Comparison of the Ensemble Transform and the Ensemble Kalman Filter Initial Perturbation Schemes in the NCEP Global Ensemble Forecast System

Xiaqiong Zhou[1], Yuejian Zhu[2], Dingchen Hou[2], and Daryl Kleist[3]

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

3）Dept. of Atmospheric and Oceanic Science, University of Maryland, College Park, MD

To be submitted to Mon. Wea. Rev.

Corresponding author:

Xiaqiong Zhou, IMSG at Environmental Modeling Center/NCEP/NOAA

5830 University Research Court
College Park, MD 20740

Email: Xiaqiong.Zhou@noaa.gov

Abstract

Two ensemble initialization schemes, the ensemble transform with rescaling (ETR), an improved operational breeding method and the ensemble Kalman filter (EnKF), are compared under the NCEP operational environment for the global ensemble forecast system (GEFS). The ensemble mean forecast is verified by using the root mean square error (RMSE), the ensemble spread, and the pattern anomaly correlation (PAC). The continuous ranked probability skill score (CRPSS) is used to verify the probabilistic forecast. The comparison shows that the amplitude of initial perturbations in EnKF is generally larger than in ETR especially in the Southern Hemisphere (SH), but both fail to generate sufficient spread at the medium-range forecasting times. EnKF outperforms ETR in terms of CRPS scores in the Northern Hemisphere for the first week (NH), whereas the ensemble mean forecast is more skillful for ETR in the southern hemisphere (SH). No significant differences are found for the tropics between ETR and EnKF.

Similar experiments are performed with the stochastic total tendency perturbation (STTP) scheme, in which the total tendency of all model variables is perturbed to parameterize the uncertainty in the forecast model itself. The inclusion of STTP increases the ensemble spread for EnKF and ETR. Better spread-error relationships are obtained for ETR and EnKF over NH, but not for EnKF in the SH. The probability forecast scores remains significantly better in the NH for EnKF compared to ETR as in the experiments without STTP, but some degradation is found in the EnKF experiments in the SH due to an over-dispersive ensemble. The results indicate the tuning of either EnKF or STTP is required for SH when STTP is applied with the EnKF-based perturbations.

The ensemble mean tropical cyclone track forecasts in two tropical cyclone seasons are compared. The EnKF scheme has similar performance as ETR in tropical cyclone track forecast when the vortex relocation scheme is applied.

1. Introduction

Small errors in initial conditions can amplify and result in significant forecast errors during a forecast (Lorenz 1969; 1982). A feasible way to improve forecast skill is to use ensemble forecasting with different initial conditions sampling the error probability density functions (PDFs) of the analyzed atmospheric state (Epstein 1969; Leith 1974). More specifically, the analysis PDF must be defined given estimates of error in observed and background fields, and then a representative but finite sample suitable for evolution must be taken from this PDF. No consensus has been reached on how to accomplish these tasks.

Of all ensemble generation schemes, the breeding method is the most computationally inexpensive. It has been used for generating initial perturbations at the U. S. National Center of Environmental Prediction (NCEP) for the operational Global Ensemble Forecast System (GEFS) since Dec. 1992 (Toth and Kalnay 1993, 1997). A hypothesis of the breeding method is that the important part of analysis errors is the dynamically constrained part contributed by errors of the forecast background. This method dynamically recycles the perturbations to simulate the development of growing errors in analysis cycles. After an infinitely long breading time and with the use of infinitesimal amplitude, the bred vectors (BV) should be identical to leading Lyapunov vectors (Corazza et al., 2001; Toth and Kalnay, 1993, 1997; Cai et al., 2003; Kalnay 2001), which provide estimates of fastest sustainable growth and thus represent probable growing analysis errors. The ensemble mean based on the BV method gives a better forecast than the control forecast as long as the ensemble represents the uncertainty in the control analysis.

In practical applications, however, BVs cannot accurately represent the true uncertainty in the analysis and have a tendency to produce analysis perturbations whose variance is

concentrated in considerably limited eigen-directions (Wang and Bishop 2003; Wei et al. 2006). The absence of variability in some eigen-directions is undesirable given that the ensemble perturbations are already too few to span all the directions in which analysis error variability really exists. Ensemble transformation (ET) with rescaling (ETR) was introduced in NCEP global forecast system in 2005 (Wei et al. 2006, 2008), in which the forecast perturbations from breeding cycling are transformed into analysis perturbations by multiplying a transformation matrix. The matrix is chosen to ensure that the ensemble-based representation would be consistent with a user-provided estimate of analysis error covariance. The ensemble variance is maintained in as many directions as possible within the ensemble subspace. Meanwhile, the ET method produces simplex, not paired members. Wei et al. (2008) compared the probabilistic scores of the BV and ETR. They found that the ETR had better performance than the BV.

A hybrid variational-ensemble data assimilation system based on the Gridpoint Statistical Interpolation (GSI; Wu et al. 2002; Kleist et al. 2009) and the ensemble Kalman filter (EnKF) has been developed and was successfully implemented operationally for the NCEP Global data assimilation System (GDAS) using the Global Forecast System model in May 2012 (Whitaker and Hamill 2002; Whitaker et al. 2008; Wang et al. 2013, Kleist and Ide 2015). In the hybrid GSI-EnKF, the flow-dependent background covariance based on the ensemble of short-range forecasts from EnKF is incorporated with the static background error covariance of GSI (a 3DVAR algorithm) during the data assimilation. The ensemble obtained at the end of the EnKF assimilation can be used as the initial conditions for global ensemble prediction. The success of EnKF in NCEP GDAS provides alternative ensemble initial conditions for the operational GEFS. It is necessary to choose one which will provide better medium-range forecast performance from the existing operational schemes.

Quantitative comparisons of the breeding scheme and EnKF perturbation strategies have been performed in simplified environments and in operational environments. Buizza et al. (2005) compared the ensemble forecast systems of European Centre for Medium-Range Weather Forecasts (ECMWF), the Meteorological Service of Canada (MSC), and NCEP. These three operational systems use the singular vector, the breeding of growing mode (BGM) and EnKF schemes respectively. No conclusion is reached as to the relative performances of the different initialization schemes, as they could only measure the overall quality of the three systems. Descamps and Talagrand (2006) and Bowler (2006) found that the EnKF performs better than the breeding method but their conclusions are limited to simple models.

The goal of this study is to perform a clear comparison of the current operational ensemble scheme (ETR) with the EnKF in the NCEP operational environment. The results in this study will provide essential guidance for the next GEFS implementation. The general experimental setup of the study is described in section 2. An overview of verification scores that are utilized in the study is presented in section 3. The verification scores from the two experiments covering two summer seasons is presented in sections 4 and 5, and ensemble track verification is summarized in section 6. The conclusions are given in section 7 together with a general discussion.

2. Experimental setup and verification method

a）Experimental setup

The forecast model used is the NCEP Global Forecast System (GFS), version 9.01 (http://www.emc.ncep.noaa.gov/?branch=GFS&tab=impl). Two sets of experiments, ETR and EnKF, use the same GFS model and same control analysis in order to ensure that the differences

between the experiments result only from the differences in the initialization methods. The ensemble for each set consists of 20 ensemble members and a control run. The analysis is truncated from the T574L64 analysis generated by the hybrid GSI and EnKF system and interpolated to a lower resolution (T254L42) which then also serves as the control member for the ensemble. The initial conditions for each ensemble member are created by applying a small perturbation to the control analysis by using either ETR or EnKF. For the ETR experiments, the generation of the initial perturbations follows the NCEP GEFS operational scheme (see Wei et al., 2006 and Wei et al., 2008 for the description of the ETR scheme in detail). In the EnKF experiment, 6-h ensemble forecast perturbations initialized from EnKF analyses are used as initial perturbations since only the EnKF from the previous cycle is available at the time when GEFS starts in the NCEP operational environment. This is due to the fact the EnKF is only run as part of the late cycle within the global data assimilation system for the purposes of prescribing background error covariances in the cycle that follows. In other words, the prior of the EnKF is utilized instead of the posterior to generate the initial ensemble for the GEFS. The model is integrated once daily (0000 UTC) at T254L42 resolution for 0-192 h lead times and lower resolution (T190L42) for 192-384 h lead times due to computational constraints.

Given that uncertainties in the forecast also arise from uncertainties in the model formulation, the performance of these initialization schemes is also compared by including model perturbations. The NCEP GEFS uses the Stochastic Total Tendency Perturbation (STTP) scheme to represent such unpredictable model uncertainties, in which stochastic forcing is added to the total tendencies of ensemble perturbations for the model variables (temperature, specific humidity and winds, Hou et al. 2006, 2008). The scheme is applied every six hour after forecast output is created. The total time tendency (from all physical and dynamical processes) for each

variable is perturbed by a random factor and rescaled to be size, region, and lead time appropriate. For example, the extratropics have larger perturbations in the tropics and perturbations should be larger with longer lead time. The calculation of the time tendency perturbations is done at the same time across all ensemble members, with the perturbations made statistically independent of each other before being applied to the variables.

Treatment of model errors has also been an area of research within the context of the EnKF (Whitaker and Hamill 2002, 2012; Whitaker et al. 2008; Houtekamer et al. 2005 and 2009; Zhang et al. 2004; Meng and Zhang 2008). Ideally, the ensemble providing background-error covariances should sample all sources of error in the forecast environment, including sampling error due to limitations in ensemble size and errors in the model itself. The EnKF may not give enough weight to observations when ensemble-estimated covariances are underestimated. This problem can progressively become worse in time if unaccounted for, potentially leading to filter divergence in which the ensemble variance becomes unrealistically small and the filter trusts its own forecast and ignores the information given by the observations.

In order for the EnKF to perform optimally in data assimilation, model errors and other error source were treated by using ensemble covariance inflation (e.g, Whitaker and Hamill 2002) and additive noise on posterior ensemble (e.g., Whitaker et al. 2008, Houtekamer et al. 2005). The particular versions of multiplicative inflation (relaxation to prior spread) and additive inflation (using lagged forecast pairs) is described in more details in Whitaker and Hamill (2012).

An interesting question that arises from the special treatment of model error in EnKF is whether the ensemble initialized with EnKF perturbations can produce sufficient spread for medium-range forecasts without model perturbations. If not, it is interesting to explore how well

it works with STTP. Two groups of experiments have been conducted. The performance of ETR and EnKF without STTP is examined during the periods of July 1- October 25 2011. Similar experiments are performed for 2012 summer except with STTP turned on.

Another component of the operational GEFS is a tropical cyclone relocation scheme. The tropical cyclone component is separated from the environment field using the scheme of Kurihara et. al. (1993 and 1995) and relocated to the observed TC location center when calculating TC perturbations. The TC relocation is always included in the ETR experiments for all ensemble members and the control as in operations. The impacts with and without this scheme are examined for the EnKF run.

b) Verification Metrics

Forecast and analysis fields are interpolated onto a common regular 2.5 by 2.5 degree lat-lon grid to assess the quality of ensuing ensemble predictions. Ensemble forecasts are verified against hybrid GSI and EnKF GFS analysis. The quality of forecasts is measured by the NCEP ensemble verification system (Zhu et al. 1997; Toth et al., 2003, 2006; Zhu and Toth 2008) which includes calculation of traditional verification measure such as Root Mean Square Error (RMSE) and Pattern Anomaly Correlation (PAC) for the ensemble mean and the measures related to two important attributes: the reliability and resolution (Toth et al., 2003, 2006), such as Continuous Ranked Probability Skill Score (CRPSS) and Ranked Probability Skill Score (RPSS) for the probability forecasts.

RMS errors of the ensemble mean measure the distance between forecasts and analyses (or observations). SPRD (ensemble spread) is calculated by measuring the deviation of ensemble forecasts from their mean. It is expected that SPRD has the same size of RMS error at the same

lead time in a good forecast when the ensemble forecast can represent full forecast uncertainty (Zhu 2005; Buizza et al. 2005).

PAC measures the ability of the ensemble mean to forecast the variation of weather patterns, which is defined as the correlation between the predicted anomaly and the observed anomaly with respect to their corresponding climatology. The maximum value of 1.0 indicates a perfect depiction of the patterns.

CRPS is used to measure the reliability and resolution of ensemble-based probabilistic forecasts by calculating the distance between the predicted and the observed cumulative distribution functions of scalar variables in terms of 10 climatologically equally likely intervals determined at each grid point separately (Toth et al. 2002). For statistics over a long period, CRPS is equivalent to Mean Absolute Error (MAE) for a deterministic forecast and it is very similar to RPSS. Therefore, we consider it possible to use either one of these two measures.

Tropical cyclone tracks are estimated by comparing forecast and observed tracks based on the best track dataset in the Atlantic, and western North Pacific, and eastern North Pacific basins. Only ensemble mean track error and track spread are verified.

The paired block bootstrap algorithm (Hamill, 1999) is used to estimate the statistical significance of differences in scores. The technique generates multiple datasets from the available data by selecting random samples with replacement, allowing one to estimate the statistical parameters regardless of the distribution of the underlying data. In this study, 95% confidence interval is computed from a bootstrap resampling using 20000 random samples of more than 90 cases.

3. Initial perturbations of ETR and EnKF

Fig. 1 shows the initial error variance of total energy averaged in the NH, SH and tropics for EnKF and ETR.

$$TE = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{1}{2}(u_i'^2 + v_i'^2 + \kappa t_i'^2)}$$

where $u_i'^2$, $v_i'^2$ and $t_i'^2$ are deviations corresponding to the ith member for the wind components and temperature. The quantity $\kappa = c_p/T_r$ equals to 4.0 J kg^{-1} K^{-2} in which c_p is the special heat at constant pressure and T_r is the reference temperature. EnKF 6-hour forecast perturbations used as the ensemble initial perturbations are compared with the analysis perturbations after multiplicative inflation and additive inflation. The inflation increases the ensemble spread but also introduces imbalance in the ensemble of analyzed states. The ensemble spread decreases significantly in 6 hour forecasts although it remains larger than the spreads before additive inflation is applied. The application of the inflation plays a certain role in parameterizing other uncertainty sources and stabilizing the assimilation method, but is not very efficient. Overall the EnKF 6-hour forecast perturbations are larger than ETR perturbations (except in tropical lower levels), which is consistent with the nature of these two techniques. The breeding cycling in ETR is explicitly designed to generate perturbations that contain fast-growing modes corresponding to the evolving atmosphere and the amplitude of initial perturbations is small and expected to grow fast with forecast lead times. In contrast, large amplitude perturbations are favorable for data assimilation to avoid filter divergence. EnKF ensemble not only samples analysis errors but also the other sources of error in the forecast environment. Meanwhile it captures the entire spectrum of analysis errors, many of which will project onto neutral or decaying modes. In addition, larger perturbation is expected when the prior ensemble perturbations instead of the posterior are used to generate initial perturbations.

There are also significant differences in the geographical distributions of the initial perturbations between ETR and EnKF (Fig. 2). The perturbation amplitude in ETR increases poleward in the Northern Hemisphere, which is consistent with its regional rescaling mask (Ma et al. 2014), (Fig. 2a). Smaller perturbations are found over the continent than the ocean. The land-sea contrast is usually considered as a result of the geographical inhomogeneity of the observation density distribution. Smaller perturbations are also observed over South Africa, South America and Australia than over the ocean. The geographic distribution of EnKF spread is similar with that of forecast ensemble, which is not only due to that the initial perturbations in the EnKF experiment come from 6-hour short-range forecasts, but also the posterior ensemble spread is relaxed to the prior during EnKF data assimilation. The land-sea contrast in EnKF is not evident (Fig. 2b). The spread distribution has same pattern as that of the ensemble mean. In other words, the initial perturbation amplitude is locally maximized in the regions where mean kinetic energy is high. For the NH, maximum perturbation centers are corresponding to the storm tracks over the Pacific and the Atlantic. Initial perturbations are zonally symmetric in the observation-scarce Southern Hemisphere and have large amplitude at strongly baroclinic high latitude around 60^{o} S.

The difference of the initial perturbation amplitude between these two experiments has a clear zonal structure (Fig. 2c). EnKF has much larger initial perturbations in middle-latitude baroclinic zones but smaller in polar region. We also note that the perturbations in baroclinic middle latitude grow rapidly in ETR. The difference between EnKF and ETR decreases with forecast time. The spatial patterns of the perturbation for ETR and EnKF at 96 hr are very similar except for the larger amplitude of EnKF in baroclinic zones (Fig. 3).

Fig. 4 examines the mean eigenvalue spectra of the ensemble covariance matrix. For each ensemble generation technique, the heights of the 19 bars correspond to 19 seasonally averaged eigenvalues of the geopotential height error covariance normalized by global mean for initial perturbations. The sum of these eigenvalues is related to the spread of the ensemble at this time. There are large amounts of ensemble variance present in all 19 uncorrelated orthogonal directions of ETR and EnKF, but the spectrum of the EnKF eigenvalues is flatter than the one from ETR. The first eigenvalue being much larger than all the others in ETR, indicating that there is a great deal of similarity between the ensemble members. As an extension of breeding method, dynamic cycling remains in ETR. The breeding technique would eventually only maintain error variance in the direction corresponding to some leading modes. As shown by Wei et al. (2006) and Ma. Et al. (2014), the ET method maintains the variance in as many directions as possible within the ensemble subspace and performs better than simple breeding, but the perturbations are orthogonal only in the infinite number of ensemble members. The small value in the trailing eigenvector indicates it would be inefficient to improve forecast with more ensemble members.

4. Experiments without STTP

All scores that are computed for 2011 summer can be found in http://www.emc.ncep.noaa.gov/gc_wmb/xzhou/EnKF_ETR_2011_Summer.HTML. The parameters evaluated include the geopetential height at 500 hPa and 1000 hPa pressure levels, wind fields at 10-m, 850 hPa and 250 hPa pressure levels, temperature at 2-m height, 850hPa pressure level. NH refers to the area north 20^{o} N to 80^{o} N, SH refers to the south of 20^{o} S to 80^{o} S and Tropics is from 20^{o} S to 20^{o} N. We will primarily show the forecast scores of 500 hPa geopotential height for NH and SH and zonal wind component at 850 hPa winds for tropical

region due to the absence of geostrophy and the low variability of the mass field (temperature and geopotential). The results for other parameters will be also summarized.

a) Skill of ensemble mean and spread

The greater the ability of an ensemble forecast to account for the likely errors in the forecast, the greater the skill of the mean of that ensemble forecast. The skill of the ensemble mean can be taken as a (crude) measure of the quality of the ensemble. Figure 5a shows the RMS error of the 500hPa height ensemble mean forecast over NH for the two perturbation strategies. RMSE in both ETR and EnKF rapidly increases in the first week and becomes almost saturated after day 10. RMSE is significantly smaller in ETR than in EnKF for the first two days but become similar at longer lead times. The advantage of ETR is more evident over SH than over NH. The ETR ensemble mean is more accurate than that of EnKF in the first four days with some suggestions that ETR slightly superior at longer lead times (Figure 5b).

Similar performance can be seen in RMSE of geopotential height at 1000 hPa (not shown). ETR has significantly smaller RMSE than EnKF in both NH and SH for the short-time forecast. The advantage preserves in longer lead times over SH but the differences become insignificant. EnKF is significantly better in the ensemble forecast of the low-level wind at 850 hPa and surface from day 1 to day 3 over NH (not shown). No significant difference is found from the temperature at the low levels (850 hPa and 2m) the wind fields at 250 hPa between these two experiments over NH. On the contrast, the degradation of EnKF is generally significant for all variables in the first 2-4 days over SH.

Ideally, the spread of ensemble forecast perturbations is equal to the RMS error of ensemble-mean forecast at all lead times. Having an underspread and overspread ensemble is not desirable for an ensemble forecasting system. Figure 5 shows that the growth of ensemble

spread is slower than that of ensemble mean forecast errors in both ETR and EnKF. For the ETR, the ensemble spread of geopotential height at 500 hPa as well as other variables over SH and NH is generally under-dispersive (Figs. 5a and 5b). EnKF has greater spread than ETR at the initial time. Over-dispersion in EnKF is found in the first several days and then gradually changes to under-dispersion with lead forecast time. The deviation between EnKF and ETR increases with lead time in the first week but decreases gradually and become negligible in the second week.

A slight overspread in the short-time forecast as in the NH is desirable. The ensemble mean forecast error is under-estimated when same model analysis is used for verification. However, the over dispersion in SH is considered to have a negative impact on the ensemble mean forecast. The geopoetential height perturbation at 500 hPa in EnKF is over-dispersive in SH in the first week, which is more evident and lasts longer than in NH (Fig. 5a and Fig. 5b). Over-dispersion is also found in the first 2-4 days in other variables over SH but not over NH. The presence of too large spread in the SH is consistent with the degradation of EnKF.

The degradation of EnKF is also found in the SH in terms of PAC. Figs. 5c and 5d shows the time evolution of the PAC for the geopotential height at 500-hPa in the NH and SH. PAC is close to perfect with values of 1 at the initial time and then decreases with the forecast lead time. No difference is observed in the NH for the geopotential height at 500 hPa (Fig 5c) as well as other variables (not shown). The PAC score is slightly degraded in EnKF for the SH. The degradation is not always significant but can be found at not only 500-hPa (Fig. d) and 1000-hPa geopotential height but also at the wind fields at 10 m, 850 hPa and 250 hPa levels and low-level temperature at 2 m and 850 hPa (not shown).

Fig. 6a shows the RMSE of 850 hPa zonal wind over tropical region. RMSE is significantly smaller in EnKF from day 1 to 3 than in ETR and the scores become similar beyond day 4. The spread in both EnKF and ETR is generally much smaller than RMSE, but large overdispersion is seen in ETR at the initial time which decays rapidly. In ETR, initial perturbation amplitude for all vertical layers is rescaled with a factor based on a 500 hPa kinetic energy regional mask. An ad-hoc tuning with an inflation factor is applied to obtain sufficient ensemble dispersion in low levels. This tuning strategy leads to over dispersion in low levels in the tropics, which decays rapidly with integration. This tuning will not be employed in a future implementation.

Fig. 6b shows that PAC for 850 hPa zonal wind over the tropics is very similar in these two experiments. No significant difference is found for other variables in the tropics (not shown).

For other variables in the tropics, no significant difference is found between EnKF and ETR except some suggestions that EnKF is more skillful from day 1-3 for the horizontal wind components (not shown). The ensemble spread in both ETR and EnKF is much smaller than the ensemble mean forecast error for all variables.

b) Ensemble probabilistic forecasts

CRPS is computed with a reference forecast based on the climatological distribution. The maximum value is one, associated with a perfect forecast while zero value indicates that the forecast is no better than climatology, the reference forecast. For 500 hPa geopotential height, EnKF CRPS is significantly smaller over NH than ETR in the first two days but becomes larger in longer lead time (Fig. 7a). For other variables in the NH the performance of EnKF is better than ETR at all lead times while CRPS is generally significantly higher in the first week (not shown). On contrast, there is no difference seen in other parameters for the SH except geopotential height at 500 hPa (Fig. 7b) and 1000 hPa in a short time range (not shown). For the tropics, EnKF performs better in the first 2-3 days and the improvement is statistically significant, but the scores generally become similar for longer lead times (Fig.6c).

Better performance of EnKF in probability forecasts is consistent with the results in spectrum analysis in Fig. 3. The flat spectrum in EnKF indicates a better estimate of analysis error variance than ETR, which could result to a higher score in probability ensemble forecast.

5. Experiments with model errors

The results in previous section show that the spread growth in both ETR and EnKF is not as fast as the growth of forecast error. Under-dispersion is quite common in both ETR and EnKF in long lead times. Large initial spread in EnKF does not generate enough spread in the medium-range forecast. In this section, the STTP model perturbation scheme that is used in the current operational GEFS is included for the 2012 summer experiments.

a) Ensemble mean and spread

The performance of EnKF and ETR for 2012 summer is similar with that for 2011 summer in terms of the RMSE of 500-hPa geopotential height. ETR is found to be superior to EnKF in the short-range forecast but similar for longer lead times. STTP adds spread for both ETR and EnKF configurations. EnKF ensemble spread remains larger than that of ETR with slightly over-dispersive spread in the first four days.

Given that STTP is well tuned for ETR in GEFS, the relationship between the ensemble spread and RMS error is close to perfect for the ETR experiment (Fig. 8a and Fig. 8b). For EnKF, STTP increases the ensemble spread which is already overdispersive. It is not only seen in geopotential height, but also in the horizontal wind components at different vertical levels (not shown). As a result, the EnKF ensemble mean forecast is less skillful than that of ETR. The RMSE of ensemble mean is greater in EnKF than in ETR in all lead times. The differences between these two experiments pass bootstrap test at 95% significant level from forecast day 1 to day 8.

Note that there is a sudden decrease of ensemble spread at day 8 (Fig. 8) as a result of the change of model resolution from T254 to T190 for computational efficiency. The discontinuity of the ensemble spread is due to this spatial truncation. This does not affect RMSE since the ensemble mean produce smooth fields where small-scale features are absent.

Consistent with RMSE, the difference of PAC between ETR and EnKF becomes more evident than in the experiments without STTP. Fig. 8a shows that the PAC of 500 hPa geopotential height becomes slightly greater with lead time over NH in ETR than in EnKF. Similar phenomenon can also found in geopotential height at 1000 hPa and the wind fields at different vertical levels (not shown). The better performance of ETR is more evident in the SH.

The PAC of 500 hPa geopotential height is significantly greater in ETR expect the forecast beyond day 14 (Fig. 8b). The advantage of ETR is also found in other verified variables except in 2m temperature. In the tropics, ETR and EnKF perform similarly (not shown).

In tropics, STTP does not increase the ensemble spread significantly except for geopotential height (not shown). A group of sensitivity experiments is performed and the results show that overdispersion in geopetential height results from the perturbed surface pressure (not shown). The performance of ETR and EnKF in 2012 summer over in the tropical region is similar with 2011 summer. The RMSE in these two experiments is similar (Fig. 9a), with the exception of the degradation of the RMSE in ETR for the low-level temperature and horizontal wind components in short-range forecasts (not shown). As discussed previously, the degradation of ETR is a result of the ad-hoc low-level inflation.

b) *Ensemble probabilistic forecasts*

The CRPS scores in ETR and EnKF for 500 hPa geopotential height over NH are similar (Fig. 10). EnKF outperforms in low levels for the first week, including 1000hPa geopotential height, temperature and wind fields near the surface and at 850 hPa. The advantage of EnKF is also found for wind probability forecasts in the tropics although the difference in CRPS is small (Fig. 8c). The results are opposite in the SH. The CRPS values in ETR are greater than in EnKF. The skill in ETR is generally significantly higher before day 8.

6. Tropical cyclone track forecast

The ensemble-mean forecast of TC track over the Western and Eastern North Pacific and North Atlantic is verified for two TC seasons (Fig. 11). For summer 2011, the impact of the TC

relocation scheme on TC track forecast is examined for the EnKF experiment, in which the TCs in each ensemble member are relocated to the observed centers. The ensemble spread at the initial time is larger in EnKF without TC relocation, but its growth rate is smaller than in ETR. The spread of the ensemble forecast in EnKF become similar with those in ETR after 72 hrs. The EnKF ensemble mean forecast exhibits larger forecast errors than ETR at all forecast time, with the differences being significant at 12 hr, 48 hr, and 120 hr

The degradation of storm track forecast in EnKF is considered as a result of the absence of TC relocation. When the EnKF 6hr-forecast perturbations are centered on analysis without specific procedure for tropical cyclones, tropical cyclones in ensemble members are probably not perturbed appropriately if the tropical cyclone center of the EnKF 6hr ensemble mean deviate from that in the control. In addition, Fig. 11 shows that the initial ensemble mean TC location has larger deviation without relocation. The initial track error and spread decreases after the TC relocation procedure is performed. The skills of TC track forecast are comparable with those in ETR and the track forecast errors decrease.

For 2012 summer, The TC relocation scheme is performed in both EnKF and ETR. There is no significant difference in the ensemble mean TC track forecast error. Different from 2011 summer, the track spread is comparable with track errors. A common feature for these two seasons is that the growth rate of track spread in EnKF with TC relocation is slightly greater than in ETR.

7. Summary and discussion

Two initialization schemes available to generate initial ensemble perturbations for the global ensemble forecast system (GEFS) in the NCEP operational environment are compared. The current operational GEFS uses the ensemble transform technique (ET) to generate initial

perturbations, which is an improved version of the breeding-vector technique (BV-ETR) implemented since 2005. An EnKF was implemented into the NCEP data assimilation system in 2012. The EnKF short-range ensemble forecasts provide flow-dependent ensemble covariances for the data assimilation system. A question that arose from the implementation of EnKF is whether the ensemble perturbations generated by EnKF can replace ETR to generate the initial ensemble for the global medium-range ensemble forecasts. The main purpose of this study is to perform a comprehensive comparison between these two initialization schemes.

Forecast errors not only come from initial conditions but also from model uncertainties. STTP is one of the important components in GEFS, which generates perturbations for model variable tendency to simulate the uncertainty existing in the model formulations. The initialization schemes are compared with and without STTP.

The comprehensive comparison shows that EnKF is comparable with ETR without STTP except that the ensemble mean forecast is slightly degraded in the SH. The EnKF performs better in terms of CRPS and ensemble-mean RMSE in the first week for some variables. Over dispersive ensemble spread is found in EnKF for the first several days, which leads to some degradation in the EnKF. Although the amplitude of initial perturbations in EnKF is much larger than ETR especially in the SH, the spread growth is slower than the growth of forecast error. Under-dispersion is common in both ETR and EnKF for longer lead times.

EnKF remains superior to ETR in the NH with STTP in terms of CRPS, similar to the group of experiments without STTP. Nevertheless, the application of STTP increases ensemble spread considerably in the SH, which results in significant degradation not only for the ensemble mean forecast but also for CRPS. In the tropics, CRPS becomes significantly higher in EnKF than ETR.

Tropical cyclone track errors are verified for two summer seasons. Inclusion of a tropical cyclone relocation scheme is beneficial to improving ensemble-mean track forecasts for EnKF. With the application of the tropical cyclone relocation scheme in both ETR and EnKF, tropical cyclone track forecasts are found to be similar.

STTP used in the current operational GEFS is tuned well with ETR, which not only improve the spread-error relationship but also improve the ensemble forecast skills for ETR. However, including STTP is beneficial for EnKF. Consistent with the nature of these two schemes, the amplitude of the initial perturbations is generally smaller in ETR than EnKF. The inclusion of STTP in EnKF results in overdispersive ensemble spread thus degrades the performance of EnKF. On the other hand, the forecast ensemble initialized by EnKF perturbations cannot generate sufficient ensemble spread in the medium- range forecasts without model perturbations.

Given that the performances of EnKF with and without STTP are not as good as the current operation GEFS, one solution is to rescale the EnKF perturbation size for GEFS. A set of experiments was performed and presented in a separated paper (Ma et al. 2014). Another solution is to reduce EnKF perturbation size in hybrid data assimilation cycles with the precondition of no negative impact on data assimilation. It is worth noting that EnKF in the next implementation will reduce additive perturbations significantly and add stochastic physics in model. The stochastic representation is made up of three components; 1) stochastic kinetic-energy backscatter (SKEB), 2) perturbed boundary-layer humidity (SHUM) and 3) stochastically perturbed physics tendencies (SPPT). The performance of the GEFS by using the updated EnKF perturbations will be discussed in another paper. The impact of replacing the current STTP with a combination of these three stochastic schemes in GEFS is also under assessment. EnKF background (prior) ensemble instead of posterior ensemble is used as the initial

perturbations in the EnKF experiments. The former represents the uncertainties of the background in forecast environment instead of the analysis after assimilation. Better performance is expected when the posterior ensemble is used, but this is not practical in the current NCEP operational environment. The potential alternative to the current configuration is to move the EnKF to the GFS cycle or to run a separate EnKF as part of the early cycle with a reduced set of observations. However, the corresponding cost-benefit analysis is required for practicality.

References

Bowler, N., 2006: Comparison of error breeding, singular vectors, random perturbations and ensemble Kalman filter perturbation strategies on a simple model. *Tellus*, 58A, 538–548.

Buizza, Roberto, P. L. Houtekamer, Gerald Pellerin, Zoltan Toth, Yuejian Zhu, Mozheng Wei, 2005: A Comparison of the ECMWF, MSC, and NCEP Global Ensemble Prediction Systems. *Mon. Wea. Rev.*, 133, 1076–1097.

Cai, M., E. Kalnay, and Z. Toth, 2003: Bred vectors of the Zebiak-Cane model and their application to ENSO predictions. *J. Climate*, 16, 40-55.

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.

Descamps, L. and O. Talagrand, 2006: On some aspects of the definition of initial conditions for ensemble prediction. *Mon. Wea. Rev.*, 135, 3260-3272.

Epstein, E. S., 1969: Stochastic dynamic prediction. *Tellus*, 21, 739-759.

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.

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.

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.

Ma, J., Y. Zhu, D. Hou, X. Zhou, and M. Peña, 2014: Ensemble transform with 3D rescaling initialization method. *Mon. Wea. Rev.*, 142, 4053–4073.

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, 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, 2014: 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.*, revised manuscript under review.

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

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

Leith, C. E. 1974: Theoretical skill of Monte-Carlo forecasts. Mon. Wea. Rev., 102, 409-418.

Lorenz, E. N. 1969: The predictability of a flow which posses many scales of motion. *Tellus* 21, *289–307*.

Lorenz, E. N. 1982. Atmospheric predictability experiments with a large numerical model. *Tellus* 34, 505–513.

Meng, Z., and F. Zhang, 2008a: Tests of an ensemble Kalman filter for mesoscale and regional-scale data assimilation. Part III: Comparison with 3DVAR in a real-data case study. *Monthly Weather Review*, 136, 522-540.

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 Guid 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 system, 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.

Zhang, F., C. Snyder, and J. Sun, 2004: Impacts of initial estimate and observation availability on convective-scale data assimilation with an ensemble Kalman filter. *Monthly Weather Review*, 132, 1238-1253.

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.

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

Figure titles

Fig. 1 The vertical profiles of initial perturbation spread in terms of total dry energy in the ETR and EnKF experiments over a) NH, b) SH and c) Tropics. Three EnKF profiles represent the spread of EnKF perturbations after multivariable inflation (blue curves), addition inflation (red) and 6-hr forecast (green). The profiles are averaged from 1 July – 17 Oct. 2011.

Fig. 2 Seasonally-averaged ensemble spread of initial kinetic energy at 500 hPa for a) ETR, b) EnKF and c) the difference for 2011 summer.

Fig. 3 Seasonally averaged ensemble spread of 96-hr forecast kinetic energy at 500 hPa for a) ETR, b) EnKF and c) the difference for 2011 summer.

Fig. 4 Seasonal mean spectra of eigenvalues of ensemble based initial spread of 500 hPa Geopotential Height for 2011 summer normalized by the global spread for ETR and EnKF

Fig. 5 Ensemble mean rms error (solid) and ensemble standard deviation (dotted) for 500-hPa geopotential height over a) NH and b) SH. The verification scores for ETR (black) and EnKF (red) are averaged during the period from 1st July 1 – 17 th October 2011. The lower panels show the difference of RMSE (EnKF minus ETR, black lines) and bootstrap significant test (green bars). The difference is significant at 95% confidence level when value is beyond the bar ranges. c) and d) as same as a) and b) except for ensemble mean predicted pattern anomaly correlations of Z500

Fig. 6 a) Ensemble mean rms error and spread, b) ensemble mean pattern anomaly correlations averaged from 1st July to 17th Oct. 2011.

Fig. 7 Continuous ranked probability skill scores for (a) 500 hpa height over NH, (b) 500 hpa geopotential height over SH and 850 hPa horizontal wind u component over Tropics .

Fig. 8 As same as Fig. 5 except for the experiments with STTP on (averaged from 1st July to 30th Sep. 2012)

Fig. 9 As same as Fig. 6 expect for the experiments with STTP averaged from 1st July to 30th Sep. 2012.

Fig. 10 As same as Fig. 7 except for the experiments with SSTP averaged from 1st July -30 th Sep. 2012

Fig. 11 Ensemble-mean tropical cyclone track error (solid) and spread (dashed line) for a) 2011 and b) 2012 summer.

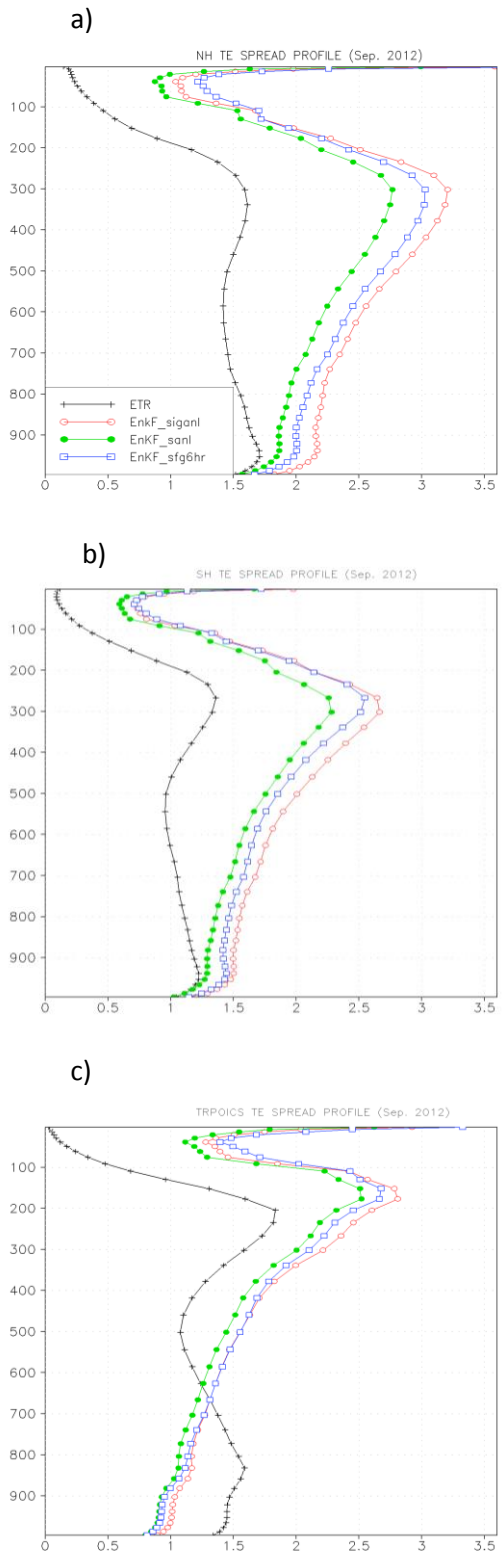


Fig. 1 The vertical profiles of initial perturbation spread in terms of total dry energy in the ETR and EnKF experiments over a) NH, b) SH and c) Tropics. Three EnKF profiles represent the spread of EnKF perturbations after multiplicative inflation (green curves), addition inflation (red) and 6-hr forecast (blue). The profiles are averaged from 1 July – 17 Oct. 2011.

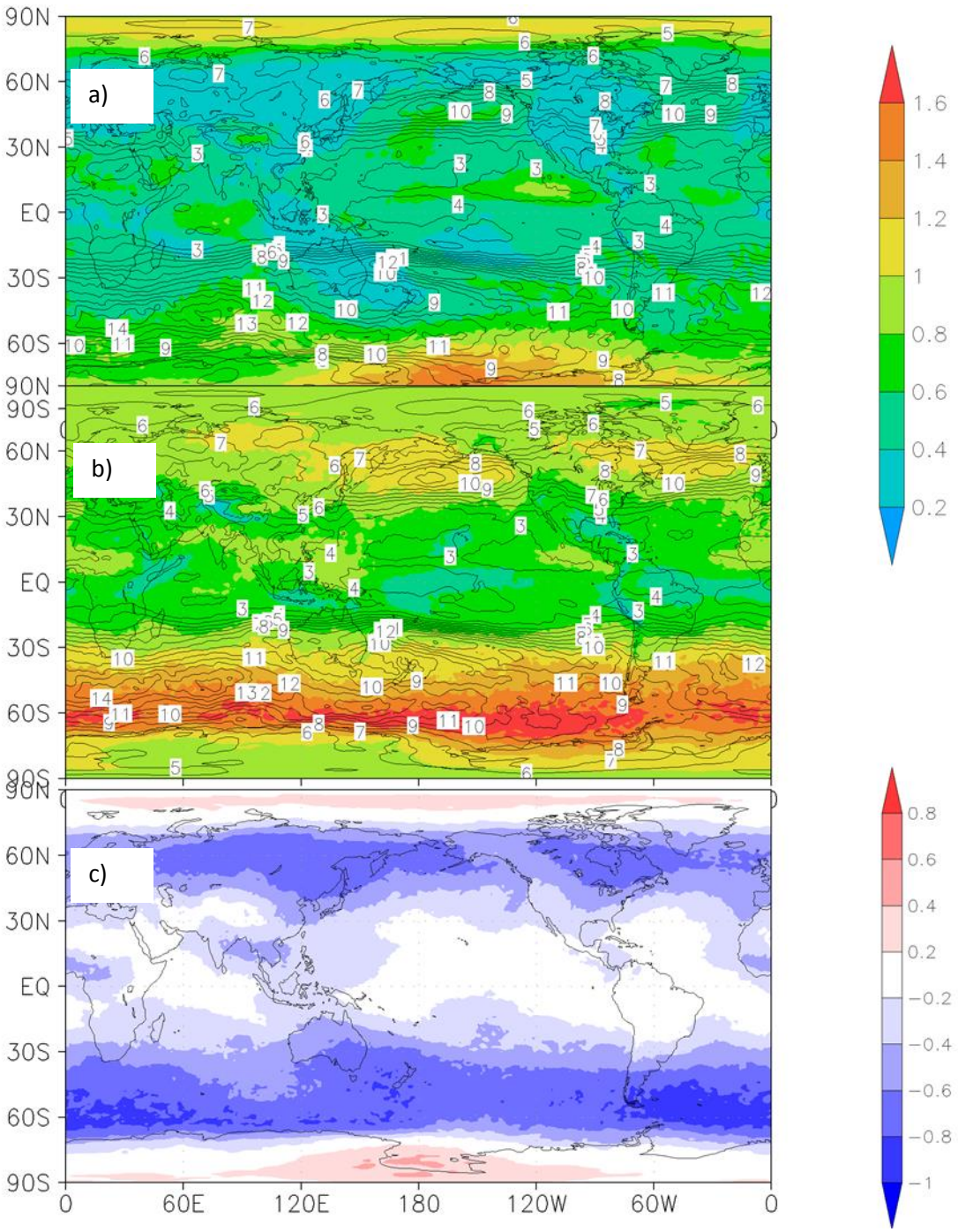


Fig. 2 Seasonally-averaged ensemble spread of initial kinetic energy at 500 hPa for a) ETR, b) EnKF and c) the difference for 2011 summer

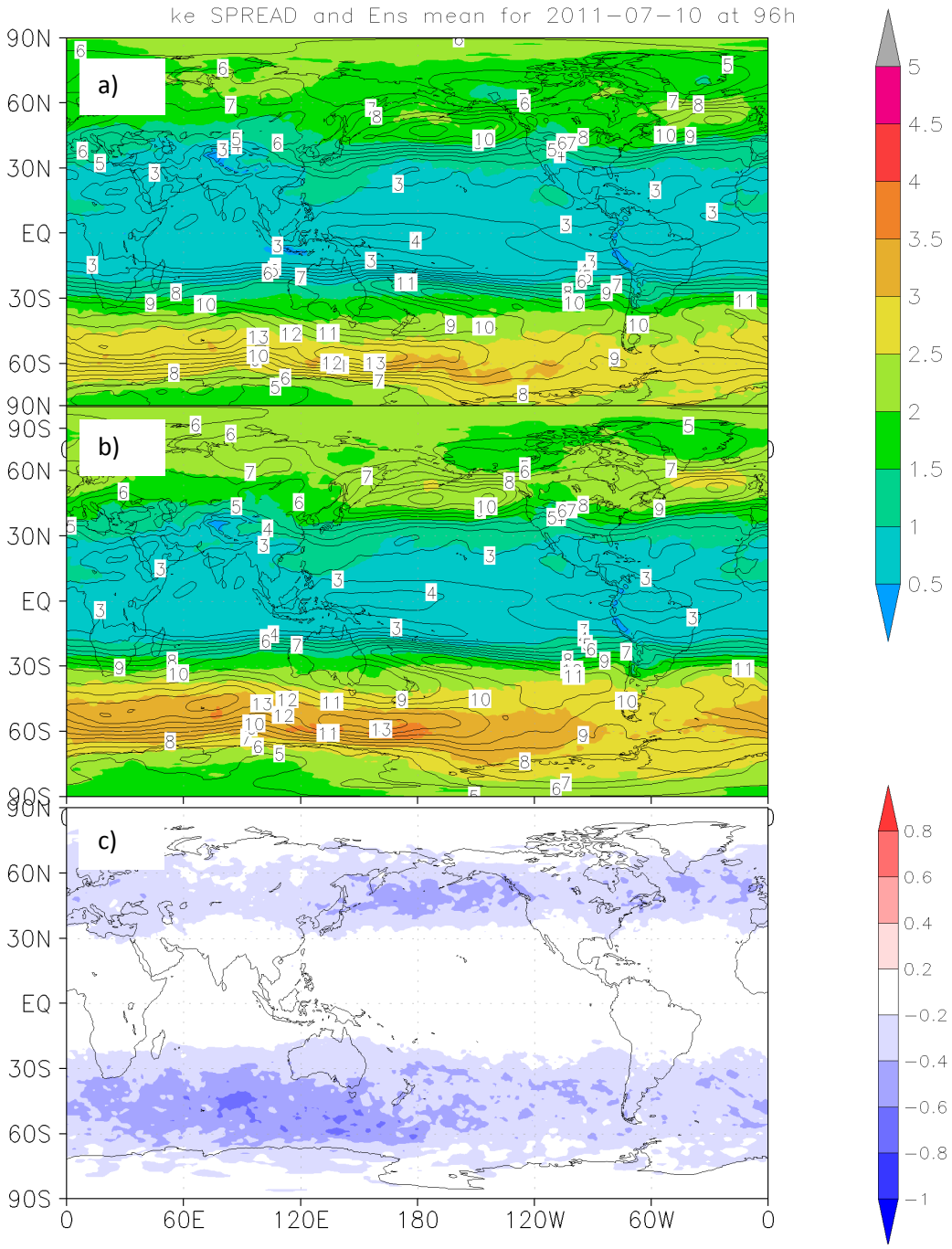


Fig. 3 Seasonally averaged ensemble spread of 96-hr forecast kinetic energy at 500 hPa for a) ETR, b) EnKF and c) the difference for 2011 summer.

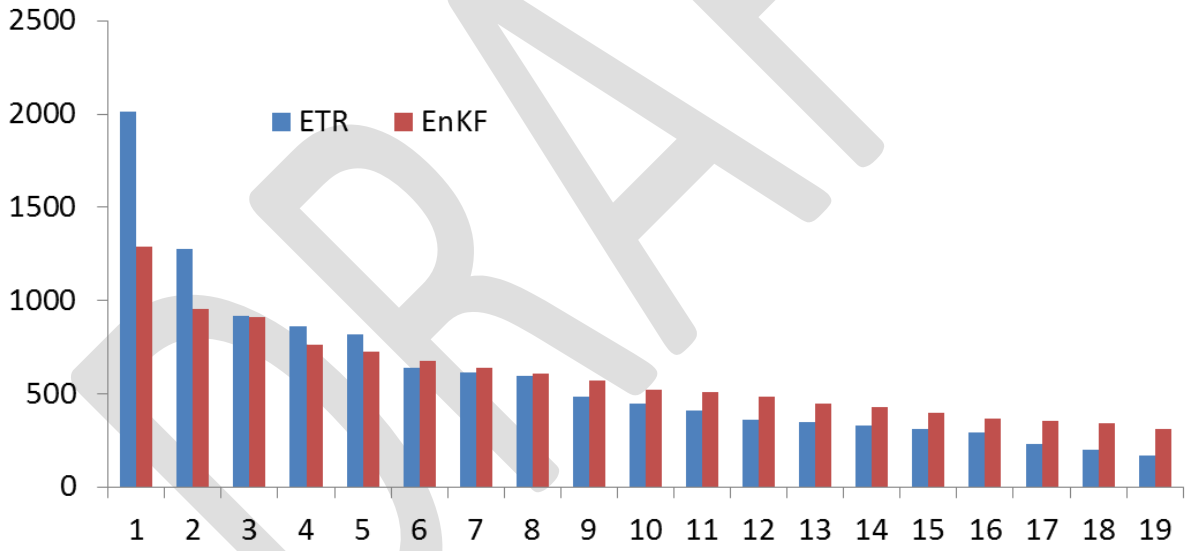


Fig. 4 Seasonal mean spectra of eigenvalues of ensemble based initial spread of 500 hPa Geopotential Height for 2011 summer normalized by the global spread for ETR and EnKF

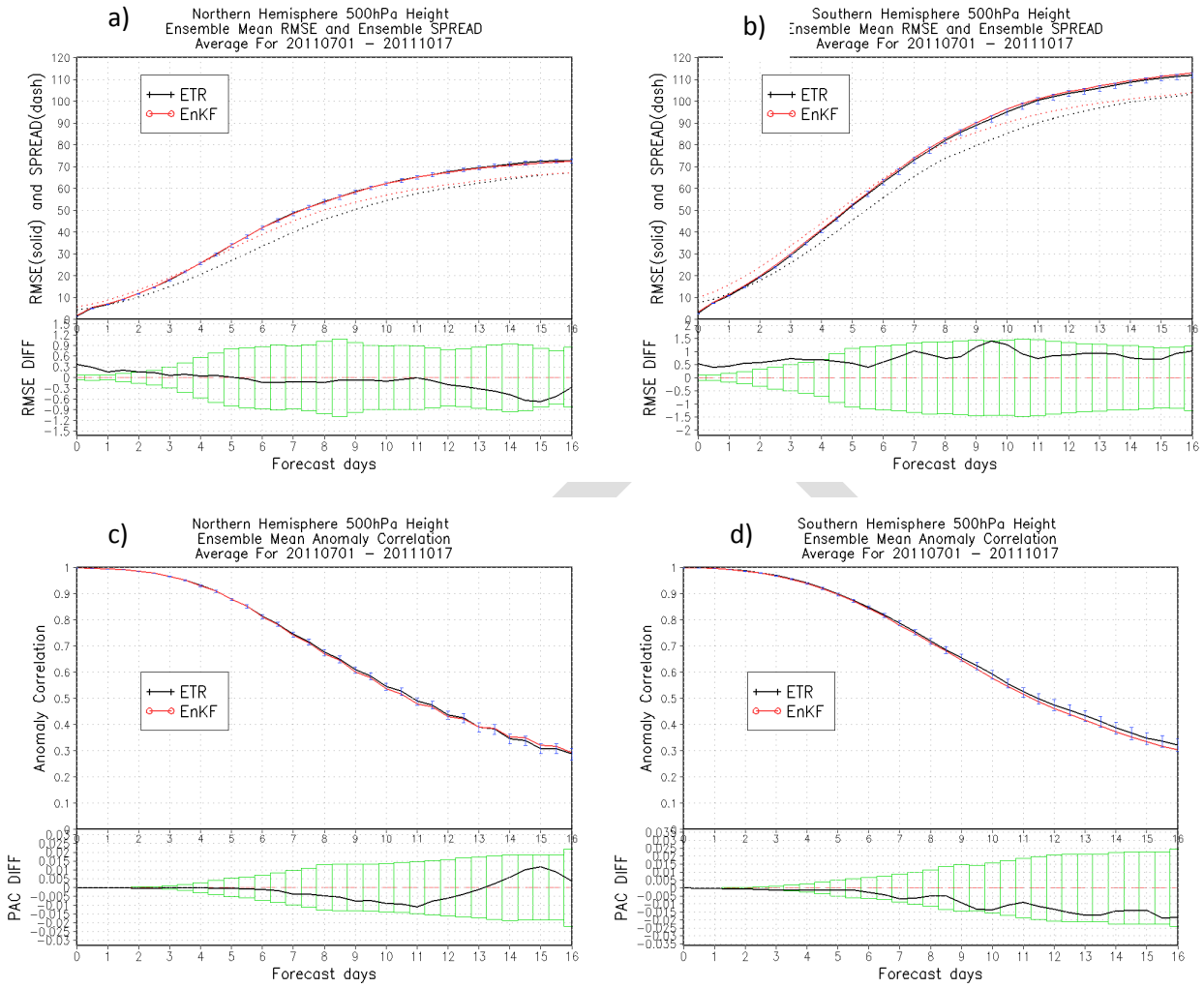


Fig.5 Ensemble mean rms error (solid) and ensemble standard deviation (dotted) for 500-hPa geopotential height over a) NH and b) SH. The verification scores for ETR (black) and EnKF (red) are averaged during the period from 1st July 1 – 17 th October 2011. The lower panels show the difference of RMSE (EnKF minus ETR, black lines) and bootstrap significant test (green bars). The difference is significant at 95% confidence level when value is beyond the bar ranges. c) and d) as same as a) and b) except for ensemble mean predicted pattern anomaly correlations of Z500.

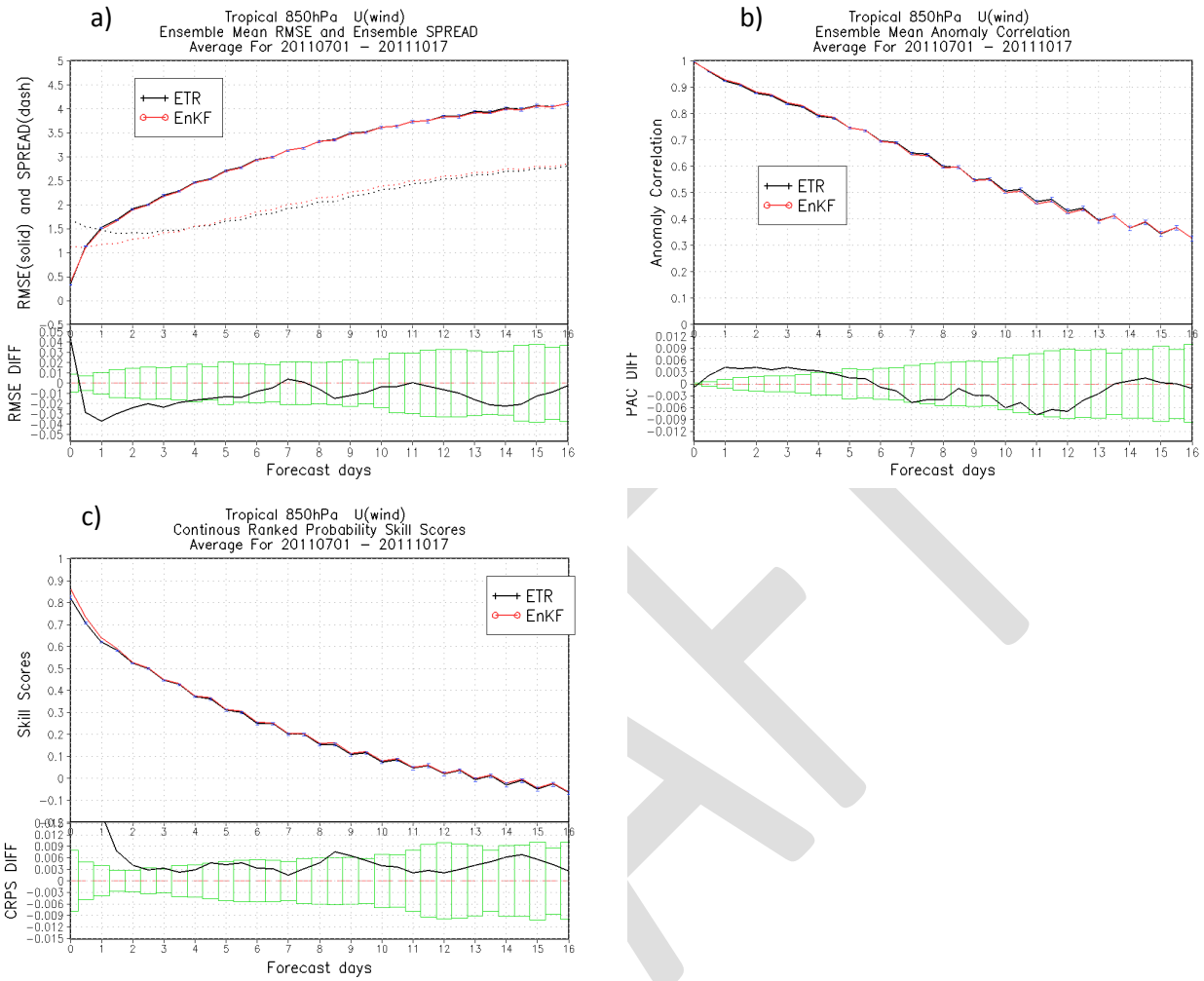


Fig. 6 a) Ensemble mean rms errors and spread, b) ensemble mean pattern anomaly correlations and c) continuous ranked probability skill scores for 850 hPa horizontal wind u component over Tropics averaged from 1st July to 17th Oct. 2011.

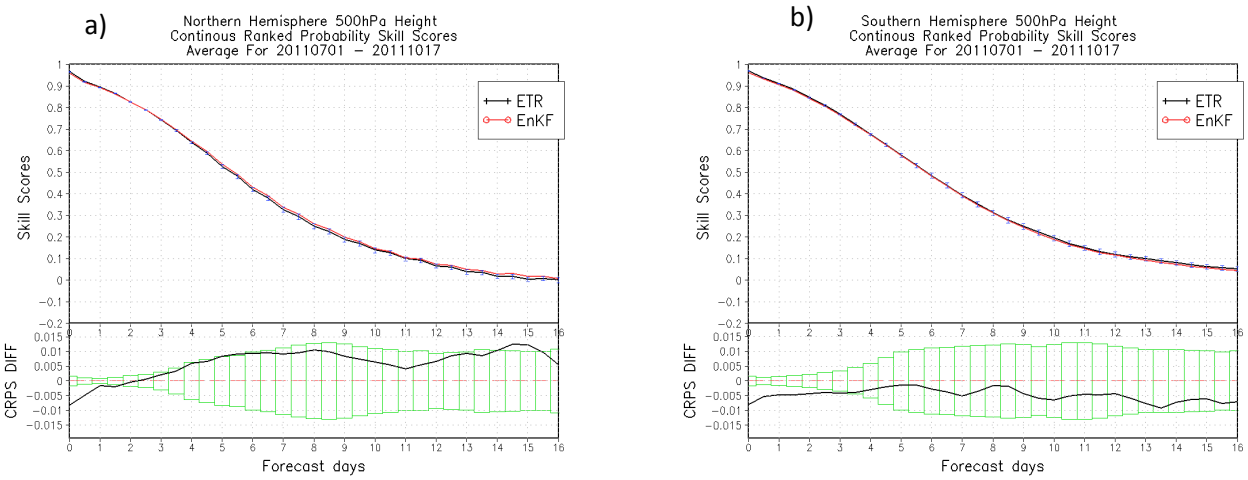


Fig. 7 Continuous ranked probability skill scores for 500 hpa height over (a) NH and (b) SH.

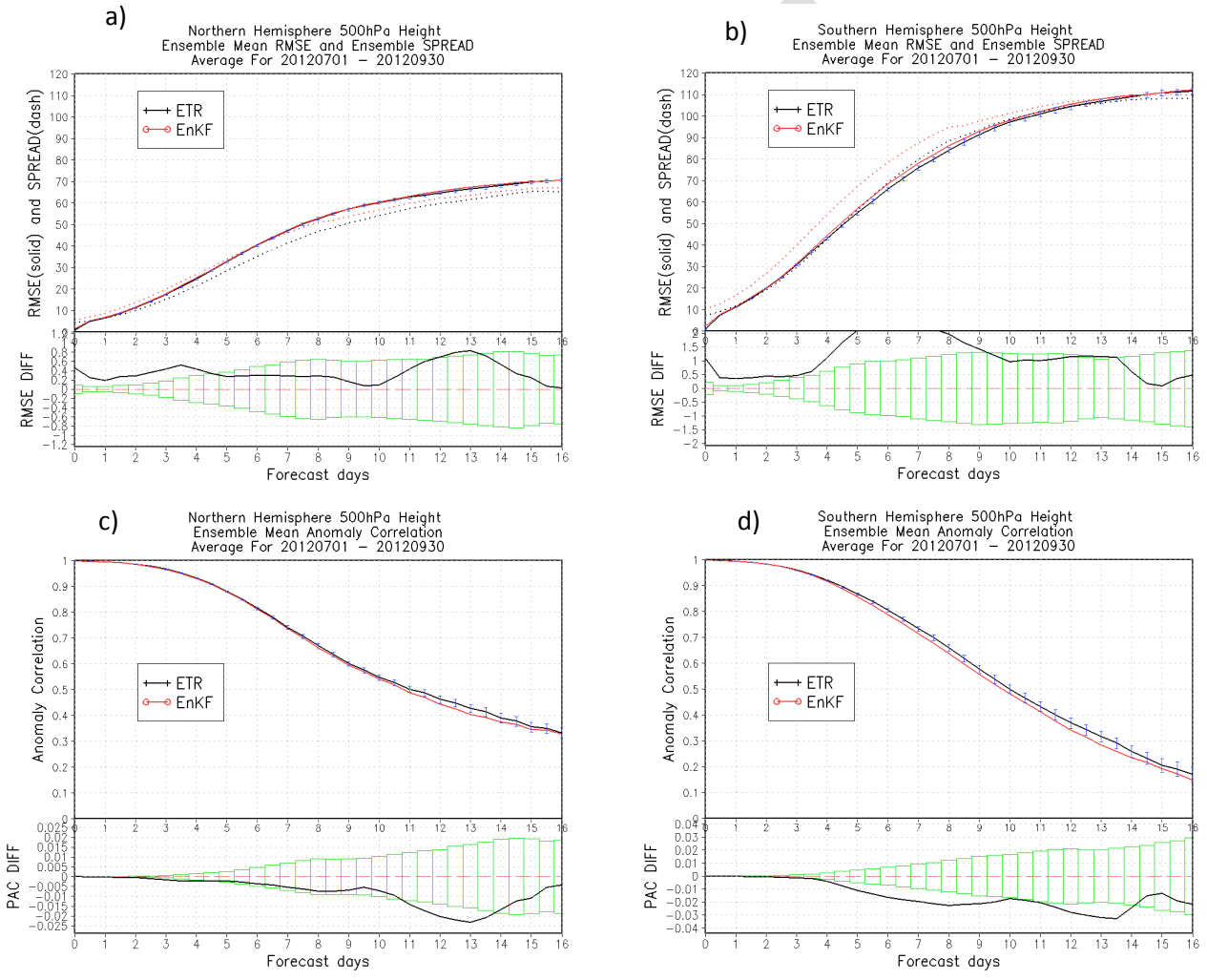


Fig. 8 As same as Fig. 5 except for the experiments with STTP averaged from 1st July to 30th Sep. 2012.

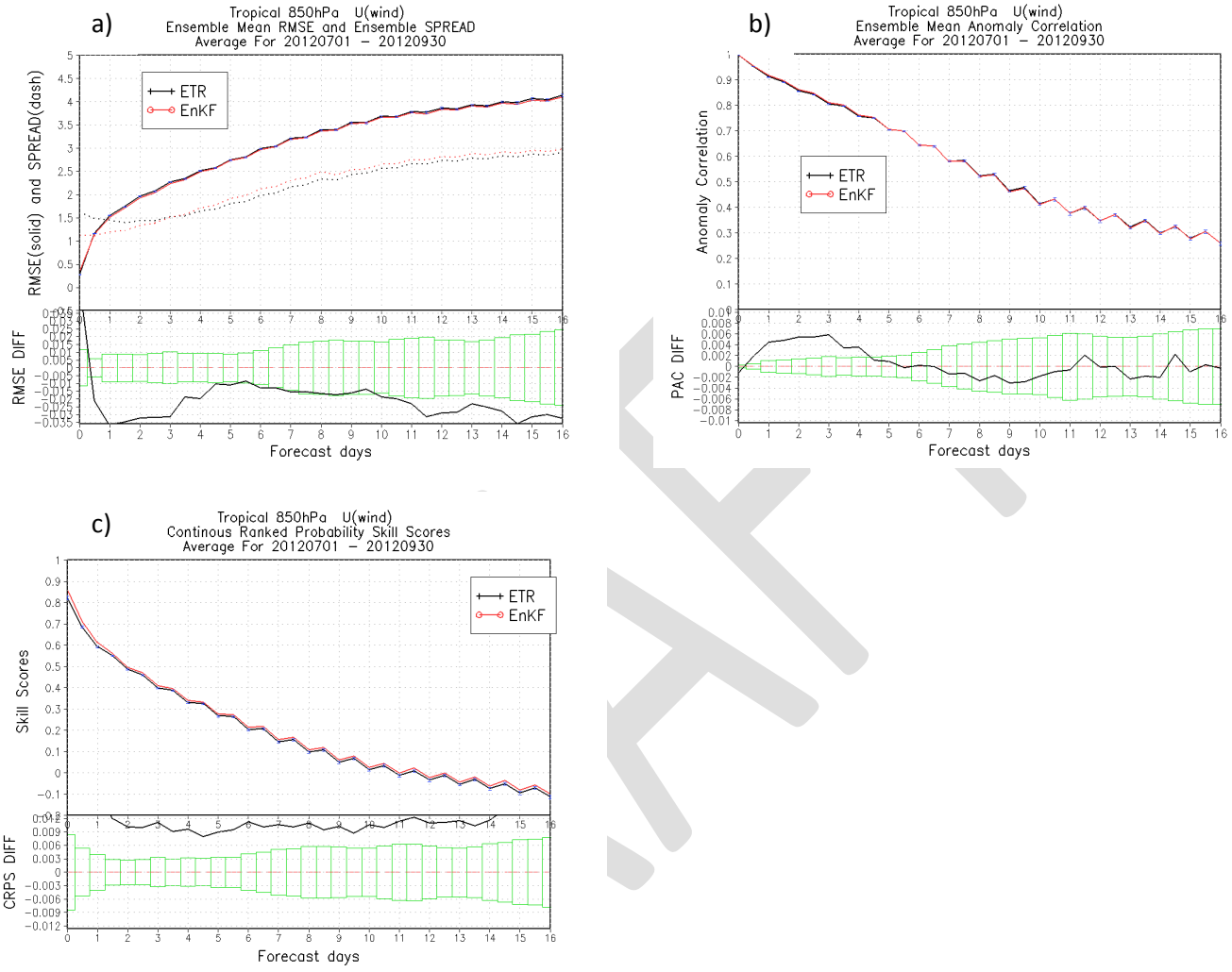


Fig. 9 As same as Fig. 6 expect for the experiments with STTP averaged from 1st July to 30th Sep. 2012.

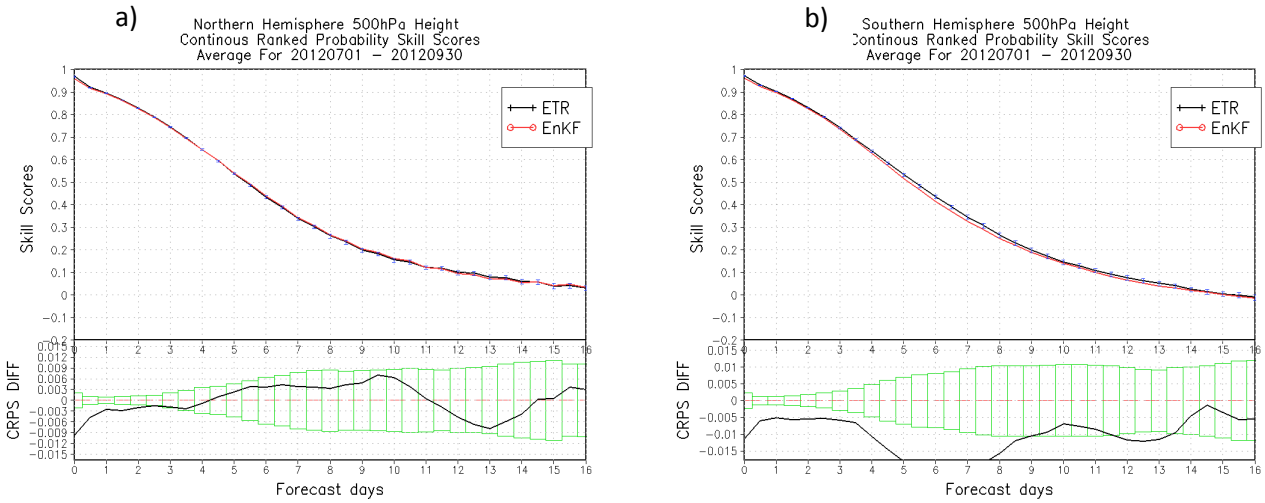
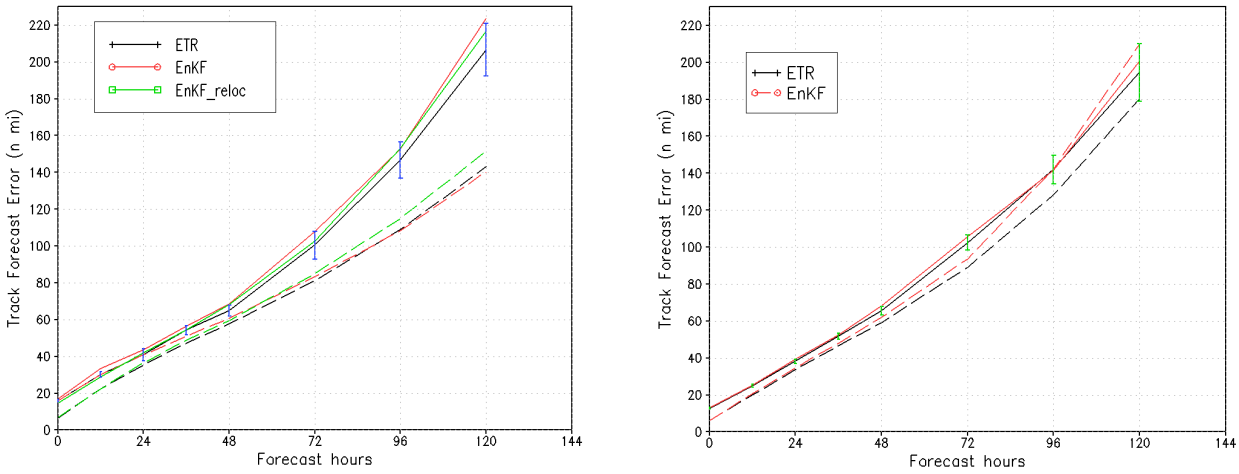


Fig. 10 As same as Fig. 7 except for 1st July -30 th Sep. 2012



Case No.232 218 187 171 150 119 89 66

Case No. 206 186 171 155 141 111 85 63

Fig. 11 Ensemble mean tropical cyclone track errors (solid) and spread (dashed line) for a) 2011 and b) 2012 summer.