one year. The parallel raw forecast degrades the surface temperature forecast compared to the operational, likely related to its larger bias errors. Bias correction improves both the operational and the parallel runs, especially the latter. RMSE in the bias corrected parallel forecast is significantly less than in the bias-corrected operational forecast or either uncorrected forecast.

4. Precipitation verification

Several verification scores are used to assess the skill of the ensemble-based probability precipitation forecast. The quantitative precipitation forecast (QPF) and probabilistic QPF with uniform 1 degree horizontal resolution is 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).

The continuous verification methods include the continuous ranked probability score (CRPS), RMSE/SPREAD, MERR/absolute error. CRPS measures the area difference between the cumulative distributions of forecasts and observations. RMSE is measured as one aspect of the comparison but the results of RMSE are not presented since it does not objectively measure the complex spatial distribution of precipitation, and it is sensitive to the discontinuities and large values of precipitation.

For categorical verification, the precipitation is categorized by the 24-hr accumulated precipitation with the threshold amount greater than 1 mm, 5 mm, 10 mm and 20 mm respectively. The evaluation methods include, Brier score/Brier skill score (BS/BSS), reliability, bias, the equitable threat score (ETS) and true skill score (TSS) (http://www.emc.ncep.noaa.gov/gmb/yluo/GEFS_VRFY/GEFS_PARA_SUMMARY.html). The

Brier score is the mean square error of a probability forecast from the observed probability which is either 1 or 0 depending on whether the categorized precipitation event occurred or not. Brier skill score calibrates the Brier score to the 10-year mean of CCPA as the climatology to avoid the dependence of BS on the frequency of the event. 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. ETS, TSS and bias are based on the 2 X 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 new system is generally similar to the operational. There is no significant difference between these two systems in terms of CRPS, ETS and TSS (not shown). RMSE and the ensemble spread of the precipitation are greater in the new version than the operational (not shown). There are some suggestions that the new version has better BS/BSS during the first three days (Fig. 9a), less bias during the first week for the light precipitation forecast (not shown) and greater reliability (Fig. 10a) for short-range forecasts (day 1-3) for precipitation greater than 1 mm/24 hr and 5 mm/24 hr. In a perfect forecast, the predicted probabilities should be exactly equal to observed probability. Fig. 10 shows that both systems are overconfident for the probability forecast of precipitation greater than 5mm/24 hr. The reliability diagram for the new version is slightly closer to the diagonal line, which is the main contribution to a better BSS for the short range forecast of precipitation. More careful examination shows that the better reliability is only seen in the warm season (April-October) for the forecast of precipitation greater than 1 mm and 5 mm in 24 hrs at day 1-3, which is also the main contributions to the higher BSS (Figs. 9ab and Figs.10ab). The performance of the

precipitation ensemble forecast is unchanged in the cold seasons (November to next March, not shown).

5. Tropical cyclone track forecast

The hurricane activities in the North Atlantic in 2013 and 2014 were well below average. There were only 8 named tropical cyclones in 2014 compared to a climatological mean of 12.1. The total number of tropical cyclones in 2013 (14) exceeds the climatology but only 2 (Hurricanes Humberto and Ingrid) reached hurricane intensity (6.4 in climatology) with no storms reaching hurricane category 2 intensity. To increase the sample, we have run the medium range forecast for the 2011 and 2012 hurricane seasons (June – October 2012).

Fig. 11 shows the ensemble mean forecast errors of tropical cyclone tracks over the North Atlantic, Eastern Pacific and Western North Pacific up to day 7. The mean track errors over Atlantic are slightly smaller in the new version up to day 6 and turn larger in longer lead times than the operational, but the difference is generally insignificant. Similar performance is found over EP. The new version is slightly better but not significantly. The ensemble-mean track forecast error is significantly reduced over WNP; the track error in day 5 is reduced 20% from 250 nmi to 200 nmi.

The performance of TC track forecast varies with season. One concern is the significant degradation of the day-6 to 7 TC track forecast in the 2012, 2013 and 2014 hurricane seasons over Atlantic (not shown). No significant difference is identified in the 2011 season. There are 77 cases for day 6 forecasts with 61 cases occurred in 2012 and 8 each in 2013 and 2014. Fig. 12 shows the displacements of forecasted TCs positions for day 6 from their observed locations for the operational and parallel. The forecast TCs spread to larger error radii in the parallel run and

more cases have track error larger than 500 km. The TCs in the parallel run tends to have larger south and east bias. Note that the degradation of the parallel run came from the forecasts of Hurricanes Nadine (2012), Michael (2012) and Edouard (2014).

As the fourth longest-lived hurricane over Atlantic on record (10 Sep. - 4. Oct. 2012), Hurricane Nadine contributes to 30 of 77 cases of 6 day forecasts. The storm moved northwestward in its early stage and then turned northward and eastward later. Thereafter, Nadine weakened and accompanied with an unusual rotating movement over the ocean. It moved clockwise and then counterclockwise before transitioned to an extratropical low system. The forecast locations for day 6 initiated from the early stage were generally located northwest of the observed locations as a result of a slow northeastward turning compared with the observed one (Figs. 13 a). The eastward deviations resulted from the poor forecasts of the TC' unusual movement at the late stage. Fig. 13a shows that the storm in the parallel run moved eastward instead of a cyclonical turning. This contributes to larger track forecast errors.

Hurricane Michael (2012) moved northward slowly after it formed from a mid- to upper-level short wave disturbance followed by zig-zag movement as a result of a change of the large scale steering flow. The TC in the parallel system moved to the east instead of the north, resulting to large track forecast errors.

Hurricane Edouard (2014) was a typical storm steered by the large-scale flow around a subtropical ridge. The recurve from the northwestward to northeastward is well predicted in both the operation and the parallel, except that the later has much slower northward speed, resulting in much larger track errors.

6. Conclusion and discussion

The primary goal of this study was to introduce the new Global Ensemble Forecast System in NCEP and provide a comprehensive verification compared to the current operational version. Identification of the aspects of the GEFS performance that can be improved is helpful for future development and a comprehensive evaluation is very useful for GEFS users.

The major upgrade to the GEFS is the ensemble initialization scheme. The ETR scheme, an updated breeding scheme implemented in the operational GEFS in 2005, will be replaced by the EnKF scheme. An operational upgrade of the NCEP GFS/GDAS was implemented on Jan. 14, 2015. The new EnKF perturbations have similar amplitude to the current operational ETR. The STTP which is used to represent the uncertainty of the model itself is slightly tuned to obtain optimal error-spread relationship. There is a direct benefit for the operational computer resource to replace ETR with EnKF since the latter has been implemented as an important part of the operational GDAS at the GEFS upstream and the computing resources for the breeding cycles and the calculation of the transform matrix in ETR can be saved.

Another upgrade is the GFS model (GSM version 12.0.2) itself and GDAS. The new GEFS uses the same version of the global model as the operational high-resolution deterministic GFS which was implemented on January 14 2015. The forecast model was changed significantly with the most notable change being the upgrade from Eulerian to Semi-Lagrangian dynamics. The model resolution increased from Eulerian T254 (~52 km) to Semi-Lagrangian T574 (~34 km) and increased the resolution from 42 to 64 vertical levels.

The new GEFS is generally more skillful than the operational up to day 8-10 over extratropical regions with respect to the ensemble mean and probability forecasts of the model variables such as geopotential height, temperature and wind fields. The improvement is

significant at the 95% confidence level as evaluated by a bootstrap test. The new system improved on the operational AC of 500hPa forecasts by extending the skillful forecast from 9 to 9.5 days.

The new GEFS has a warm surface temperature bias over Central America in summer. The GFS development group suggested that the land surface update in GFS 2015 implementation results in lower soil moisture due to a change of soil climatology data. This systematic temperature bias can be corrected by subtracting the decaying averaged bias from the ensemble raw forecasts (Cui et al. 2012).

The new GEFS outperformed the operational in forecasting tropical wind fields 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 new system has a positive impact on the probability forecasts over tropics.

The probabilistic forecasts of precipitation over CONUS were evaluated against CCPA with a set of verification tools including both continuous methods and categorical methods. The performance of the precipitation forecast is unchanged overall compared with the operational, except for higher reliability and BSS scores in the short-range forecast. The similar performance for precipitation forecasts is likely related to the similar physics in these two systems. The major upgrade to the GFS model is its dynamics and there are no significant changes to physical process.

Based on the statistics of four hurricane seasons (2011-2014), we found that the upgraded GEFS improved the tropical cyclone track forecasts over WNP significantly but the improvement over EP and Atlantic is very limited. Instead, there is a slight degradation in day 6-

7 track forecast, especially for 2012, 2013 and 2014. The large track forecast errors mainly came from the forecasts of three long-lived hurricanes especially Hurricane Nadine (2012).

This performance of GEFS depends strongly on the skill of the GFS model. The warm surface bias over CONUS is consistent with the performance of the high-resolution deterministic GFS in spite of lower horizontal resolution in the ensemble forecast. This systematic bias in the operational GFS was noticed in the surface temperature forecast over Central America and a fix was proposed and implemented in new GFS. The degradation of 6-7 day track forecast in GEFS is also consistent with the performance of GFS. The reason of the degradation is not clear. We noticed that the GEFS TC track forecasts for 2015 Atlantic hurricane season is much better than operational. With limited sample size, it is not clear that the degradation in the long lead time for previous seasons is just a bad luck or a systematic deficiency.

Reference:

Cui, B. Z. Toth, Y. Zhu, and D. Hou, 2012: Bias Correction for Global Ensemble Forecast. *Wea. Forecasting, 27, 396-410*.

Corazza, M., E. Kalnay, DJ Patil, E. Ott, J. Yorke, I Szunyogh and M. Cai, 2001: Use of the breeding technique in the estimation of the background error covariance matrix for a quasigeostrophic model. *AMS Symposium on Observations, Data Assimilation and Predictability*, Preprints volume, Orlando, FA, 14-17 January 2002.

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

Buizza, R., M. Miller, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System, Quart. J. Roy. Meteorol. Soc., 125, 2887–2908.

Hamill T. M. 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Weather Forecast.* 14: 155–167.

Hamill, T. M., Snyder, C. and Morss, R. E. 2000: A comparison of probabilistic forecasts from bred, singular-vector, and perturbed observation ensembles. *Mon. Wea. Rev.* 128, 1835–1851.

Han, J., and H.-L. Pan, 2011: Revision of Convection and Vertical Diffusion Schemes in the NCEP Global Forecast System. Weather and Forecasting, 26, 520-533.

Harold Ritchie, Clive Temperton, Adrian Simmons, Mariano Hortal, Terry Davies, David Dent, and Mats Hamrud, 1995: Implementation of the Semi-Lagrangian Method in a High-Resolution Version of the ECMWF Forecast Model. *Mon. Wea. Rev.*, 123, 489–514.

Hou, D., Z. Toth, and Y. Zhu, 2006: A stochastic parameterization scheme within NCEP global ensemble forecast system. *18th AMS Conference on Probability and Statistics*, 29 January –2 February 2006, Atlanta, Georgia.

Hou, D., Z. Toth, Y. Zhu, and W. Yang, 2008: Impact of a stochastic perturbation scheme on NCEP global ensemble forecast system. *19th AMS Conference on Probability and Statistics*, 21–24 January 2008, New Orleans, Louisiana.

Hou, D, M. Charles, Y. Luo, Z. Toth, Y. Zhu, R. Krzysztofowicz, Y. Lin, P. Xie, Dong-Jun Seo, M. Pena, and B. Cui, 2014: Climatology-Calibrated Precipitation Analysis at Fine Scales: Statistical Adjustment of Stage IV toward CPC Gauge-Based Analysis. *J. Hydrometeor*, **15**, 2542–2557.

Houtekamer, P. L., Herschel L. Mitchell, Gérard Pellerin, Mark Buehner, Martin Charron, Lubos Spacek, Bjarne Hansen, 2005: Atmospheric Data Assimilation with an Ensemble Kalman Filter: Results with Real Observations. *Mon. Wea. Rev.*, 133, 604–620.

Houtekamer, P. L., Herschel L. Mitchell, Xingxiu Deng, 2009: Model error representation in an operational ensemble Kalman filter. *Mon. Wea. Rev.*, 137, 2126–2143.

Li, X., M. Charron, L. Spacek, and G. Candille (2008), A regional ensemble prediction system based on moist targeted singular vectors and stochastic parameter perturbations, Mon. Weather Rev., 136, 443–462.

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. [Available online at http:// www.nws.noaa.gov/om/tpb/472.htm.]

Liu, Q., S. J. Lord, N. Surgi, Y. Zhu, R. Wobus, Z. Toth and T. Marchok, 2006: Hurricane Relocation in Global Ensemble Forecast System, Preprints, 27th Conf. on Hurricanes and Tropical Meteorology, Monterey, CA, *Amer. Meteor. Soc.*, P5.13.

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. Monthly Weather Review, 139, 1583-1607.

Juang, H.-M. H., 2014: A discretization of deep-atmospheric nonhydrostatic dynamics on generalized hybrid vertical coordinates for NCEP global spectral model. NCEP Office Note 477, 40pp.

Juang, H.-M. H., and S.-Y. Hong, 2010: Forward semi-Lagrangian advection with mass conservation and positive definiteness for falling hydrometeors. Monthly Weather Review, 138, 1778-1791.

Kalnay, E 2001: Atmospheric modeling, data assimilation and predictability. Chapter 6. Cambridge University Press, UK.

Kleist, D.T., D.F. Parrish, J.C. Derber, R. Treadon, R.M. Errico, and R. Yang, 2008: Improving incremental balance in the GSI 3DVAR analysis system. Mon. Wea. Rev., 137, 1046-1060.

Kleist, D.T., D.F. Parrish, J.C. Derber, R. Treadon, W.-S. Wu, and S. Lord, 2008: Implementation of a new 3DVAR analysis as part of the NCEP global data assimilation system. Wea. Forecasting., 24, 1691-1705.

Kleist, D. T., D. F. Parrish, J. C. Derber, R. Treadon, W. S. Wu, and S. Lord, 2009: Introduction of the GSI into the NCEP Global data assimilation system. *Wea. Forecasting,* 24, 1691–1705.

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.

Kurihara, Yoshio, Morris A. Bender, Rebecca J. Ross, 1993: An Initialization Scheme of Hurricane Models by Vortex Specification. *Mon. Wea. Rev.*, 121, 2030–2045.

Kurihara, Y., Bender, M. A., Tuleya, R. E., Ross, R. J., 1995: Improvements in the GFDL hurricane prediction system. *Mon. Wea. Rev.*, 123, 2791-2801.

Palmer, T. N. (1997), On parametrizing scales that are only somewhat smaller than the smallest resolved scales, with application to convection and orography, in Proceedings of the ECMWF Workshop on New Insights and Approaches to Convective Parametrization, 4-7 November 1996, pp. 328–337, ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, UK.

Palmer, T. N. 2001: A nonlinear dynamical perspective on model error: A proposal for non-local stochasticdynamic parametrization in weather and climate prediction models, Q. J. R. Meteorol. Soc., 127, 279–304.

Shutts G. 2005. A kinetic energy backscatter algorithm for use in ensemble prediction systems. Q. J. R. Meteorol. Soc. 131: 3079–3102.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteorol. Soc.*, 74, 2317-2330.

Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, 127, 3297-3318.

Toth, Z., Talagrand, O., Candille, G. and Zhu, Y. 2003. Probability and ensemble forecasts. In: *Forecast Verification: A Practitioner's Guide in Atmospheric Science* (eds Ian T. Jolliffe and David B. Stephenson). John Wiley & Sons Ltd., England, 137–163.

Toth, Z., O. Talagrand, and Y. Zhu, 2006: The attributes of forecast systems, In book of: *Predictability of Weather and Climate,* Ed.: T.N. Palmer and R. Hagedorn, Cambridge University Press, 584-595.

Wang, X. and C. Bishop, 2003: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.* 60, 1140–1158.

Wang, X., Bishop, C. and Julier, S. 2004. Which is better, an ensemble of positive-negative pairs or a centered spherical simplex ensemble? *Mon. Wea. Rev.* 132, 1590–1605.

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).

Whitaker, Jeffrey S., Thomas M. Hamill, 2002: Ensemble Data Assimilation without Perturbed Observations. *Mon. Wea. Rev.*, 130, 1913–1924.

Whitaker, Jeffrey S., Thomas M. Hamill, Xue Wei, Yucheng Song, Zoltan Toth, 2008: Ensemble Data Assimilation with the NCEP Global Forecast System. *Mon. Wea. Rev.*, 136, 463–482.

Whitaker, J. S., and T. M. Hamill, 2012: Evaluating methods to account for system errors in ensemble data assimilation. *Mon. Wea. Rev.*, 140, 3078-3089.

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 A,* 58:1, 28-44.

Wei, M., Toth, Z., Wobus, R. and Zhu, Y. 2008: Initial perturbations based on the ensemble transform (ET) technique in the NCEP global ensemble forecast systems. Tellus 60A, 62–79.

Wu, W., R. J. Purser, and D. F. Parrish, 2002: Three dimensional variational analysis with spatially inhomogeneous covariances. *Mon. Wea. Rev.*, 130, 2905-2916.

Zhou X. et al. 2015: Comparisons of the Initialization Schemes between the Ensemble Transform and the Ensemble Kalman Filter for the NCEP Global Ensemble Forecast System.

Zhu, Y., 2005: Ensemble forecast: A new approach to uncertainty and predictability, *Advance in Atmospheric Sciences, Vol. 22,* No. 6, 781-788.

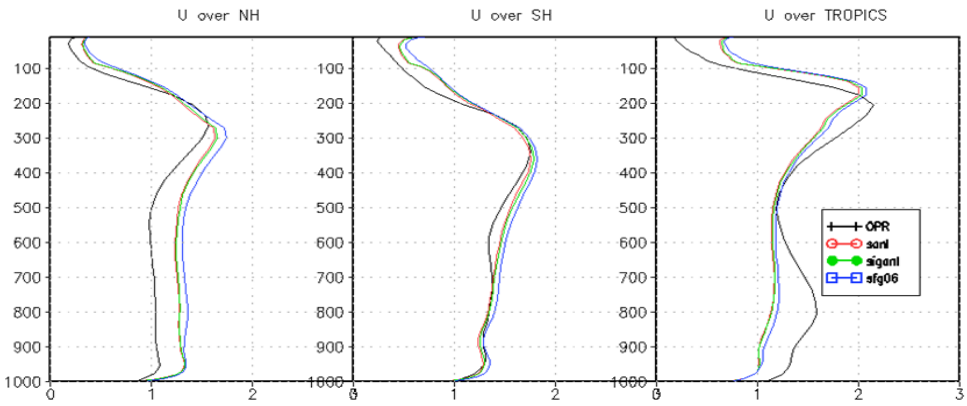
Zhu, Y., G. Iyengar, Z. Toth, M. S. Tracton, and T. Marchok, 1996: Objective evaluation of the NCEP global ensemble forecasting system. Preprints, *15th Conf. on Weather Analysis and Forecasting*, Norfolk, VA, Amer. Meteor. Soc., J79–J82.

Zhu, Y. and Z. Toth, 2008: Ensemble Based Probabilistic Forecast Verification Preprints, *19th Conf. on Probability and Statistics*, New Orleans, LA, Amer. Meteor. Soc., 2.2.

Table 1. The GEFS Configuration in operational and parallel runs

	V10.0.0 (OPR)	V11.0.0 (PARA)
GFS Model	Euler, 2012	Semi-Lagrangian, 2015
Resolution 0-192 h	T254 (52km) L42 (hybrid)	T_L574 (34km) L64 (hybrid)
Resolution 192-384 h	T190 (70km) L42 (hybrid)	T_L382 (52km) L64 (hybrid)
Computational Cost	84 nodes (+ post process)	300 nodes 1st segment 250 nodes 2nd segment
Execution time	55 min	35 min 1st segment 30 min 2nd segment
Output resolution	1^O x 1^O	0.5^O x 0.5^O for 0-8 days 1^O x 1^O the rest
Output frequency	6h	3h the first 8 days; 6h the rest
Initialization Scheme	ETR	EnKF

(a)



(b)

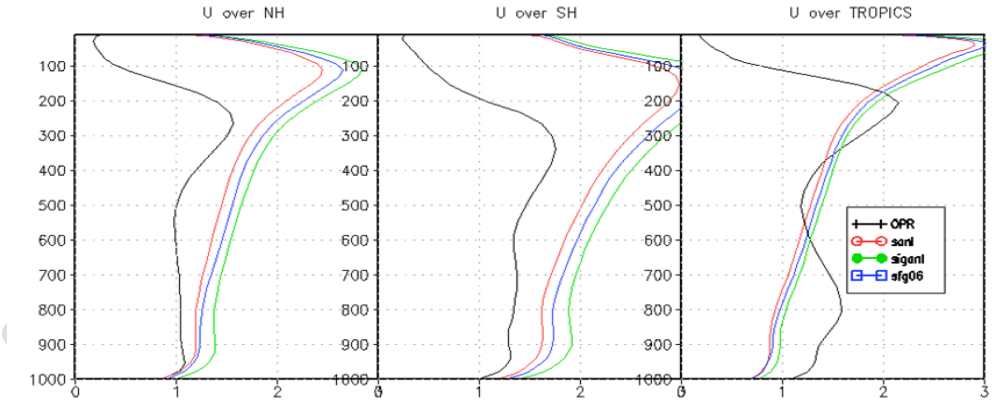


Fig. 1 The vertical profiles of initial perturbation spread for the horizontal wind component averaged over NH, SH and tropics in the 2015 (a) and 2012 (b) implementations.

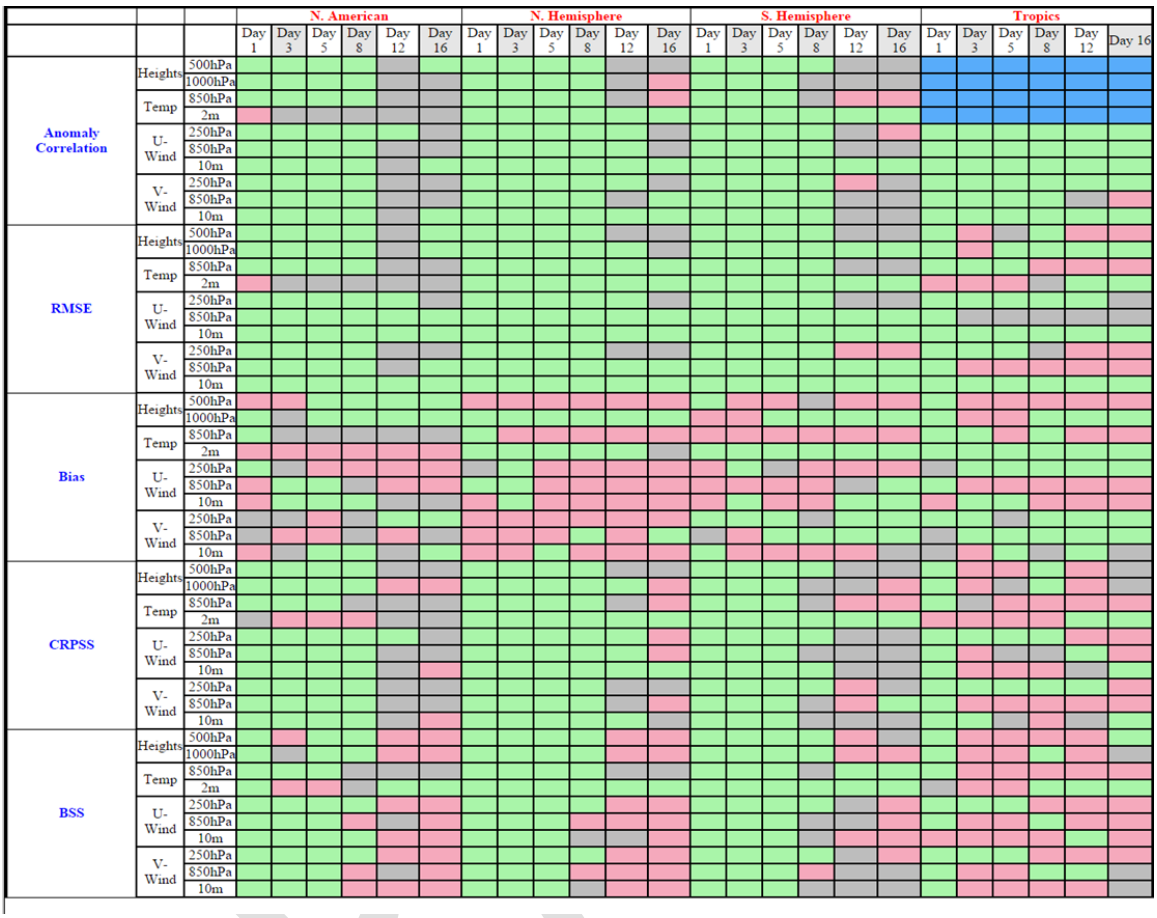


Fig.2 Scorecard for the parallel run versus the operational, verified against their respective analyses

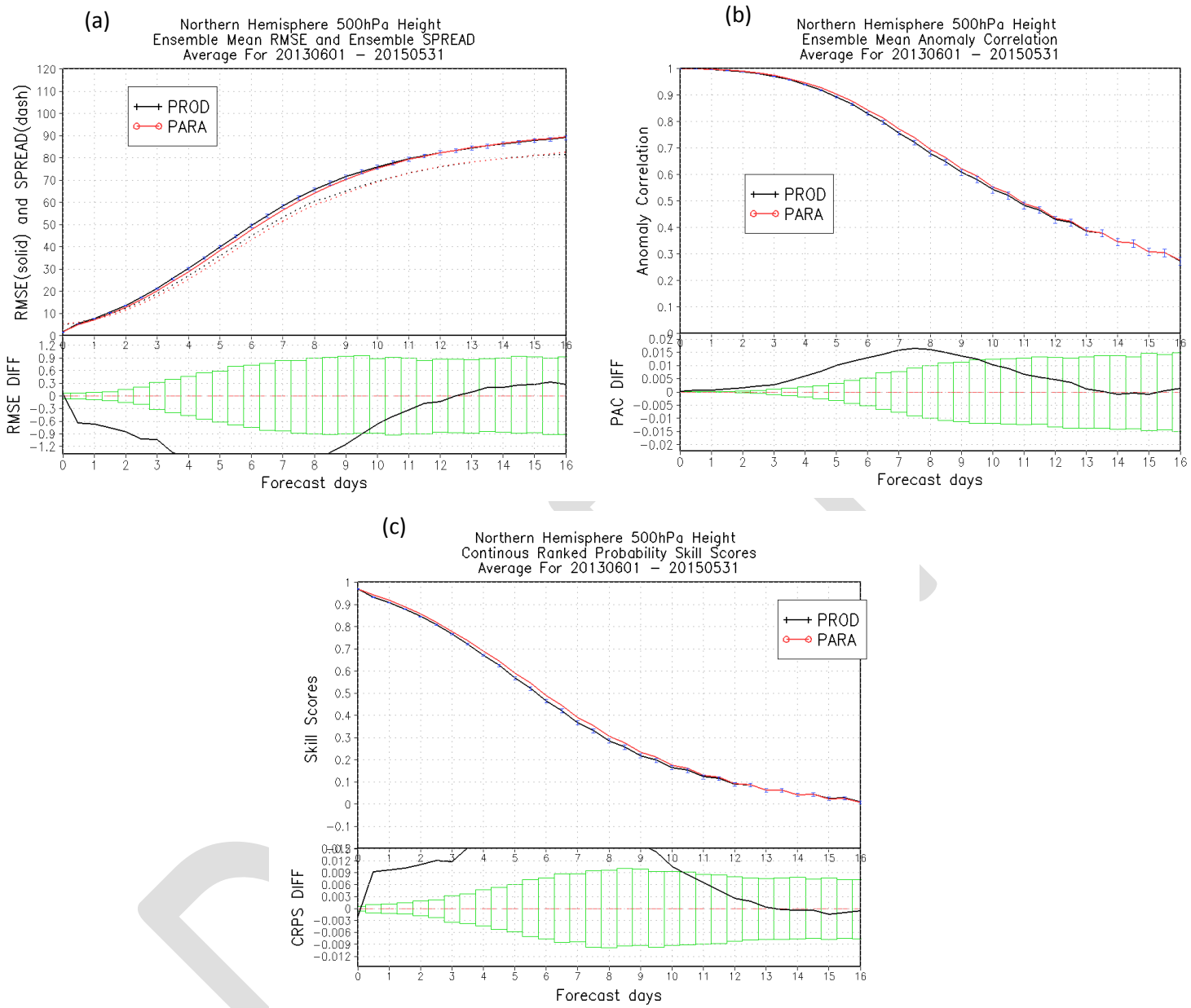


Fig. 3 Two-year average verification scores (1st June 2013 to 31st May 2015) for the 500hPa geopotential height over NH: (a) Ensemble mean RMS error (solid) and ensemble standard deviation (dotted), b) anomaly correlation and c) CRPSS. The lower panels show the difference and bootstrap significance test (green bars). The difference is significant at the 95% confidence level when value is outside the bars.

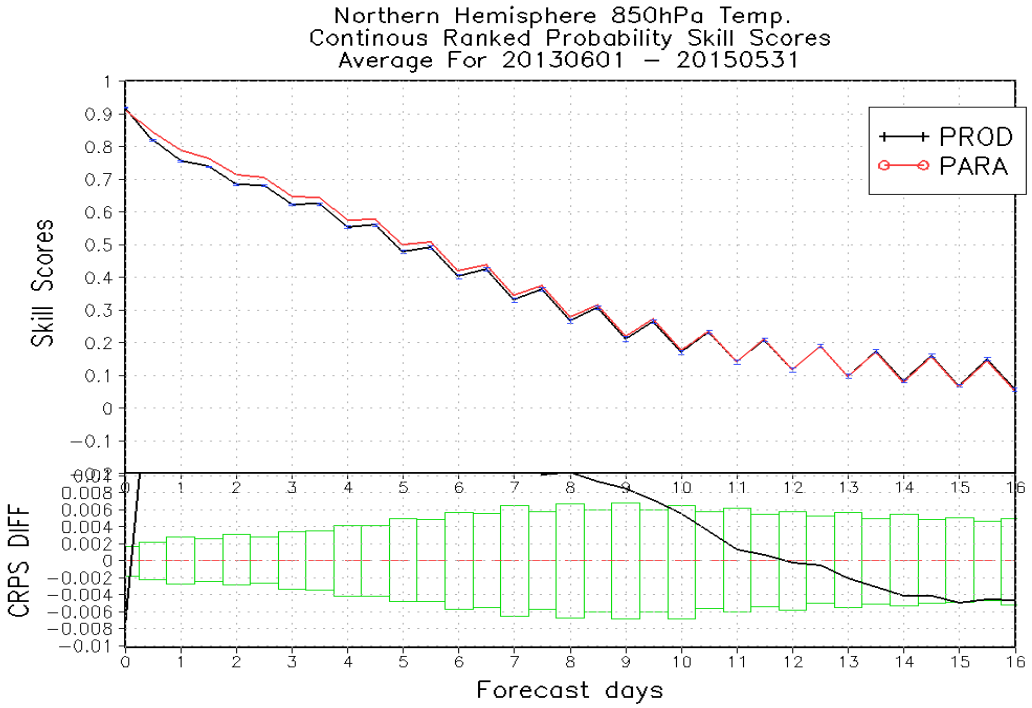


Fig. 4 As Fig. 3c except for 850hPa temperature.

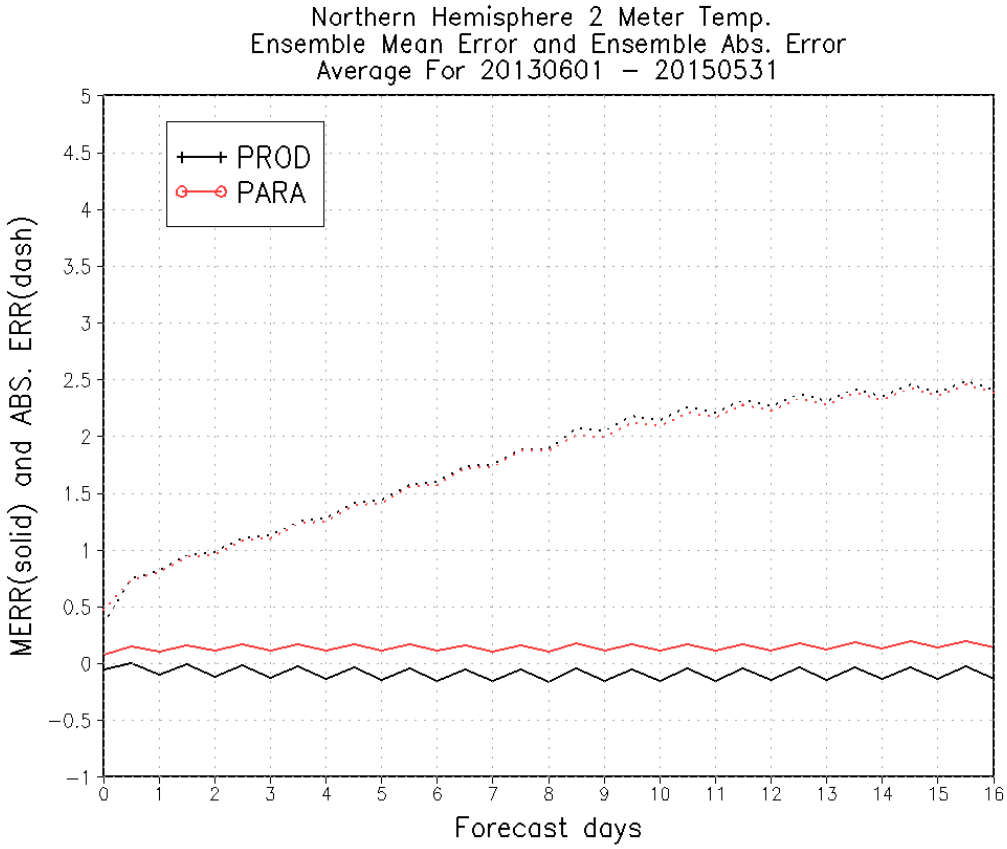


Fig.5 Ensemble mean bias (solid line) and absolute error (dashed line) of 2m temperature over NH.

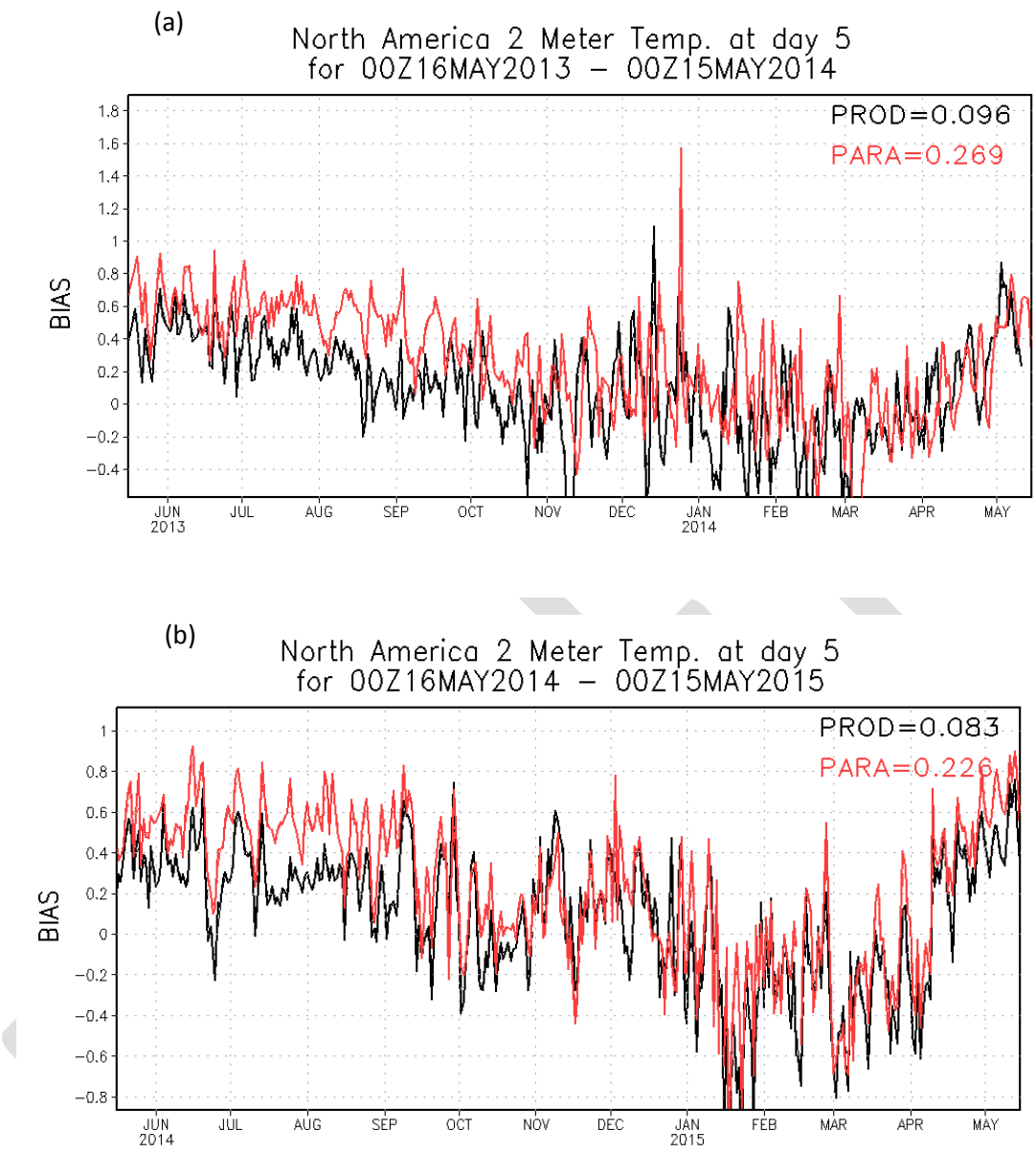


Fig. 6 The time series of 2m temperature bias over Northern America from May 15 2013 to May14 2014 (a) and from May 15 2014 to May14 2015 (b).

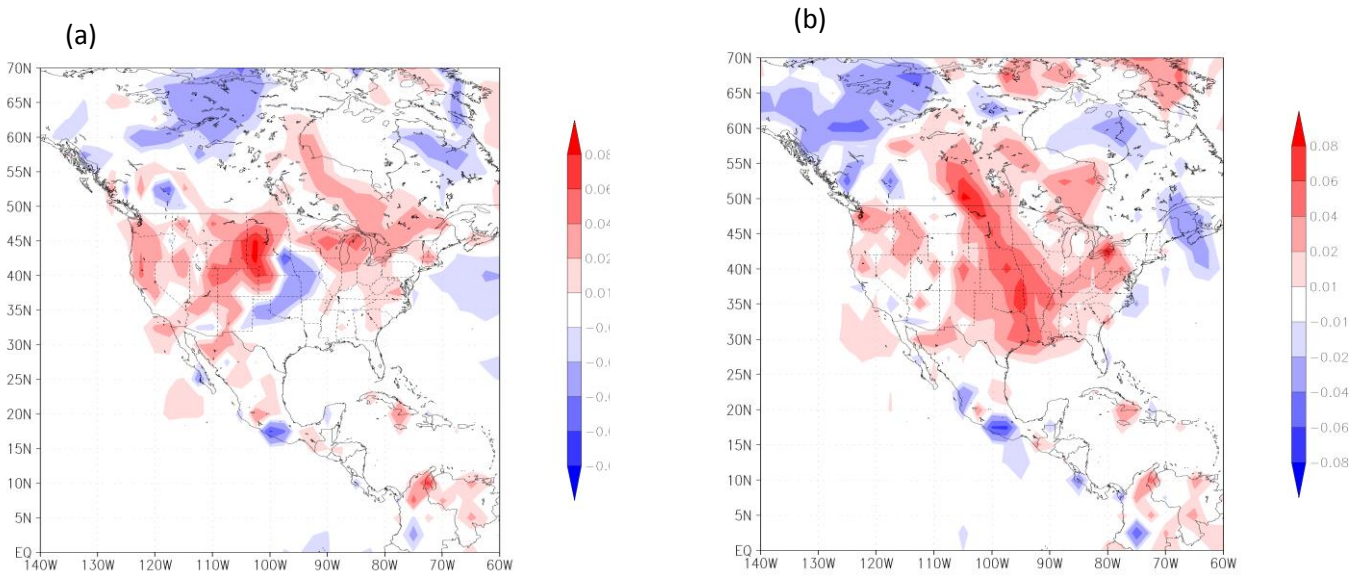


Fig. 7 The bias of surface temperature over North America for the parallel run averaged over June-August 2013 (a) and 2014 (b).

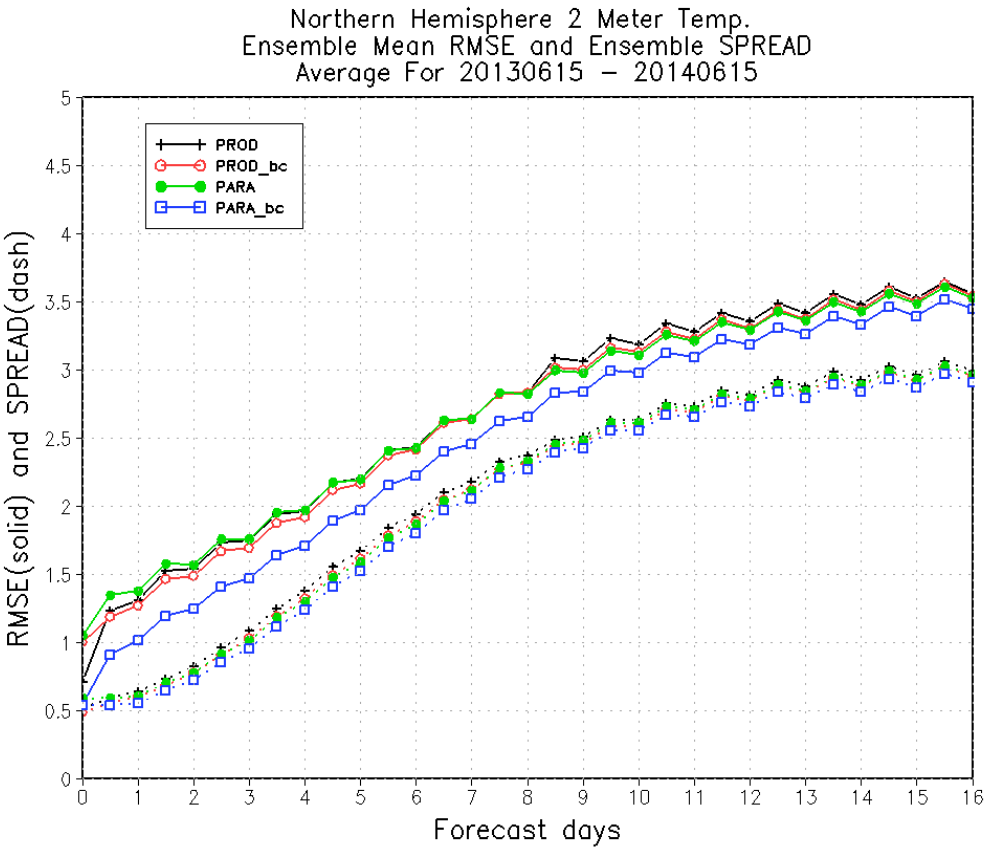


Fig. 8 The RMSE and spread of raw and bias-corrected 2m temperature over NH for the operational and parallel runs.

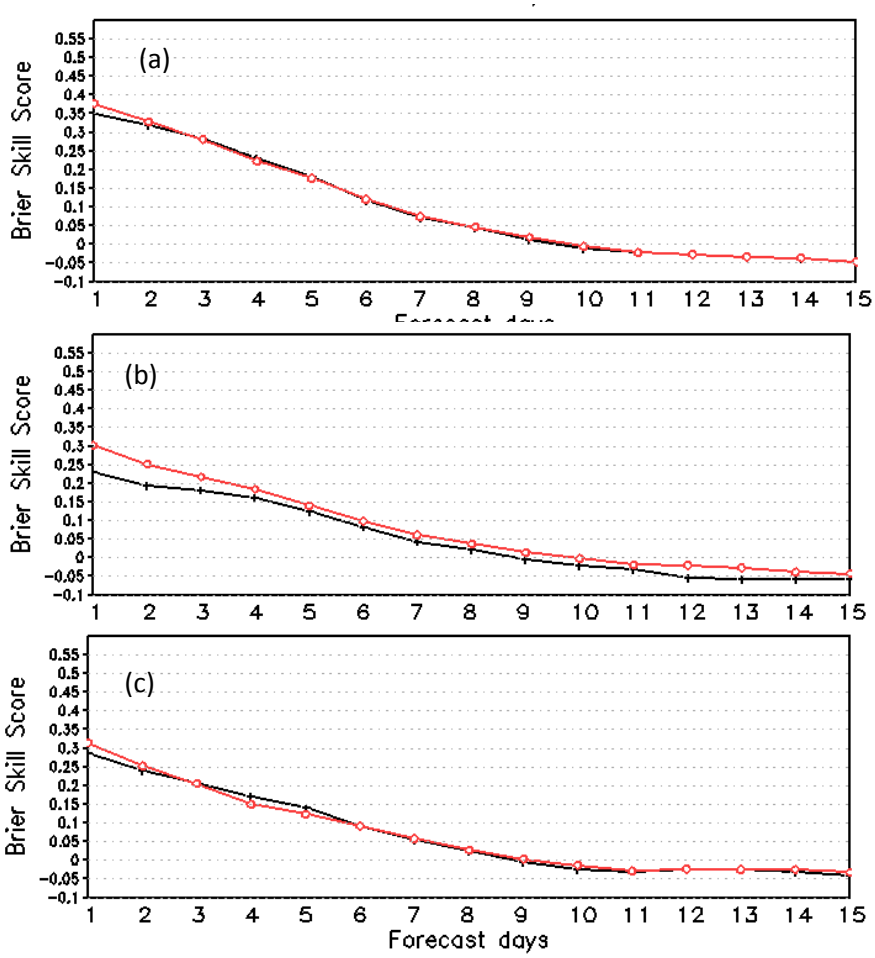
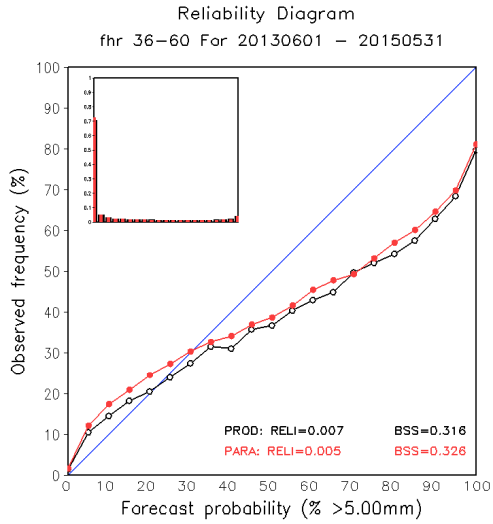
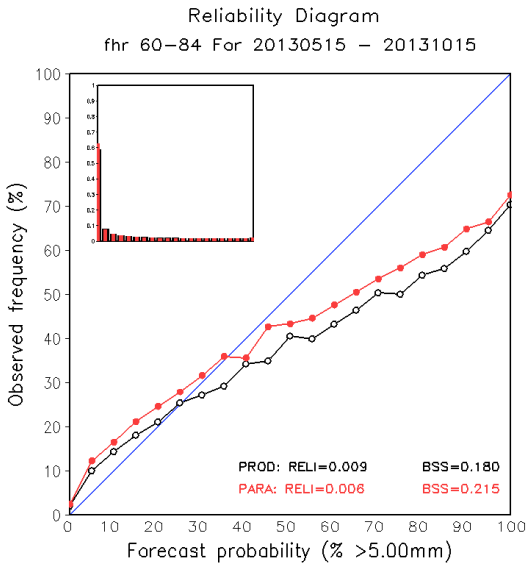
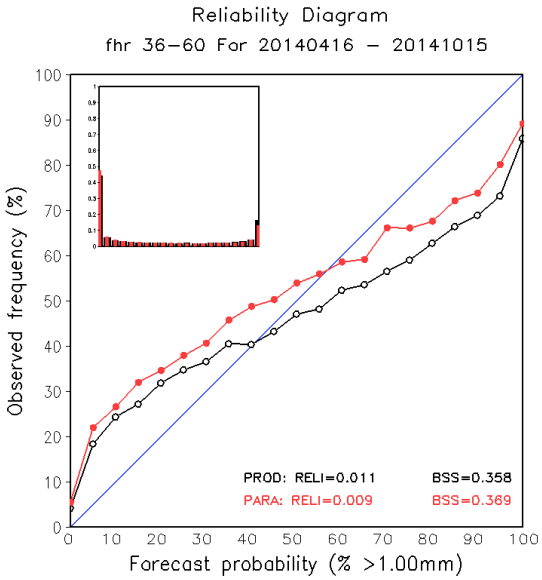


Fig. 9 BSS for the precipitation greater than 5mm/24 hours averaged over two years (a), from May 15 to Oct. 15 2013 (b) and May 15 to Oct. 15 2014.







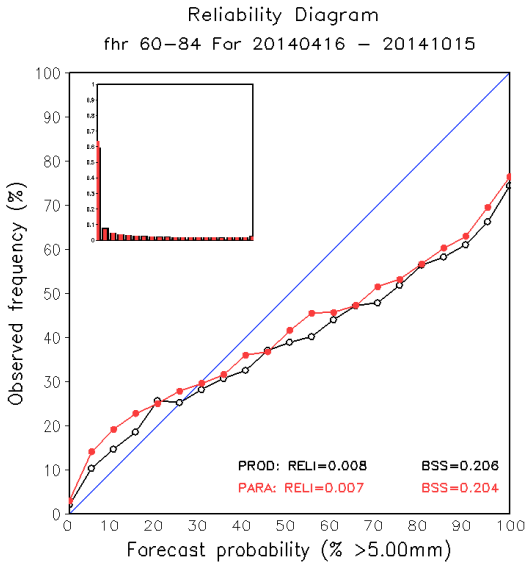


Fig.10 Reliability diagram of the 21 probability categories from 21-member ensemble. The upper left inset in each plot shows the proportion of cases in each probability category

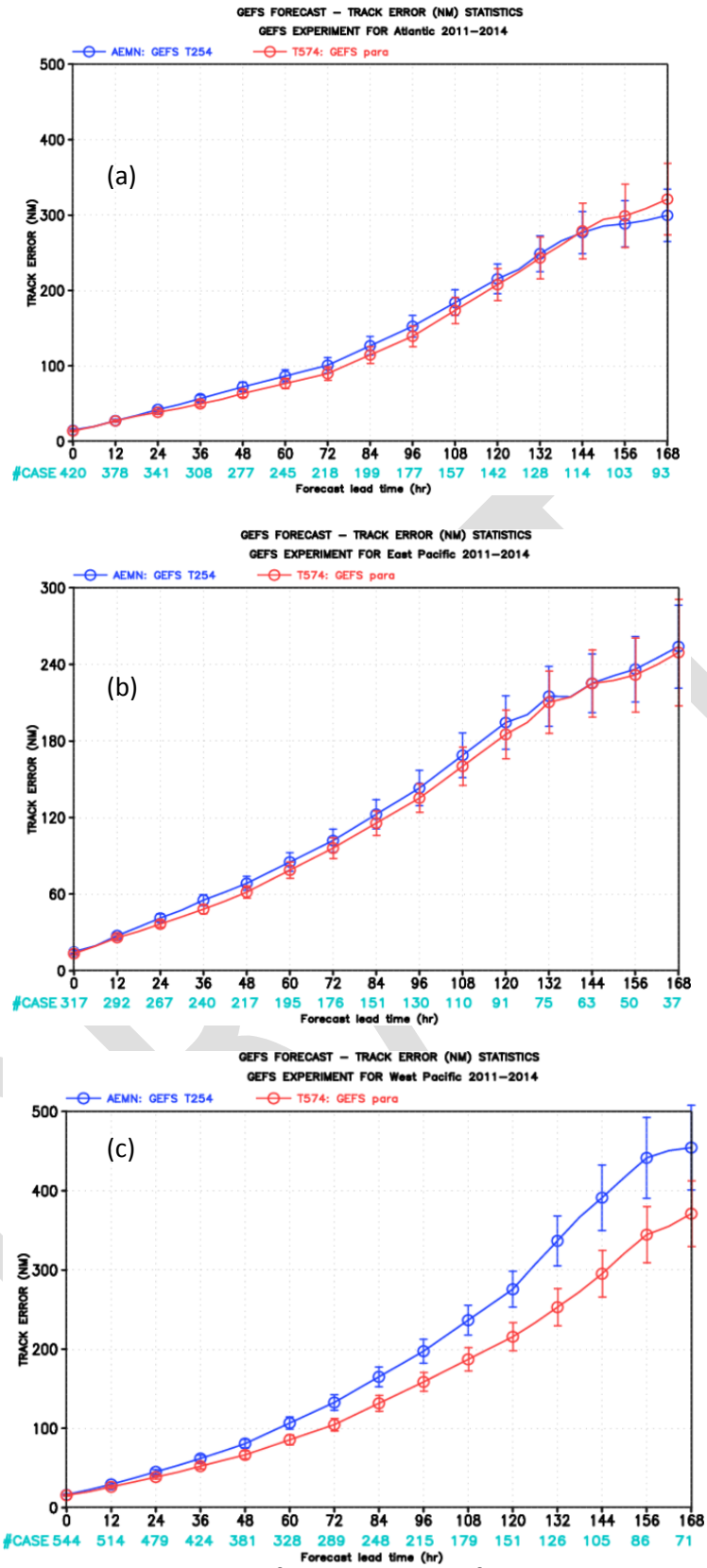


Fig. 11 Ensemble mean track forecast errors for 2011-2014 hurricane seasons over Atlantic (a), EP (b) and WNP (c).

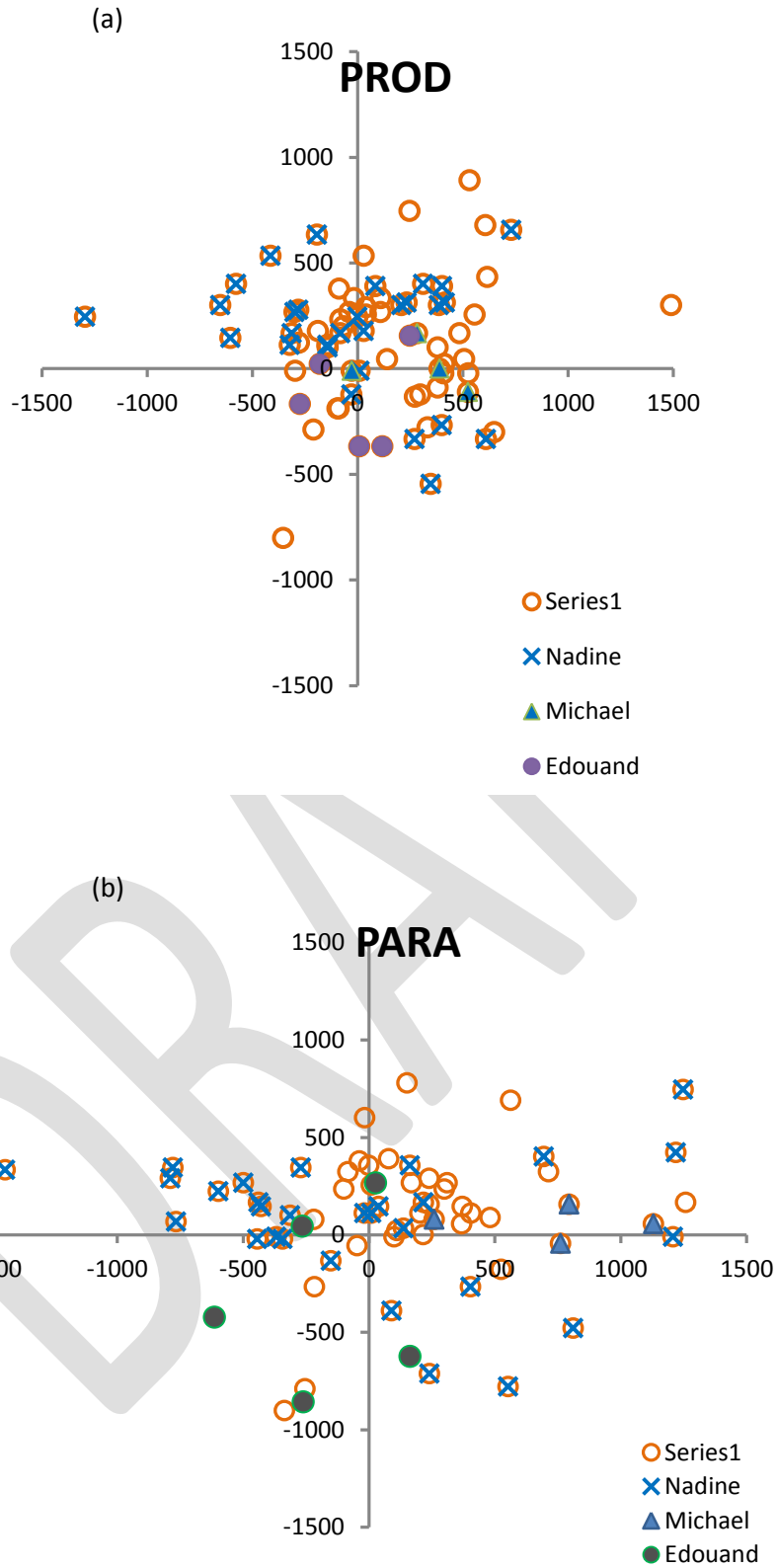


Fig. 12 The deviation of the 6-day forecast TC locations from the OBS over Atlantic (2012-2014) for the production (a) and parallel (b).

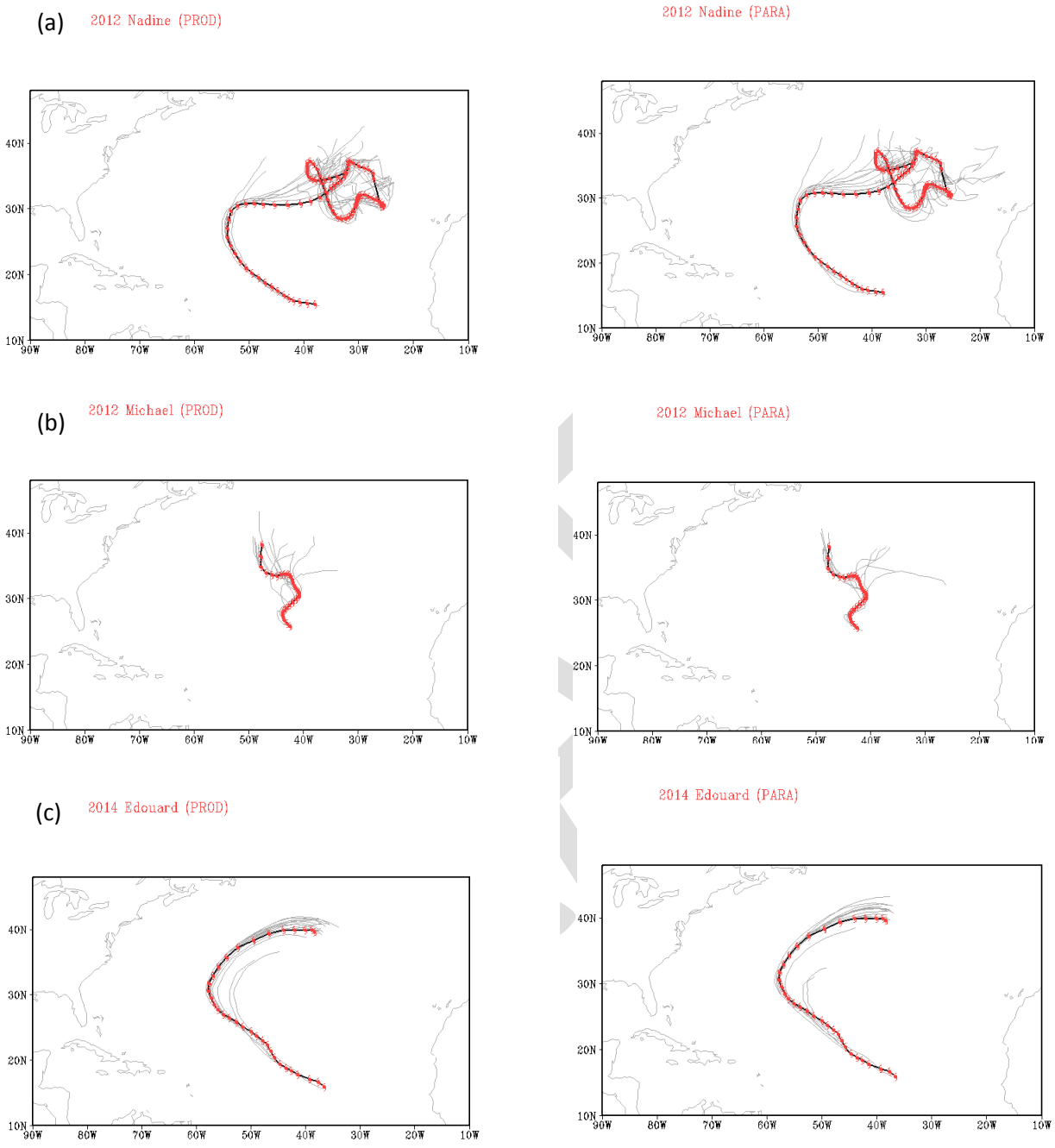


Fig. 13 The observed TC tracks (hurricane signal) and the 168-hr ensemble-mean forecast tracks for Hurricane Nadine (a), Michael (b) and Edouard (c) in the production (left panel) and the parallel run (right panel)