Sensitivities of the NCEP Global Forecast System

Jih-Wang A.Wang[1,2], Prashant D. Sardeshmukh[1,2], Gilbert P. Compo[1,2], Jeffrey S. Whitaker[2], Laura C. Slivinski[1,2], Chesley M. McColl[1,2], and Philip J. Pegion[1,2]

[1]University of Colorado, CIRES, Boulder CO
[2]NOAA Earth System Research Laboratory, Physical Sciences Division, Boulder CO

22 January 2019

*Corresponding Author: Dr. Jih-Wang Aaron Wang, NOAA Earth System Research Laboratory, Physical Sciences Division, 325 Broadway, Boulder, CO, 80305.

Email: Aaron.Wang@noaa.gov

Abstract

An important issue in developing a forecast system is its sensitivity to additional observations for improving initial conditions, to the data assimilation (DA) method used, and to improvements in the forecast model. These sensitivities are investigated here for the Global Forecast System (GFS) of the National Centers for Environmental Prediction (NCEP). Four parallel sets of 7-day ensemble forecasts were generated for 100 forecast cases in mid-January to mid-March 2016. The sets differed in their 1) inclusion or exclusion of additional observations collected over the eastern Pacific during the El Niño Rapid Response (ENRR) field campaign, 2) use of a Hybrid 4D-EnVar versus a pure EnKF DA method to prepare the initial conditions, and 3) inclusion or exclusion of stochastic parameterizations in the forecast model. The Control forecast set used the ENRR observations, hybrid DA, and stochastic parameterizations. Errors of the ensemble-mean forecasts in this Control set were compared with those in the other sets, with emphasis on the upper tropospheric geopotential heights and vorticity, mid-tropospheric vertical velocity, column-integrated precipitable water, near-surface air temperature, and surface precipitation. In general, the forecast errors were found to be only slightly sensitive to the additional ENRR observations, more sensitive to the DA methods, and most sensitive to the inclusion of stochastic parameterizations in the model, which reduced errors globally in all the variables considered except geopotential heights in the tropical upper troposphere. The reduction in precipitation errors, determined with respect to two independent observational datasets, was particularly striking.

1. Introduction

The large improvement in weather prediction skill over the past several decades has been described as a "quiet revolution" resulting from many small steps rather than a few dramatic leaps (Bauer *et al.,* 2015). One has now apparently entered a stage of diminishing returns in skill improvement, with no clear guidance as to improving which aspects of current forecast systems will yield the greatest benefit. Broadly speaking, forecast systems have three basic elements: 1) the input observations, 2) the data assimilation (DA) method used to merge those observations with model-generated guess fields to create the forecast initial conditions, and 3) the forecast model itself. As forecast systems continue to evolve, their relative sensitivities to these three elements will evolve as well, and it will remain important to identify the element with the largest sensitivity to help set priorities in system development.

After decades of progress, both in-situ and remotely sensed observations available for forecast initialization have become plentiful, albeit with important gaps in the tropics and polar regions (see http://www.wmo.int/pages/prog/www/OSY/GOS.html). DA techniques have also improved, in both theory and implementation. In particular, two commonly used DA methods – Ensemble Kalman Filter (EnKF; Evensen, 2003) and Four-Dimensional Variational Data Assimilation (4DVar; Lewis and Derber, 1985; Courtier *et al.*, 1994) – and their various hybrids (e.g., 4D-EnVar; see Section 2.2) have matured in merging observations with model-generated first-guess fields to provide more accurate initial conditions for forecasts. The forecast models themselves have also improved, both in their representation of dynamical and physical tendencies and their use of much higher horizontal and vertical resolution (e.g., references in http://www.emc.ncep.noaa.gov/GFS/ref.php). These developments, together with expanding computing resources, now enable several operational weather forecasting centers around the world to generate ensembles of high-quality 10-day global forecasts on a 50 km or finer mesh every 12 hours.

Despite this, weather forecasts continue to be far from perfect. There is room for improvement in each of the three basic forecast system elements. The question is in which element to invest the most effort to gain the greatest benefit. A first step toward addressing this is to identify the element to which the forecasts are most sensitive. We will adopt this approach here for the Global Forecast System (GFS) used at the National Centers for Environmental Prediction (NCEP). Specifically, we will focus on its forecast performance and sensitivities in the mid–January to mid-March 2016 period during the mature phase of the 2015-16 El Niño event. An intensive observational El Niño Rapid Response (ENRR) field campaign was conducted by the National Oceanic and Atmospheric Administration (NOAA) over the tropical and subtropical eastern Pacific during the period (Dole *et al*., 2018), and the impact of the additional observations on GFS performance is of particular interest.

Section 2 provides relevant details of the additional ENRR observations, followed by a description of the numerical experiments performed to test the sensitivity of the GFS forecasts. Briefly, four parallel sets of 7-day 80-member ensemble forecasts were generated for 100 forecast cases in the period, differing in their 1) inclusion or exclusion of the additional ENRR observations, 2) use of a Hybrid 4D-EnVar versus a pure EnKF DA method to prepare the initial conditions, and 3) inclusion or exclusion of stochastic physical parameterizations in the forecast model. The Control forecast set used the ENRR observations, hybrid DA, and stochastic parameterizations. Section 3 compares the errors of the ensemble-mean forecasts in this Control set with those in the other sets, with emphasis on the errors of upper tropospheric geopotential heights and vorticity, mid-tropospheric vertical velocity, column-integrated precipitable water, near-surface temperature, and surface precipitation. A summary and concluding remarks follow in Section 4, emphasizing that although only a limited set of GFS sensitivities were investigated here, our methodology could also be fruitfully applied to investigate the sensitivities of other forecast systems to their three basic elements.

2. Additional observations and experimental design

2.1 ENRR Field Campaign

As discussed by Dole *et al*. (2018), a strong El Niño event was projected to occur in the northern winter and spring of 2015-16 based on observed tropical Pacific sea surface temperature (SST) anomalies in the preceding summer. NOAA seized this opportunity to undertake the ENRR field campaign to record the event while it was ongoing. The extra observations collected included 1) dropsonde, radar, and microwave radiometer observations from campaign flights (mostly within 180º-135ºW and between Honolulu and the equator), 2) radiosonde and surface observations from campaign cruises (Honolulu to

San Diego), 3) radiosonde and surface observations from Kiritimati Island (1.9˚N, 157.4˚W), and 4) radar observations from the U.S. west coast. These ENRR observations, together with the far more numerous routine conventional and satellite observations over the globe, provide an excellent opportunity to examine the impact of such event-oriented field campaign observations on weather forecast skill. The upper-air radiosonde and dropsonde observations covered most of the ENRR campaign area; there were 22,510 humidity observations, 33,646 temperature observations, and 35,943 wind observations by radiosondes and dropsondes from January 20 to March 16, 2016. We focus here on the forecast impact of only the upper-air radiosonde and dropsonde observations from the campaign, referring to them as "the ENRR observations". Full details of the campaign can be found in Dole *et al*. (2018) and at https://www.esrl.noaa.gov/psd/enso/rapid_response/, as well as in Slivinski *et al.* (2018).

2.2 *Analyses – Initial Conditions and "Truth"*

For clean comparisons, we generated our own analyses to provide initial conditions for our 7-day forecasts. We used the same 64-level version of NCEP's GFS model (Environmental Modeling Center, 2003) operational in April 2016 but at a lower horizontal resolution (spectral truncation of 254, approximate grid spacing of 50 km) for all the analyses and forecasts. To generate the analyses using NCEP's Global DA system, we performed sequential 6-hourly forecast-analysis cycles comprising the following steps:

Step 1: Combine an 80-member ensemble of 0- to 6-hr forecasts with observations in that 6-hour window to generate an 80-member ensemble of preliminary analyses.

Step 2: Perform IAU (incremental analysis update; see below for more details) from hr-0 to hr-6 to generate the “ultimate” analyses and continue running the 80-member ensemble for the next 6-hr background (i.e, first guess) ensemble of forecasts.

Step 3: Repeat Steps 1 through 2 for the next cycle.

In Step 1, we used either the Ensemble Kalman Filter method (EnKF; Evensen, 2003) or the Hybrid Four-Dimensional Ensemble Variational method (Hybrid 4D-EnVar; Buehner *et al.*, 2013; Kleist and Ide, 2015). The EnKF method is a Monte Carlo approximation of the Kalman Filter. It uses a model ensemble of finite size to approximate the probability distribution of predicted states, and updates the model-generated *a priori* state variables to *a posteriori* variables by using the model ensemble covariance to estimate the Kalman gain (Evensen, 2003). A reasonably large ensemble size is required for this purpose, and also to avoid abrupt imbalances among the state variables being updated. The problem of abrupt imbalances is partly overcome in Step 2 through an incremental analysis update (IAU; Bloom *et al.*, 1996; Lei and Whitaker, 2016; Takacs *et al.*, 2018), which divides the analysis increment from a preliminary analysis cycle into small portions and repeats the background forecast by adding the portions as extra forcing to the forecast at every time step. The final background forecast is the ultimate analysis, which closely resembles the preliminary analysis at the end of the forecast-analysis cycle but does not have abrupt imbalances, and is continued as the preliminary forecast for the next forecast-analysis cycle. For the present study, each analysis that we used for model initialization and verification purposes was the preliminary analysis (i.e., the output of EnKF or Hybrid DA before application of the IAU forcing) in the current forecast-analysis cycle, but it had the IAU forcing from the beginning of the experiment period (i.e., Jan 20, 2016; see Fig. 1 and context) up to the previous forecast-analysis cycle. There are two options in the NOAA EnKF code: the serial Ensemble Square Root Filter (EnSRF) and the Local Ensemble Transform Kalman Filter (LETKF). The EnSRF used here is also implemented operationally in the atmospheric GFS at NOAA. It is based on the serial EnSRF described in Whitaker and Hamill (2002) and uses the parallel algorithm described in Anderson and Collins (2007) for computational efficiency.

The Hybrid 4D-EnVar is a combination of EnKF and 4DVar (Four Dimensional Variational method; Lewis and Derber, 1985; Courtier *et al.*, 1994) which aims (a) to combine the time-varying ensemble covariances with static background error covariances to estimate the total background error contribution to the cost function being minimized, and (b) to eliminate the use of tangent-linear (TL) and adjoint (AD) models used in pure 4DVar (Wang *et al.*, 2008; Buehner *et al.*, 2013; Kleist and Ide, 2015).

In addition to the inclusion of a static background error covariance, the Hybrid 4D-EnVar differs from the EnKF in the way ‘covariance localization’ is performed. Covariance localization is a method for dealing with spurious covariances at large spatial lags that result from using small ensemble sizes. In the Hybrid 4D-EnVar system, covariance localization is performed in model space (Houtekamer and Mitchell, 2001) instead of observation space (Gaspari and Cohn, 1999; see summary of both in Lei and Whitaker, 2015). This can significantly impact the assimilation of observations such as satellite radiances, which involves using complicated forward observation operators to link the model state to the radiances (Campbell *et al.*, 2009). In the global numerical weather prediction (NWP) system of the National Weather Service (NWS), an 80-member EnKF is run operationally to initialize the Global Ensemble Forecast System (GEFS) and to provide ensemble covariances for the Hybrid 4D-EnVar data assimilation (Kleist and Ide, 2015) used by the Grid-point Statistical Interpolation (GSI) analysis system that generates the high-resolution deterministic analysis for the high-resolution GFS forecasts. In our analyses, we did not separately perform high-resolution deterministic analyses or forecasts; instead, we substituted the ensemble mean as the deterministic solution so that the interpolation from one resolution to another was avoided.

We performed the DA in Step 1 by using either the EnKF or Hybrid method, and either including or excluding the ENRR observations, thus generating four separate sets of 80-member ensemble analyses for the ENRR period. Given computing and storage constraints, we worked mainly with the Hybrid-with-ENRR set (hereafter the Control analysis set), the Hybrid-without-ENRR set (hereafter the Denial analysis set), and the EnKF-with-ENRR observations (hereafter the EnKFonly analysis set). These three sets of analyses were then used as initial conditions for three separate sets of 7-day 80-member ensemble forecasts. For forecast verification, we could have used any one of these three analysis sets as “truth”. However, we chose the Control analysis set for this purpose as our “best” analysis product, both because of its assimilation of all observations (including the ENRR observations) and its improved quality resulting from the hybridization. Using the EnKFonly or Denial analyses instead of the Control analyses for forecast verification did not affect any of our findings for forecasts beyond 24 hours.

2.3 Forecasts and Evaluations

The three analysis sets were used to initialize three sets of 7-day forecasts every 12 hours in the 57-day (20 January to 16 March) ENRR period. We will henceforth refer to these as Control, Denial, and EnKFonly forecasts, respectively. Their performance was evaluated by comparing them with the verifying Control analyses, and with independent observational estimates in the case of precipitation. The impact of the ENRR observations was gauged by comparing the skill of the Control and Denial forecasts, and the impact of the DA method by comparing the skill of the Control and EnKFonly forecasts. Table 1 lists these three sets of forecasts and their relevant characteristics.

All three forecast sets used stochastic parameterizations (SPs) to perturb the deterministic physical tendencies in the model. The use of SPs in operational forecasts is usually motivated by a need to increase the ensemble spread to make it more consistent with the generally larger root-mean-square error (RMSE) of ensemble-mean forecasts. Such a consistency is also implicitly assumed in the EnKF. The GFS SP module can employ three different types of SPs, namely SPPT (Stochastically Perturbed Physical Tendencies; Palmer *et al.*, 2009; Shutts *et al.*, 2011), SHUM (Stochastic HUMidity perturbations in the boundary layer; Tompkins and Berner, 2008), and SKEB (Stochastic Kinetic Energy Backscatter; Berner *et al.*, 2009), to increase the ensemble spread. The SPPT scheme has the following general form for the tendency perturbation:

$$\dot{x}_p = (1 + r\mu)\dot{x}_c \,,$$

where $\dot{x}_c$ and $\dot{x}_p$ are the physical tendencies of the state variable before and after applying the stochastic perturbation, respectively; r is a stochastic horizontal weight that is bounded in the interval [-1,1] by using an inverse logit transform of a Gaussian distribution, and μ is a vertical weight that is 1 between the surface and 100hPa and is tapered to zero at 25hPa. The horizontal weight r can be represented in terms of spherical harmonics as

$$r = \sum_{mn} \hat{r}_{mn}\, Y_{mn} \,,$$

where $\hat{r}_{mn}$ is the spherical harmonic coefficient of r for total wavenumber n and zonal wavenumber m. This enables the tendency perturbation to be made scale-aware and smoothed in space to the degree desired. Palmer *et al.* (2009) (see also Sardeshmukh, 2005) represented $\hat{r}_{mn}$ as a combination of a first-order autoregressive AR(1) process and spatially smoothed white noise as

$$\hat{r}_{mn}(t + \Delta t) = \phi \hat{r}_{mn}(t) + \sigma_n \eta_{mn}(t) \,,$$

where Δt is the model time step, $\phi = exp(-\Delta t/\tau)$ is the AR(1) coefficient, σ_n is the standard deviation (i.e., strength) of the tendency perturbation, and $\eta_{mn}(t)$ is a Gaussian random number with zero mean and unit variance. σ_n is a function of total wavenumber n and spatial autocorrelation length scale L such that the variance in grid space $Var(r)$ is uniform and the spatial pattern has a spatial autocorrelation corresponding to the equivalent of a Gaussian function on the sphere (Palmer *et al.*, 2009; Sardeshmukh, 2005; Weaver and Courtier, 2001). The SPPT scheme is applied to the tendencies of zonal wind, meridional wind, specific humidity, and temperature induced by the GFS physics package, but not to the tendencies induced by the clear-sky radiation scheme.

The SHUM perturbations are similar to the SPPT perturbations, except that they are applied to the humidity itself and not the humidity tendency (although they may be interpreted as perturbations to the humidity tendency integrated over a model time step), and only in the lower troposphere. The formula is

$$q_p = (1 + r\mu)q_c \ ,$$

where q_c and q_p are the specific humidity before and after the stochastic perturbation respectively. The vertical weight μ decays exponentially in pressure away from the surface. The scheme additionally constrains the specific humidity to remain positive.

We used SPPT and SHUM perturbations (but not SKEB perturbations) in all three sets of forecasts. We could have specified multiple values of the AR(1) e-folding time scale τ, spatial variance *Var(r)*, and spatial autocorrelation scale L to avoid the early saturation of ensemble spread at small scales. However, for simplicity we chose fixed values of τ= 6 hours, *Var(r)* = 0.8 and L = 500 km for the SPPT, and τ= 6 hours, *Var(r)* = 0.005 and L = 500 km for the SHUM perturbations.

Finally, in order to quantify the impact of the SPs, we generated a fourth set of 7-day forecasts similar to the Control forecasts but without SPs (labeled noSP; see Table 1). As with the other three forecast sets, the skill of the noSP forecasts was evaluated by comparing with the verifying Control analyses, and the impact of the SPs was gauged by comparing the skill of the Control and noSP forecasts.

To summarize, the Control, Denial, EnKFonly and noSP forecasts were each 7-day 80-member ensemble forecasts, started twice a day at 00Z and 12Z in the 57-day ENRR period. There were thus 114 forecast cases in each set. The forecast output frequency was 3 hours (i.e. 3, 6, 9, ..., 168 hours). To ensure the same number of forecast verifications for all forecast lead times, we only evaluated forecasts valid between January 27 and March 16. As illustrated in Fig. 1, this verification period spans 50 days and contains 100 verification cases (with each case corresponding to one initialization time) for each forecast lead time. Overall, for each forecast lead time we thus had 4 sets × 80 forecasts × 100 cases = 32,000 forecasts of all model variables at all grid points. We shall show below that these large sample sizes enable us to quantify the impacts of the ENRR observations, DA methods, and SPs on the forecast skill with statistical confidence.

3. Forecast Evaluation and Comparisons

3.1 Forecast Errors

We define the forecast error as the RMSE of the *M*=80 member ensemble-mean forecast with respect to the 80-member ensemble-mean Control analysis, determined over all *N*=100 forecast cases as

$$RMSE(t) = \left\{ \frac{1}{N} \sum_{n=1}^{N} V'^{2}_{n,t} \right\}^{1/2},$$

where

$$V'_{n,t} = V_{f,n,t} - V_{a,n} = \frac{1}{M}\sum_{m=1}^{M} V_{f,n,t}^{m} - \frac{1}{M}\sum_{m=1}^{M} V_{a,n}^{m}$$

Here subscript *t* refers to forecast lead time, *f* and *a* to the forecast or verifying analysis of variable *V*, *n* to the forecast case number, and *m* to the ensemble member number. This expression was used to calculate *RMSE(t)* for selected variables at each grid point. An analogous expression, with the area-weighted gridpoint values of $V'^{2}_{n,t}$ averaged additionally over the globe as well as over some specific regions, was used to calculate global and regional values of *RMSE(t)*. We focus here on the forecast errors of geopotential height at 200 hPa (Z_{200hPa}), relative vorticity at 200 hPa (ξ_{200hPa}), vertical velocity at 500 hPa (ω_{500hPa}), column-integrated precipitable water (PWAT), and 2-meter air temperature (T_{2m}). The RMSEs for a few additional variables were also examined but are not shown here due to their similar behavior.

For precipitation, we compared forecasts of 12-hour accumulated precipitation values (AP12HR) with two independent observational datasets: the NASA (National Aeronautics and Space Administration) GPM (Global Precipitation Measurement) dataset (Huffman *et al.,* 2014) and the PERSIANN (Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks) dataset (Sorooshian *et al.,* 2014; Ashouri *et al.,* 2015). For brevity, we only show the comparison with the NASA GPM dataset, since the comparison with the PERSIANN dataset yielded similar results.

Fig. 2 shows the area-weighted global RMSEs of the Control, Denial, EnKFonly, and noSP forecasts of Z_{200hPa}, ξ_{200hPa}, ω_{500hPa}, PWAT, and T_{2m} at 12-hourly intervals up to 7 days (hr-

168), as well as the RMSEs of AP12HR between 20ºS and 20ºN and between 60ºS and 60ºN. The initial (hr-0) error of the Denial forecasts reflects the difference between the Control and Denial analyses (not shown). The Control forecasts have slightly smaller errors than the Denial forecasts until hr-24 but show no discernible impact thereafter, at least in this global metric, of including the ENRR observations in the initial conditions.

In contrast, the global RMSEs of the EnKFonly forecasts are larger than those of the Control and Denial forecasts throughout the forecast period. Indeed, the EnKFonly forecasts are worse than the Control forecasts beyond Day 1 even when both are verified against the EnKFonly analyses (not shown) instead of the Control analyses as in Fig. 2. We should stress that this result does not imply that an EnKF method is inferior to a Hybrid method in general. One can think of several ways in which our particular implementation of the EnKF algorithm could have been improved, such as by adjusting the vertical covariance localization of the satellite radiance observations, by improving the balance constraints on analysis increments, and by increasing the ensemble size of the ensemble Kalman Filter. Nevertheless, Fig. 2 clearly demonstrates the greater sensitivity of the forecast errors to initial conditions prepared using different DA methods than to the inclusion or exclusion of the ENRR observations in those initial conditions.

The global RMSEs of the Control forecasts are smaller than those of noSP forecasts for ω_{500hPa}, ξ_{200hPa}, and PWAT throughout the 7-day forecast range, demonstrating the beneficial impact of including SPs in the model. Similar reductions in ensemble-mean forecast errors have been reported in other forecast systems (e.g., Leutbecher *et al.,* 2017).

The global RMSEs of the noSP forecasts are larger than those of the EnKFonly forecasts after Day 3 for ω_{500hPa}, Day 6 for ξ_{200hPa}, and Day 5 for PWAT. In other words, beyond Day 3 these forecasts errors are more sensitive to including or not including SPs in the forecast model than they are to the use of the Hybrid versus EnKF DA method to prepare the forecast initial conditions. The ω_{500hPa} errors saturate by about Day 6 (Fig. 2c), but interestingly the PWAT errors do not saturate even by Day 15 (not shown). The precipitation errors (Fig. 2f) saturate at an intermediate lead time of about Day 7. Although ω_{500hPa} and PWAT are both important for determining precipitation strength, the near-simultaneity of ω_{500hPa} and precipitation error saturation suggests that ω_{500hPa} has a stronger control than PWAT on determining precipitation variations on the time scales of synoptic weather (see also Sardeshmukh *et al.*, 2015).

The error growth curves of T_{2m} (Fig. 2e) and precipitation (Fig. 2f) in the Control, Denial, EnKFonly, and noSP forecasts have a similar general character to that of the other variables, with little or no sensitivity to the ENRR observations, considerably higher sensitivity to the choice of the Hybrid versus EnKF DA method, and greatest sensitivity to the use of SPs in the model. For all variables in Fig. 2 except Z_{200hPa}, the Control forecasts are the best and the noSP forecasts are the worst by Day 7. The impact of the SPs is evidently cumulative over time, resulting by Day 7 in a reduction of the precipitation forecast error in the Control forecasts by ~4.3% in the 20ºS-20ºN latitude domain and by ~3% in the 60ºS-60ºN latitude domain.

Note that the errors of the 12-hour accumulated precipitation amounts in all four forecast sets, measured with respect to the observational GPM values, are already quite large (> 6.5 mm) at hr-12. The GPM precipitation is a blend of radar-reflection and radiance based precipitation estimates from multiple satellites, and is calibrated against in-situ ground observations. For a cleaner comparison with the precipitation forecasts, we integrated the 30-minute 0.1° resolution GPM values to 12-hr 0.5° resolution values. Given that precipitation is a positive semi-definite quantity, its substantial error even at short forecast ranges suggests that there are precipitation events of which locations and large magnitude (> 100mm accumulations in 12 hours) are not captured by our forecasts.

The general conclusions drawn from the global forecast error growth curves in Fig. 2 are also valid for limited regions. To illustrate this, Fig. 3 shows the RMSEs of ω_{500hPa} in the Northern Hemisphere (20^{o}N-90^{o}N), Southern Hemisphere (20^{o}S-90^{o}S), Tropics (20^{o}S-20^{o}N), and the contiguous United States (CONUS; 125^{o}W-66^{o}W, 24^{o}N-50^{o}N). The errors saturate in the Northern Hemisphere, Southern Hemisphere, and Tropics by Day 7, and nearly saturate in the CONUS region by the end of Day 7. Geographically, the errors are largest in the extratropical storm track regions and in areas of tropical deep convection (Fig. 4a). They are particularly large over the CONUS region, not surprisingly because the region overlaps strongly with the northern hemispheric storm track at those longitudes, but also possibly because of erroneous model representations of the influence of the Rocky Mountains on synoptic weather systems.

A beneficial impact of the ENRR observations on the regional ω_{500hPa} forecasts is not discernible in Fig. 3 beyond Day 1, which reflects an average of small differences of mixed signs between the Control and Denial forecasts. For instance, small positive and negative impacts on Day 7, likely not statistically significant, are scattered around the globe (Fig. 4b) with no coherent geographical structure. On the other hand, using the Hybrid versus the EnKF initial conditions leads to smaller Day-7 errors in many though not all regions (Fig. 4c). However, including SPs in the model unambiguously reduces the ω_{500hPa} error almost everywhere on the globe (Fig. 4d). The improvement is particularly clear in the Northern Hemisphere storm track and tropical convective regions.

Given the strong link between ω_{500hPa} and precipitation on synoptic time scales, the results for the precipitation errors in the Control forecasts and how they differ from the errors in the other three forecast sets (Fig. 5) are highly consistent with the results for the ω_{500hPa} errors in Fig 4. Similar to the ω_{500hPa} errors, the precipitation errors are least sensitive to including or excluding the ENRR observations, more sensitive to the choice of the Hybrid versus EnKF DA method used to initialize the forecasts, and most sensitive to using or not using the SPs in the forecast model.

Fig. 6 shows the errors of near-surface air temperature (T_{2m}) in the Control forecasts and how they differ from the errors in the other three forecast sets. Note that the prescribed SST boundary conditions are updated daily in the analyses but not in the 7-day forecasts. Still, because the SSTs vary little and the T_{2m} values over the ocean are tightly linked to them, the T_{2m} RMSE over the oceans remains relatively small over the 7-day forecast range. Also, because the prescribed SSTs are identical in all the four forecast sets, the differences of the T_{2m} errors over the oceans among the forecast sets are small as well. The Control forecast errors are larger over land and largest in high latitudes (Fig. 6a). The differences between the RMSEs of the Control and Denial forecasts are also large over high-latitude land, but with mixed signs (Fig. 6b). The impact of the choice of the Hybrid over the EnKF DA method is stronger than the impact of the ENRR observations (cf. Figs. 6c and 6b). Including the SPs again has the largest impact (Fig. 6d), with an unambiguous reduction of the T_{2m} error almost everywhere, but especially over land areas.

Using SPs is clearly beneficial for the ω_{500hPa}, precipitation, and T_{2m} forecasts over most of the globe. For upper tropospheric geopotential heights (Z_{200hPa}), however, the benefit is not so clear-cut. The impact is negligible in the extratropics and negative in the tropics, as shown in Fig. 7 for the same four regions as in Fig. 3. The Control and Denial forecast errors are again very similar, except in the CONUS region where the Control errors are slightly smaller than the Denial errors on Days 3-5 (Fig. 7d). Perhaps this is to be expected, given that the CONUS region is downstream of the region of the ENRR observations. We also show below in Section 3.2 that even though the positive impact of the ENRR observations is weak, there is a recognizable enhancement of El Niño-related features over North America in Z_{200hPa} due to the ENRR observations.

It is evident that the Z_{200hPa} RMSE sensitivity to the DA methods is different in the Northern Hemisphere, Southern Hemisphere and Tropics (cf. Figs. 7a, 7b, 7c). Using the Hybrid versus the EnKF method has a large positive impact on the Z_{200hPa} forecasts in the Southern

Hemisphere, a weaker positive impact in the Northern Hemisphere, but a negative impact in the Tropics starting from about Day 2. Interestingly, using the Control (Hybrid DA) versus the EnKFonly analyses as initial conditions also increases the positive tropical bias of the Day-7 Z_{200hPa} Control forecasts (cf. Figs. 9a, 9c). The EnKFonly analyses have lower Z_{200hPa} than the Control analyses in the tropics, resulting from several methodological differences in the EnKF algorithm, including (a) covariance localization of satellite radiances (see Lei *et al.* (2019) for a recent study); (b) lack of additional balance constraints on analysis increments; (c) no static background error covariances; and (d) use of maximum likelihood versus minimum variance estimation as in 4D-EnVar. While both Control and EnKFonly forecasts develop positive tropical biases over 7 days, the EnKFonly forecasts are closer to the truth and have smaller RMSEs. The forecast model drift toward higher Z_{200hPa} in the tropics is worthy of further investigation. With regard to the impact of SPs on the Z_{200hPa} forecasts, their positive impact does not become clear in the global RMSE metric until the end of Day 7 (Fig. 2a), because of cancellations between the positive impacts in the extratropics and negative impacts in the tropics seen in Fig. 8d.

Fig. 8 shows the Day-7 errors of the Control Z_{200hPa} forecasts and how they differ from the errors in the other three forecast sets. The impact of the ENRR observations is relatively small in the tropics and mixed in the extratropics (Fig. 8b). Using the Hybrid versus EnKF initialization yields a similarly mixed impact in the extratropics, and a small but clear degradation in the tropics (Fig. 8c). Using the SPs in the forecast model yields a more consistent beneficial impact in the extratropics, but also a much stronger degradation of the Z_{200hPa} forecasts in the tropics (Fig. 8d). Interestingly, this degradation occurs not just over the tropical convective areas but also over clear-sky areas in the descending branch of the Pacific Walker cell, in which one would expect scant local SPPT tendencies of radiative heating.

3.2 Forecast biases

Thus far, we have considered GFS forecast sensitivities to the ENRR observations, data assimilation method, and stochastic parameterizations in terms of RMSE measures of ensemble-mean forecasts. It is also relevant to consider how these three factors affect the mean forecast drift, i.e., the systematic bias at each forecast lead time of the ensemble-mean forecasts averaged over all 100 forecast cases. Fig. 9a shows the biases of the Day-7 Z_{200hPa} Control forecasts. Note that unlike the RMSEs, which are positive at all locations, the biases can be positive or negative. Some prominent features in Fig. 9a, such as the positive biases over North America, East Asia, Europe, and the tropics, and the negative biases over the northwest Pacific, northeast Pacific, and northeastern U.S., appear early in the forecasts and are evident throughout the 7-day forecasts (not shown).

The other panels of Fig. 9 show the systematic differences of the ensemble-mean Z_{200hPa} Control forecasts from the ensemble-mean forecasts in the other three forecast sets. They may also be interpreted as the impacts of the ENRR observations (Fig. 9b), Hybrid vs. EnKF initial conditions (Fig. 9c), and stochastic parameterizations (Fig.9d) on the Control forecast biases. The impact of the ENRR observations is apparently to intensify El Niño-related features in the Day-7 Z_{200hPa} forecasts: a low along the Canadian West Coast and

U.S. Pacific Northwest, a high to the west of the Great Lakes, and another high off the Northeast U.S. coast. Although this impact is not statistically significant (see Fig. 11), it is not inconsistent with the response to an anomalous equatorial heat source located east of the dateline (Ting and Sardeshmukh, 1993) during El Niño events. The impact is likely due to a slight but systematic strengthening of the tropical upper tropospheric convective outflow in the Control analyses using the ENRR wind observations (Slivinski *et al.,* 2018) and consequently the Rossby wave source associated with the El Niño-related tropical heating (Sardeshmukh and Hoskins, 1988).

The impacts of the DA method and SPs on the ensemble-mean Z_{200hPa} Control forecast biases in Fig. 9c are much larger than those of the ENRR observations. Both increase the ensemble-mean Z_{200hPa} in the tropics and subtropics, and contribute to the positive bias of the Control Z_{200hPa} forecasts over these large regions covering more than 50% of the globe. The negative impact of the SPs is especially strong and remarkable, considering that the Control forecast biases are determined with respect to analyses which include SPs in the DA model. This degradation is evident as early as Day 1 in the tropics, spreading thereafter to higher latitudes (not shown). A preliminary diagnosis suggests that it originates largely from a nonlinear response of convection to the SHUM perturbations, which are themselves unbiased (i.e., have zero mean). The impact of using the Hybrid versus EnKF initial conditions is more mixed in this regard, with alternating positive and negative impacts along the Northern Hemisphere extratropical jet stream waveguide.

Fig. 10 shows similar bias results for ω_{500hPa} in an identical format to Fig. 9. To focus on larger-scale features, we smoothed the fields using the spatial filter described in

Sardeshmukh and Hoskins (1984), retaining scales corresponding to total spherical wavenumbers 15 and lower. Even so, the fields remain noisy, but with a clear suggestion of a wave-train of alternating positive and negative Control forecast biases along the extratropical jet stream waveguide. This wave-train is also evident in the other panels of Fig. 10 showing the bias impacts of the ENRR observations, using the different DA methods, and SPs. Inspection of maps similar to those in Fig 10, but for earlier forecast lead times (not shown) reveal this wave-train to be a remarkably robust eastward propagating feature of the Control forecast biases and bias impacts. Note that the bias impacts of the ENRR observations and DA method stem only from differences in the forecast initial conditions, whereas the bias impacts of the SPs result from changes to the forecast model. The impact of the ENRR observations occurs initially as westward propagating tropical waves that provide perturbations in sensitive regions for exciting the mid-latitude wave-train. The impact of the DA method is stronger than that of the ENRR observations, because the systematic differences between the Hybrid and EnKF DA (see Section 2.2 for the DA method description) are larger than those between the Control and Denial analyses. The impact of the SPs is different in being much stronger in the tropics, and with a slower emergence of the midlatitude wave-train. This slower emergence is not unexpected, since the SPs provide new perturbations throughout the forecast and prevent the occurrence of coherent optimal conditions for exciting the wave-train.

The bias results in Figs. 9 and 10 have a dynamically meaningful interpretation in at least the extratropics. The extratropical wave-train is highly reminiscent of the most unstable (or least damped) perturbation eigenmode of the extratropical circulation investigated by Hall and Sardeshmukh (1998). On the other hand, since almost any perturbation can set off such an unstable eigenmode with arbitrary amplitude and phase, its appearance in our bias impact statistics makes it harder to distinguish among our estimated bias sensitivities to the ENRR observations, DA methods, and SPs and to establish their statistical significance.

Indeed, it turns out that the bias impacts in Figs. 9b, 9d, 10b, and 10d are generally not statistically significant in the extratropics. This is shown in Fig.11 for Z_{200hPa} and ω_{500hPa} in terms of the Student's *t* scores of the estimated bias differences. The details of these significance calculations are provided in Appendix A. The impact of the ENRR observations on the Day-7 forecast biases is insignificant almost everywhere on the globe. While the bias impacts of the hybrid DA are significant in some scattered areas in the extratropics, the bias impacts of the SPs are generally insignificant outside the tropics. However, they are both highly significant in the tropics.

4. Summary and concluding remarks

In our forecast sensitivity experiments, the impact of the ENRR observations on the RMSEs of the ensemble-mean forecasts was relatively large at short forecast lead times (about 1 day) whereas the impact of using the Hybrid versus EnKF DA method lasted throughout the forecast period (7 days). This was evident for all the six variables examined (Z_{200hPa}, ξ_{200hPa}, ω_{500hPa}, PWAT, T_{2m}, and AP12HR). The impact of the SPs was to reduce the RMSEs of the ensemble-mean forecasts of all these variables, except Z_{200hPa} in the tropics. Furthermore, this generally positive impact of the SPs grew with forecast lead time.

The mechanisms through which SPs reduce the errors of ensemble-mean forecasts are worthy of a more detailed investigation, which will be reported elsewhere.

To varying degrees, the ENRR observations, DA method, and SPs also impacted the forecast biases. The impact of the ENRR observations was the weakest and not statistically significant over most of the globe. The impacts of the DA method were statistically significant in the tropics and in some scattered areas in the extratropics, while the impacts of the SPs were highly significant and generally concentrated in the tropics. The impact of the SPs was stronger than that of the DA method.

In summary, our goal in this study was to assess the relative sensitivities of global GFS forecasts during late winter/early spring 2016 to the additional ENRR observations collected during the period, to the DA method used to provide the forecast initial conditions, and to the use of SPs in the forecast model. Of these, the sensitivity to the additional ENRR observations, in terms of both biases and RMSEs of the ensemble-mean forecasts, was found to be the weakest, and that to the SPs the strongest, in the 100 forecast cases investigated. The generally positive impact of the SPs on the ensemble-mean forecasts, and also their strongly negative impact on the tropical Z_{200hPa} forecasts, are noteworthy and require further investigation.

Modern forecast systems are sensitive to many system elements, and our investigation was certainly not meant to be exhaustive in this regard. Rather, our goal was to provide a sense of the relative sensitivities to the three principal types of development activities that are of current interest at major forecasting centers: collecting and using more observations, developing better data assimilation methods, and improving the forecast models.

As far as we are aware, our study is the first to perform sensitivity tests of sufficient size simultaneously on all the three basic elements of an ensemble forecast system to produce statistically meaningful results for intercomparisons. Even so, the generalizability of our results is limited. For example, our result that the additional ENRR observations did not significantly improve the GFS forecast skill does not necessarily imply that additional observations will have little impact on forecast skill in general. It is well known that short-range forecasts of high-impact weather events benefit from additional in-situ observations (e.g., NOAA Sensing Hazards with Operational Unmanned Technology project). Clearly, the impact of additional observations depends on their relative augmentation of pre-existing observational networks as well as on the types and scales of target weather events.

Our investigation of forecast sensitivities to DA methods was likewise not exhaustive, as we only compared one implementation of the Hybrid 4D-EnVar to one implementation of the EnKF. We might have obtained different results by using, for example, a different relative weighting of the static and time-varying background error covariances in the cost function of the Hybrid filter (see Section 2.2), or by further optimizing the EnKF parameters. Adopting another distinct DA method might also have yielded different results in this regard.

Perhaps the strongest robust conclusion of our study is that utilizing even simple types of stochastic parameterizations (SPs) in the forecast model can have stronger and generally beneficial impacts on forecast skill than tinkering with other elements of current forecast systems. However, even this conclusion comes with a caveat that we did not exhaustively investigate forecast sensitivities to other types of stochastic parameterizations. Nonetheless, the main positive result from including stochastic parameterizations seems clear.

We end with a cautionary note that state-of-the-art forecast systems are now sufficiently advanced and finely tuned that establishing the impacts of forecast system changes on forecast skill with statistical confidence requires careful numerical experimentation with large forecast ensemble sizes. The fact that even with 8,000 (= 100 forecast cases × 80 ensemble members for each case) 7-day forecasts in each of our four forecast sets (Control, Denial, EnKFonly, noSP), the apparently large impacts on the extratropical biases in Figs. 9 and 10 turned out to be not statistically significant in the Northern Hemisphere upper tropospheric waveguide provides a sobering reminder in this regard.

Appendix A

To test the statistical significance of the forecast differences in Figs. 9 and 10, we used the Student's *t* test (see Fig. 11 for their *t* values), assuming that the variables are normally distributed. Specifically, at each gridpoint we computed the t-statistic

$$t = \frac{\overline{x_1} - \overline{x_2}}{\left(\frac{\sigma_1^2}{n_1^*} + \frac{\sigma_2^2}{n_2^*}\right)^{1/2}},$$

where $\overline{x_1}$ and $\overline{x_2}$ are the means of 8,000 (= 100 forecast cases × 80 ensemble members/forecast case) valid forecast values from two different forecast sets, σ_1^2 and σ_2^2 are the variances of the 8,000 values in the two forecast sets, and n_1^* and n_2^* are the estimated degrees of freedom (DOF) or effective sample sizes.

The DOF are smaller than 8,000, because the I=80 ensemble values for each forecast case are not truly independent, and the J=100 forecast cases also have some serial dependence since they are initialized only 12 hours apart. We estimated the DOF as follows. Let z_{ij} be the forecast from the *i*-th ensemble member and *j*-th forecast case. One can group z_{ij} by ensemble member or case number so that

$$\{z_{ij}\} = \{x_i\} = \{y_j\},$$

where x_i is the case series of the *i*-th ensemble member, and y_j is the ensemble member series of the *j*-th case. One can think of x and y as the row and column vectors, respectively, of the matrix z. Then one can write

$$Var\left(\sum_{i=1}^{I} x_i\right) = \sum_{i=1}^{I} Var(x_i) + \sum_{i \neq k} Cov(x_i, x_k).$$

This variance has two contributions: 1) the sum of the variances of the individual ensemble members, and 2) the sum of covariances between any two distinct ensemble members. This may also be expressed as

$$Var\left(\sum_{i=1}^{I} x_i\right) = Var(IM_x) = I^2 Var(M_x),$$

where $M_x = \frac{1}{I}\sum_{i=1}^{I} x_i$ is the case series of the ensemble means. By combining the two equations above, and assuming that all the z_{ij} are independent and identically distributed (i.i.d.), the variance of the ensemble-mean forecasts, from the Law of Large Numbers (LLN), may be written as

$$Var(M_x) = \frac{\sum_{i=1}^{I} Var(x_i) + \sum_{i \neq k} Cov(x_i, x_k)}{I^2} = \frac{Var(z_{ij})}{I}.$$

However, the z_{ij} are not independent, because of the non-zero covariance between any two distinct ensemble members ($\sum_{i \neq k} Cov(x_i, x_k) \neq 0$). If positive, this covariance makes the ratio

$$r_x = \frac{[\sum_{i=1}^{I} Var(x_i)]/I^2}{Var(z_{ij})/I} = \frac{[\sum_{i=1}^{I} Var(x_i)]/I}{Var(z_{ij})}$$

less than 1. The DOF in the ensemble member dimension (i.e. the effective ensemble size) is then not I but $I \times r_x$ since

$$Var(M_x) = \frac{Var(z_{ij})}{I \times r_x}$$

agrees with the LLN. Similarly, the ratio

$$r_y = \frac{\left[\sum_{j=1}^{J} Var(y_j)\right]/J}{Var(z_{ij})},$$

provides an estimate of the dependency among the different forecast cases. The overall DOF is then $(I \times r_x) \times (J \times r_y) = 8{,}000 \times r_x \times r_y$.

Fig. A1 shows maps of $Var(z_{ij})$, $\sum_{i=1}^{I} Var(x_i)\,/I$, and $\sum_{j=1}^{J} Var(y_j)\,/J$ for the spatially smoothed Day-7 ω_{500hPa} Control forecasts. If all the forecasts were independent, the three maps would be identical. The results show that r_x is a nearly uniform 0.8 everywhere over the globe, while r_y is generally between 0.3 and 0.9. The overall DOF ω_{500hPa} in the Control forecasts is thus generally between 2,500 and 5,000 for our samples of size 8,000.

The variance of the ensemble members is clearly representative of the total variance over the whole globe, except that the magnitude is smaller because the ensemble members are still not completely independent by Day 7 (Fig. A1 middle). On the other hand, the case variance is not as representative, and the variance ratios are especially noisy in tropical areas (Fig. A1 bottom).

Appendix B

The RMSEs in this study were defined as the square root of case-mean and area-mean squared errors of ensemble-mean forecasts with respect to *truth* (see Section 2.2 and 3.1). Because parametric forms of the probability distributions of RMSEs or RMSE differences (hereafter ΔRMSEs) are generally unknown, we used a Bootstrap method to estimate the sampling distributions of ΔRMSEs to assess the significance of ΔRMSEs obtained between any two forecast sets. To this end we combined the 100 forecast cases in each set into a pool of 200 cases. By randomly drawing with replacement from the pool, two new separate 100-case samples were made, and their ΔRMSE was calculated. Repeating this process 1000 times yielded 1000 values of ΔRMSE for estimating the sampling ΔRMSE distribution. The statistical significance of the actual ΔRMSE was then judged by whether it ranked above the 97.5 percentile or below the 2.5 percentile of this constructed ΔRMSE distribution for a two-sided statistical test. This process was repeated for each 12-hourly forecast lead time up to 168 hours (7 days).

Figs. B1-B3 show global and regional ΔRMSEs between the Control and the other three (Denial, EnKFonly, and noSP) forecasts, corresponding to Figs. 2, 3, and 7 respectively, as well as the 97.5% and 2.5% percentiles of the ΔRMSEs of their respective sampling distributions. Fig. B1 shows that the Control global RMSEs are significantly smaller than the Denial only for ξ_{200hPa} and ω_{500hPa} in the first 24 hours of the forecasts, confirming that the ENRR observations only benefit short-term forecasts at smaller spatial scales. The general pattern in Figs. B1-B3 shows that Hybrid initialization (Control forecasts) significantly lowers the RMSEs in the first few days, compared to EnKF initialization (EnKFonly forecasts). Also, using SPs (Control forecasts) significantly lowers the RMSEs in the later part of the 7-day forecast evolution, compared to not using SPs (noSP forecasts). The exceptions are AP12HR ΔRMSEs between 60ºS and 60ºN (Fig. B1f), which do not ever exceed the confidence interval, and Z_{200hPa} $\Delta RMSE_{Control\text{-}noSP}$ (Fig. B3d), which shows larger errors when using SPs especially in the tropics.

Acknowledgments

This research was supported by the Physical Sciences Division of NOAA's Earth System Research Laboratory. Support was also provided by the NOAA Climate Program Office. Computing was performed on NOAA's Remotely Deployed High-Performance Computing Systems. The scientific results and conclusions, as well as any views or opinions expressed herein, are those of the authors and do not necessarily reflect the views of NOAA or the Department of Commerce.

References

Anderson, J. L., and N. Collins (2007), Scalable implementations of ensemble filter 247 algorithms for data assimilation, Journal of Atmospheric and Oceanic Technology, 248 24 (8), 1452–1463, doi:10.1175/JTECH2049.1.

Ashouri H., K. Hsu, S. Sorooshian, D. K. Braithwaite, K. R. Knapp, L. D. Cecil, B. R. Nelson, and O. P. Prat, 2015: PERSIANN-CDR: Daily Precipitation Climate Data Record from Multi-Satellite Observations for Hydrological and Climate Studies. Bull. Amer. Meteor. Soc., doi: 10.1175/BAMS-D-13-00068.1.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, doi:10.1038/nature14956

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

Bloom, S. C., L. L. Takacs, A. M. da Silva, and D. Ledvina, 1996: Data Assimilation Using Incremental Analysis Updates. *Mon. Weather Rev.*, 124, 1256-1271.

Buehner, M., J. Morneau, and C. Charette, 2013: Four-dimensional ensemble-variational data assimilation for global deterministic weather prediction. *Nonlinear Processes Geophys.*, **20**, 669–682.

Campbell, W.F., C.H. Bishop, and D. Hodyss, 2010: Vertical Covariance Localization for Satellite Radiances in Ensemble Kalman Filters. *Mon. Wea. Rev.,* **138**, 282–290, https://doi.org/10.1175/2009MWR3017.1.

Courtier, P., J.-N. Thépaut, and A. Hollingsworth, 1994: A strategy for operational implementation of 4D-Var, using an incremental approach. *Quarterly Journal of the Royal Meteorological Society*, Vol. **120**, Issue 519, Part B, 1367-1387.

Dole, R.M., and Co-Authors, 2018: Advancing Science and Services during the 2015-16 El Niño: The NOAA El Niño Rapid Response Field Campaign. *Bulletin of the American Meteorological Society*, Vol. **99**(5), 975-1002, DOI:10.1175/BAMS-D-16-0219.1.

Efron, B., 1982: *The Jackknife, the Bootstrap, and Other Resampling Plans*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 92 pp.

Efron, B. and R. J. Tibshirani, 1993: *An Introduction to the Bootstrap.* Chapman & Hall, New York, 456 pp.

Environmental Modeling Center, 2003: The GFS Atmospheric Model. NCEP Office Note 442, Global Climate and Weather Modeling Branch, EMC, Camp Springs, Maryland.

Evensen, G., 2003: The Ensemble Kalman Filter: theoretical formulation and practical implementation. *Ocean Dynamics*, 53:343-367, doi:10.1007/s10236-003-0036-9.

Gaspari, G., and S. E. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.*, **125**, 723-757, doi:10.1002/qj.49712555417.

Hall, N.M.J., and P. D. Sardeshmukh, 1998: Is the Time-Mean Northern Hemisphere Flow Baroclinically Unstable? *J. Atmos. Sci.*, Vol. **55**, 41-56.

Houtekamer, P. L., and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.*, **126**, 796-811, doi:10.1175/1520-0493(1998)126<0796:DAUAEK>2.0.CO;2.

Huffman, D., Bolvin, D. Braithwaite, K. Hsu, R. Joyce, P. Xie, 2014: Integrated Multi-satellite Retrievals for GPM (IMERG), version 4.4. NASA's Precipitation Processing Center, accessed 31 March, 2015, ftp://arthurhou.pps.eosdis.nasa.gov/gpmdata/

Kleist, D.T. and K. Ide, 2015: An OSSE-Based Evaluation of Hybrid Variational–Ensemble Data Assimilation for the NCEP GFS. Part II: 4DEnVar and Hybrid Variants. *Mon. Wea. Rev.,* **143**, 452–470, https://doi.org/10.1175/MWR-D-13-00350.1

Lei, L., and J. S. Whitkaer, 2015: Model space localization is not always better than observation space localization for assimilation of satellite radiances. *Mon. Wea. Rev.*, **143**, 3948-3955, doi: 10.1175/MWR-D-14-00413.1.

Lei, L., and J. S. Whitaker, 2016: A Four-Dimensional Incremental Analysis Update for the Ensemble Kalman Filter. *Mon. Wea. Rev.*, **144**, 2605-2621, DOI: 10.1175/MWR-D-15-0246.1.

Lei, L., J. S. Whitaker and C. Bishop, 2018: Improving assimilation of radiance observations by implementing model space localization in an ensemble Kalman filter. *Journal of Advances in Modeling Earth Systems*, **10**, 12, 3221-3232. DOI: 10.1029/2018MS001468.

Leutbecher, M., and Co-authors, 2017: Stochastic representations of model uncertainties at ECMWF: state of art and future vision. *Quarterly J. Royal Met. Soc.*, **143**, 707, 2315-2339. DOI: 10.1002/qj.3094

Lewis, J. M. and J. C. Derber, 1985: The use of adjoint equations to solve a variational adjustment problem with advective constraints. *Tellus*, **37A**, 4, 309-322.

Palmer, T. N., R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. J. Shutts,M. Steinheimer, and A.Weisheimer, 2009: Stochastic parametrization and model uncertainty. ECMWF Tech. Memo. 598, 42 pp.

Sardeshmukh, P. D., G. P. Compo, M. C. Penland, 2015: Need for Caution in Interpreting Extreme Weather Statistics. *J. Climate*, Vol. **28**, 23, 9166-9187.

Sardeshmukh, P. D., 2005: Issues in stochastic parameterization. *Proc. Workshop on Representation of Sub-Grid Processes Using Stochastic-Dynamic Models*, Shinfield Park, Reading, ECMWF, 5–12.

Sardeshmukh, P. D. and Hoskins, B. J., 1984: Spatial Smoothing on the Sphere. Mon. Wea. Rev., Vol. **112**, 2524-2529.

Sardeshmukh, P.D., and B. J. Hoskins, 1988: The Generation of Global Rotational Flow by Steady Idealized Tropical Divergence. *J. Atmos. Sci.*, Vol. **45**, No. 7, 1228-1251.

Shutts, G., M. Leutbecher, A. Weisheimer, T. Stockdale, L. Isaksen, and M. Bonavita, 2011: Representing model uncertainty: stochastic parameterizations at ECMWF, *ECMWF Newsletter*, **129**, 19-24.

Sorooshian, S., K. Hsu,D. Braithwaite,H. Ashouri, and NOAA CDR Program (2014): NOAA Climate Data Record (CDR) of Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks (PERSIANN-CDR), Version 1 Revision 1. NOAA National Centers for Environmental Information. doi:10.7289/V51V5BWQ [access date: April 27th, 2017]

Slivinski, L.C., G.P. Compo, J.S. Whitaker, P.D. Sardeshmukh, J.H.A. Wang, K. Friedman, and C. McColl, 2018: What is the impact of additional tropical observations on a modern data assimilation system? *Mon. Wea. Rev.*, in review.

Takacs, L. L., M. J. Suárez, and R. Todling, 2018: The Stability of Incremental Analysis Update. *Mon. Wea. Rev.,* **146**, 3259–3275, https://doi.org/10.1175/MWR-D-18-0117.1.

Ting, M., and P.D. Sardeshmukh, 1993: Factors Determining the Extratropical Response to Equatorial Diabatic Heating Anomalies. *J. Atmos. Sci.*, Vol. **50**., No. 6, 907-918.

Tompkins, A. M., and J. Berner, 2008: A stochastic convective approach to account for model uncertainty due to unresolved humidity variability. *J. Geophys. Res.*, Vol. **113**, D18101, doi:10.1029/2007JD009284.

Weaver, A., and P. Courtier, 2001: Correlation modelling on the sphere using a generalized diffusion equation. *Quart. J. Roy. Meteorol. Soc.*, **127**, 1815–1846.

Whitaker, J. S., and T. M. Hamill (2002), Ensemble data assimilation without 289 perturbed observations, *Mon. Weather Rev.*, **130** (7), 1913–1924, doi:10.1175/1520-0493(2002)130<1913:EDAWPO>2.0.CO;2.

Figure Captions

Figure 1. Schematic depiction of the 7-day forecasts generated and verification period used. Each arrow represents one forecast case, and only the portion in the verification period is evaluated for this study. Note that there are 80 members in the ensemble forecast for each forecast case.

Figure 2. Global RMSEs of the Control (solid gray), Denial (dashed blue), EnKFonly (dotted green) and noSP forecasts (dash-dot red), determined with respect to the Control

analyses for (a) 200hPa heights (Z_{200hPa}), (b) 200hPa vorticity (ξ_{200hPa}), (c) 500hPa vertical p-velocity (ω_{500hPa}), (d) precipitable water (PWAT), and (e) 2-meter air temperature (T_{2m}). (f) The RMSE of 12-hr accumulated precipitation (AP12HR) averaged in the 20°S to 20°N domain (thin upper curves) and the 60°S to 60°N domain (thick lower curves), determined with respect to NASA GPM observational dataset. Note the ordinate for the precipitation RMSE starts at 6 mm.

Figure 3. Domain ω_{500hPa} RMSEs of the Control, Denial, EnKFonly and noSP forecasts with respect to the Control analyses in the (a) Northern Hemisphere (20°N-90°N), (b) Southern Hemisphere (20°S-90°S), (c) Tropics (20°S-20°N), and (d) Contiguous United States (CONUS; 125°W-66°W, 24°N-50°N).

Figure 4. (a) The ω_{500hPa} RMSEs of the Day-7 Control forecasts; (b) The differences of the ω_{500hPa} RMSEs between the Day-7 Control and Denial forecasts; (c) Similar to (b), but between the Control and EnKFonly forecasts; (d) Similar to (b), but between the Control and noSP forecasts.

Figure 5. (a) The AP12HR RMSEs of the Control forecasts with respect to independent NASA GPM product at the end of Day 7; (b) The AP12HR RMSE differences between the Control and Denial forecasts at the end of Day 7; (c) Similar to (b), but between the Control and EnKFonly forecasts; (d) Similar to (b), but between the Control and noSP forecasts. The valid geographic domain is between 60°S and 60°N. If there exist only missing values

in a grid box (0.5°×0.5°) at any moment during the verification period, that box is painted gray in (b)-(d).

Figure 6. As in Fig. 4, except for T_{2m}.

Figure 7. As in Fig. 3, but for Z_{200hPa}.

Figure 8. (a) The Z_{200hPa} RMSEs of the Control forecasts at the end of Day 7; (b) The Z_{200hPa} RMSE differences between the Control and Denial forecasts at the end of Day 7; (c) Similar to (b), but between the Control and EnKFonly forecasts; (d) Similar to (b), but between the Control and noSP forecasts.

Figure 9. (a) Bias of case-mean ensemble-mean Day-7 Z_{200hPa} Control forecasts with respect to the Control analyses; (b) Difference of case-mean ensemble-mean Control and Denial forecasts; (c) Difference of case-mean ensemble-mean Control and EnKFonly forecasts; (d) Difference of case-mean ensemble-mean Control and noSP forecasts. Note that the contour interval in panel (a) is 4.5 times that in the other panels.

Figure 10. As in Fig. 9 except for ω_{500hPa}. Note that the contour interval in panel (a) is five times that in the other panels. The additional thick black curves in the extratropical Northern Hemisphere enclose the region of 200hPa mean zonal winds stronger than 30m/s in the Control analysis, which is a good proxy of the extratropical baroclinic waveguide.

Figure 11. Left panels: The Student's t scores for the Day-7 Z_{200hPa} bias differences between (top) the Control and Denial forecasts, (middle) the Control and EnKFonly forecasts, and (bottom) the Control and noSP forecasts. A value of ±1.645 is 10% significant in two-tailed test, ±1.96 is 5% significant, and ±2.576 is 1% significant. Right Panels: Similar to left panels but for ω_{500hPa} fields. The thick black 30m/s contour of the 200hPa zonal winds in the Northern Hemisphere shows the approximate location of the upper tropospheric jet stream waveguide, as in Fig. 10.

Figure A1. (top) The total variance of the spatially smoothed Day-7 ω_{500hPa} Control forecasts; (middle) the sum of the variances within the individual ensemble members across the cases, divided by group size 100; (bottom) the sum of the variances within the individual cases across the ensemble members, divided by group size 80 (color shaded), and the ratio of the values of the sum of the variances to the total variance (contours). The contour interval in the bottom panel is 0.1, and the 1 contour is thickened. The variance ratio in the middle panel is ~0.79 almost uniformly over the globe and hence no contour is plotted. Note that if all the forecasts were independent, the values in the middle and bottom panels would be equal to those in the top panel.

Figure B1. Global RMSE differences between the Control and Denial forecasts (solid blue), between the Control and EnKFonly forecasts (solid green), and between the Control and noSP forecasts (solid red) for (a) 200hPa geopotential heights (Z_{200hPa}), (b) 200hPa vorticity (ξ_{200hPa}), (c) 500hPa vertical p-velocity (ω_{500hPa}), (d) precipitable water (PWAT), and (e) 2-meter air temperature (T_{2m}). (f) Similar to panel (a)-(d), except for 12-hr accumulated precipitation (AP12HR) RMSE differences averaged in the 20ºS to 20ºN (thin curves) and the 60ºS to 60ºN (thick curves) latitude domains. The dotted lines represent the 2.5% (below ΔRMSE=0) and 97.5% (above ΔRMSE=0) of the constructed distributions for Control-Denial (blue), Control-EnKFonly (green), and Control-noSP (red), derived from the Bootstrap method.

Figure B2. Similar to Fig. B1, except for ω_{500hPa} in (a) Northern Hemisphere, (b) Southern Hemisphere, (c) Tropics, and (d) Contiguous United States. See Fig. 3 and context for domain definitions.

Figure B3. Similar to Fig. B2, except for Z_{200hPa}.

Table 1: List of forecast ensembles generated

Label	**Initial Condition**	**Data Assimilation Method**	**Forecast Model**
Control	Includes ENRR obs	Hybrid	Includes Stochastic Physics
Denial	Excludes ENRR obs	Hybrid	Includes Stochastic Physics
EnKFonly	Includes ENRR obs	EnKF	Includes Stochastic Physics
noSP	Includes ENRR obs	Hybrid	No Stochastic Physics

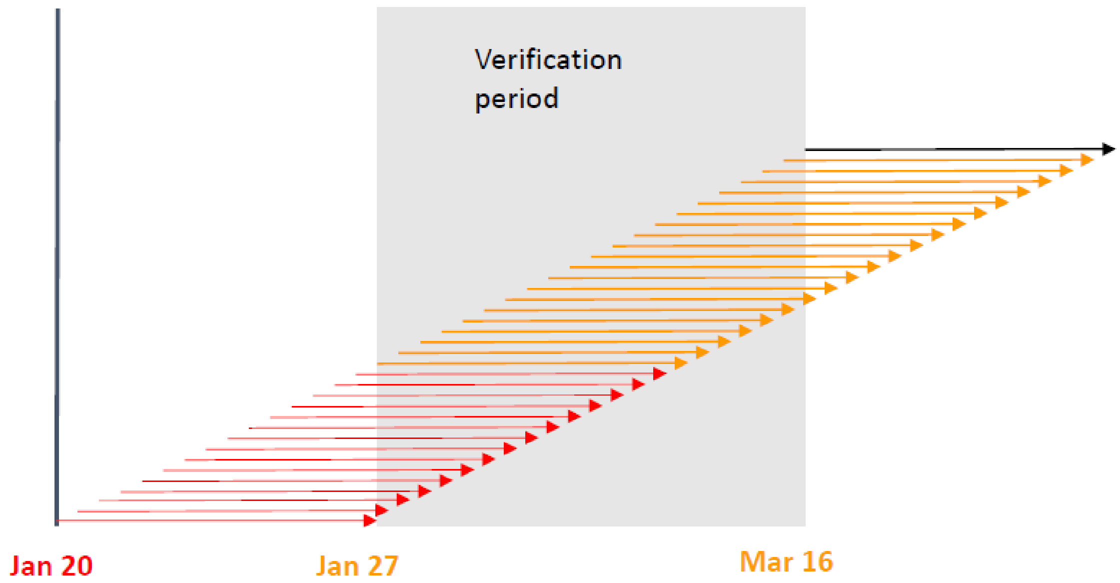

Figure 1. Schematic depiction of the 7-day forecasts generated and verification period used. Each arrow represents one forecast case, and only the portion in the verification period is evaluated for this study. Note that there are 80 members in the ensemble forecast for each forecast case.

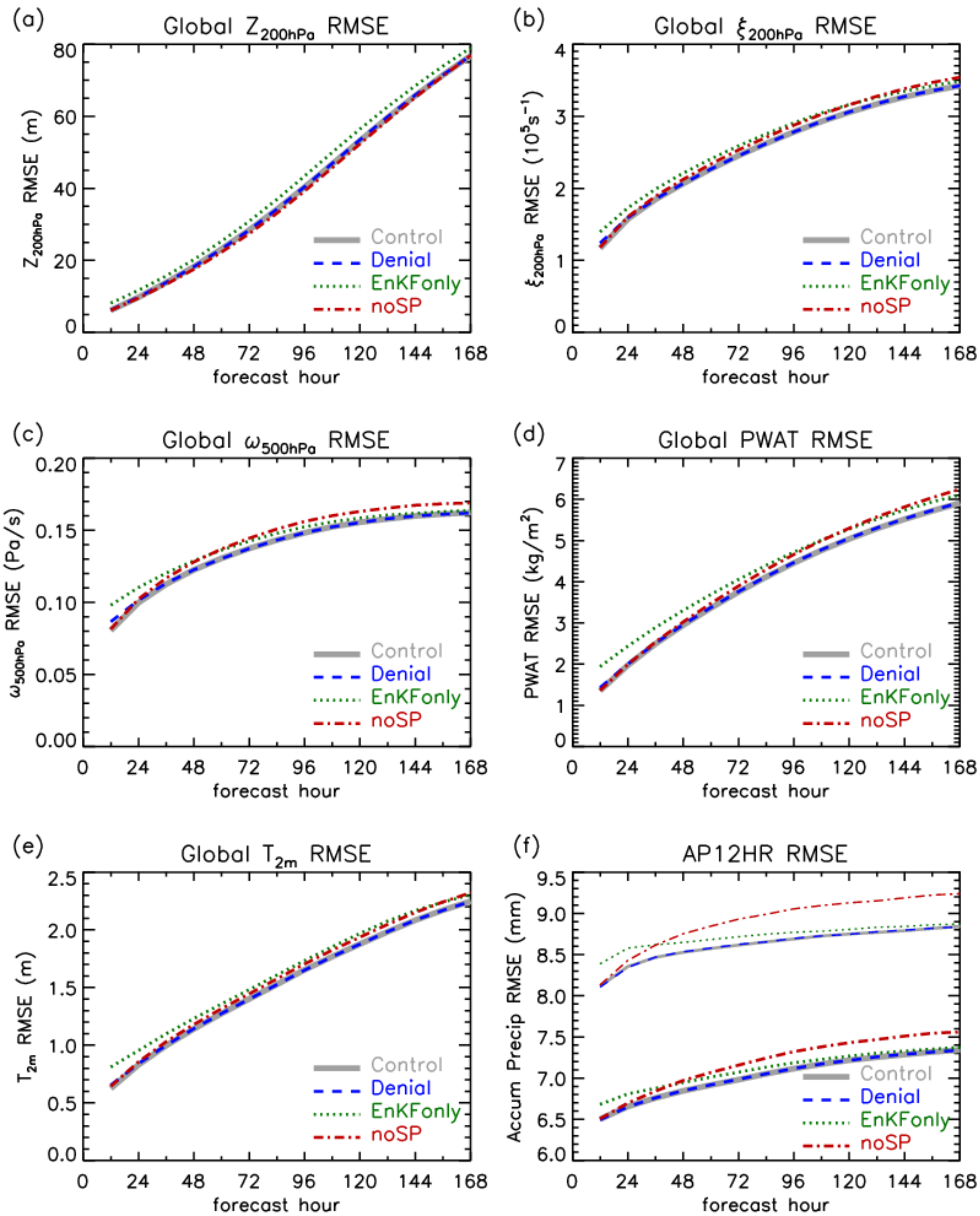

Figure 2. Global RMSEs of the Control (solid gray), Denial (dashed blue), EnKFonly (dotted green) and noSP forecasts (dash-dot red), determined with respect to the Control analyses for global (a) 200hPa heights (Z_{200hPa}),(b) 200hPa vorticity (ξ_{200hPa}), (c) 500hPa vertical p-velocity (ω_{500hPa}), (d) precipitable water (PWAT), and (e) 2-meter air temperature (T_{2m}). (f) The RMSE of 12-hr accumulated precipitation totals in the 20ºS to

20ºN domain (thin upper curves) and the 60ºS to 60ºN domain (thick lower curves), determined with respect to NASA GPM observational dataset. Note the ordinate for the precipitation RMSE starts at 6 mm.

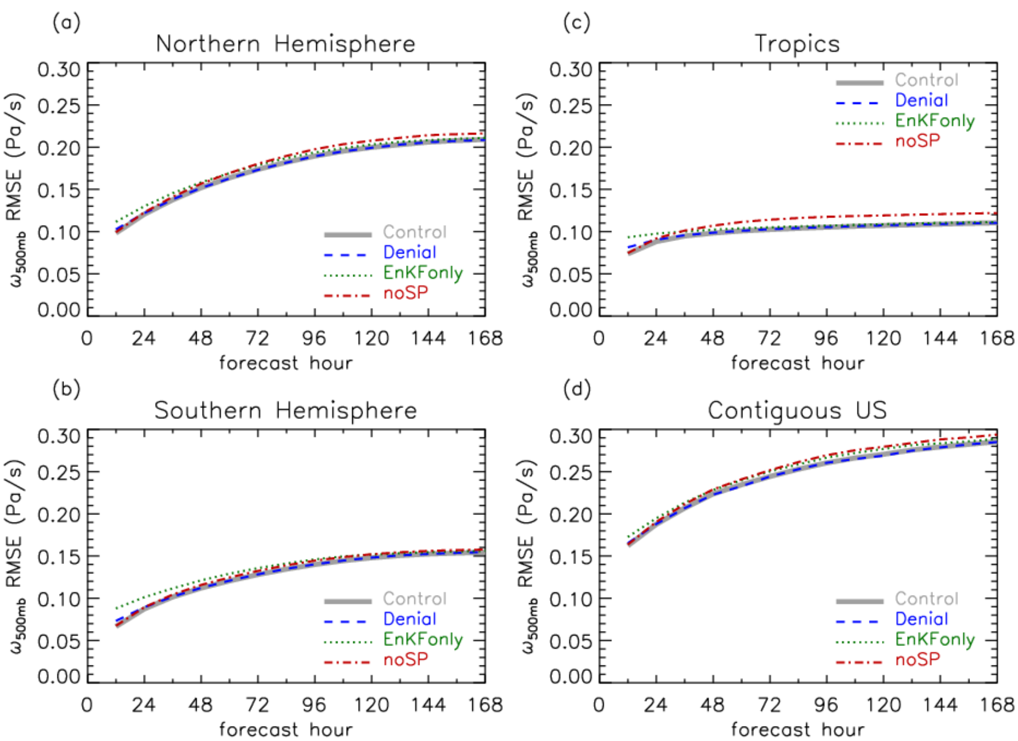

Figure 3. Domain ω_{500hPa} RMSEs of the Control, Denial, EnKFonly and noSP forecasts with respect to the Control analyses in the (a) Northern Hemisphere (20ºN-90ºN), (b) Southern Hemisphere (20ºS-90ºS), (c) Tropics (20ºS-20ºN) and (d) Contiguous United States (CONUS; 125ºW-66ºW, 24ºN-50ºN).

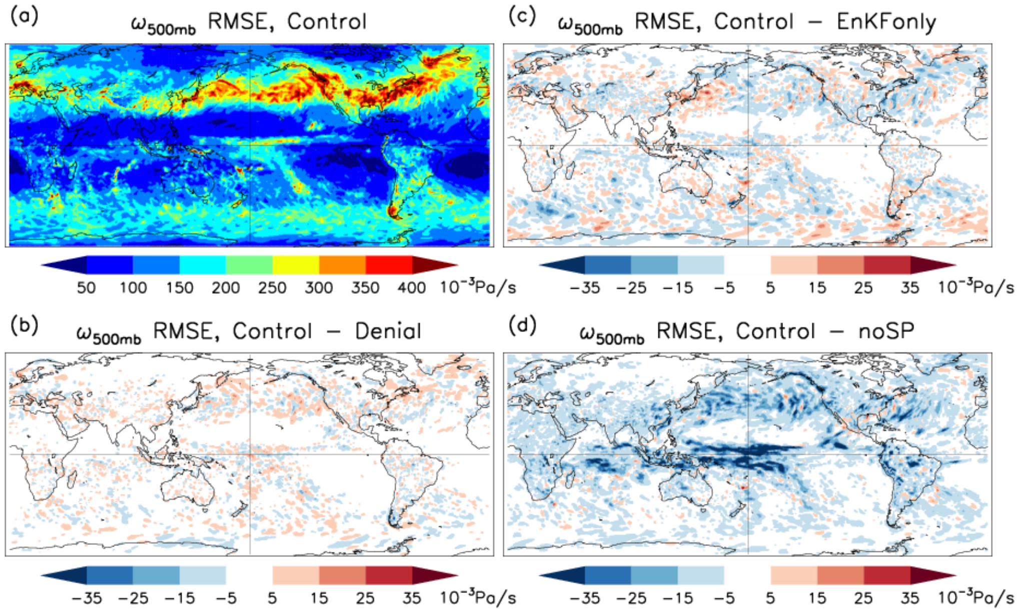

Figure 4. (a) The ω_{500hPa} RMSEs of the Day-7 Control forecasts; (b) The differences of the ω_{500hPa} RMSEs between the Day-7 Control and Denial forecasts; (c) Similar to (b), but between the Control and EnKFonly forecasts; (d) Similar to (b), but between the Control and noSP forecasts.

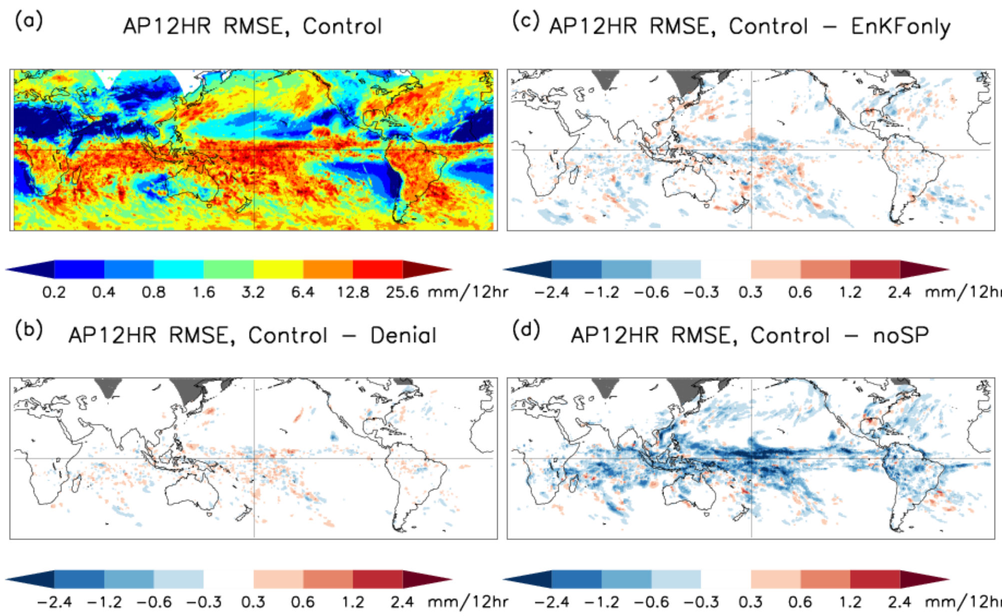

Figure 5. (a) The AP12HR RMSEs of the Control forecasts with respect to independent NASA GPM product at the end of Day 7; (b) The AP12HR RMSE differences between the Control and Denial forecasts at the end of Day 7; (c) Similar to (b), but between the Control and EnKFonly forecasts; (d) Similar to (b), but between the Control and noSP forecasts. The valid geographic domain is between 60ºS and 60ºN. If there exist only missing values in a grid box (0.5º×0.5º) at any moment during the verification period, that box is painted gray in (b)-(d).

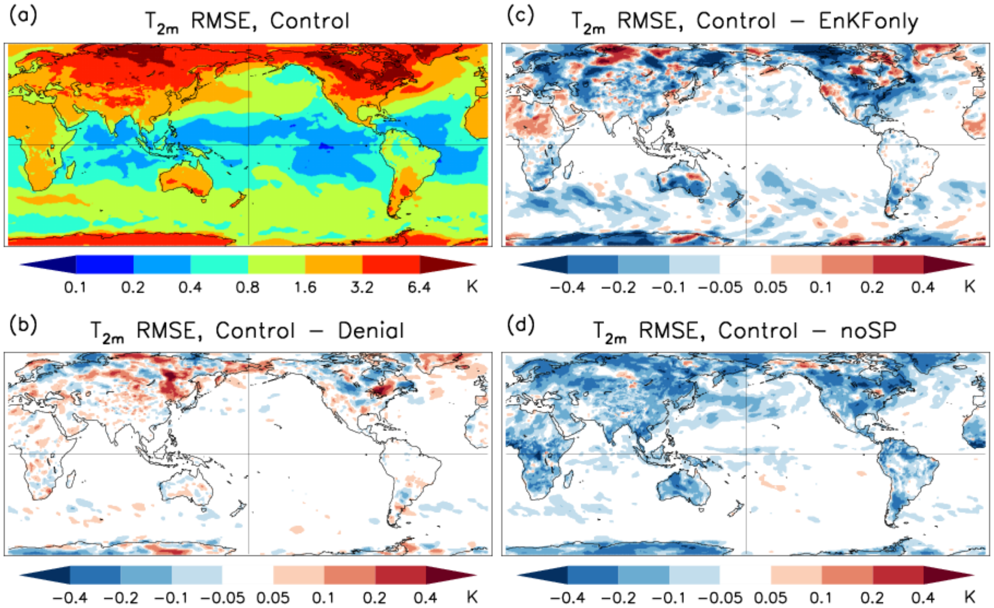

Figure 6. As in Fig. 4, except for T_{2m}.

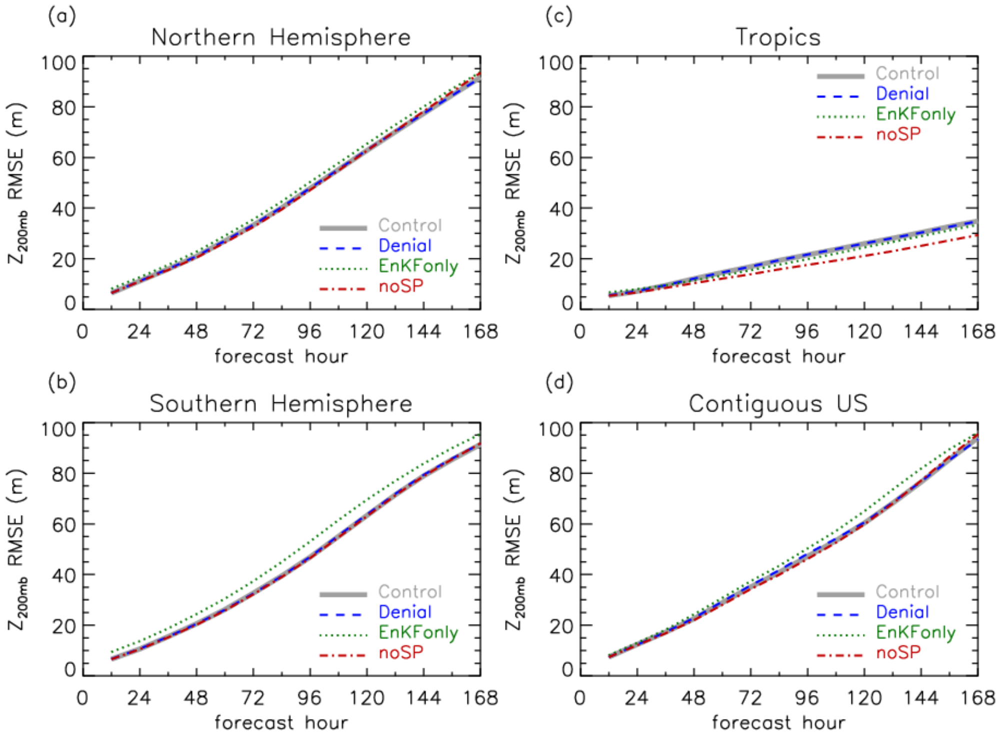

Figure 7. As in Fig. 3, but for Z_{200hPa}.

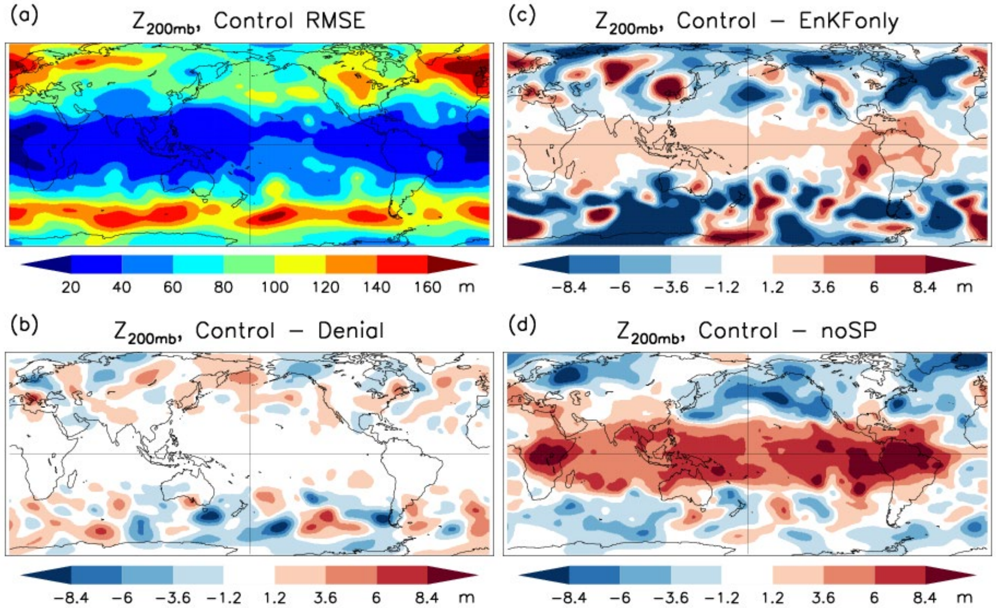

Figure 8. (a) The Z_{200hPa} RMSEs of the Control forecasts at the end of Day 7; (b) The Z_{200hPa} RMSE differences between the Control and Denial forecasts at the end of Day 7; (c) Similar to (b), but between the Control and EnKFonly forecasts; (d) Similar to (b), but between the Control and noSP forecasts.

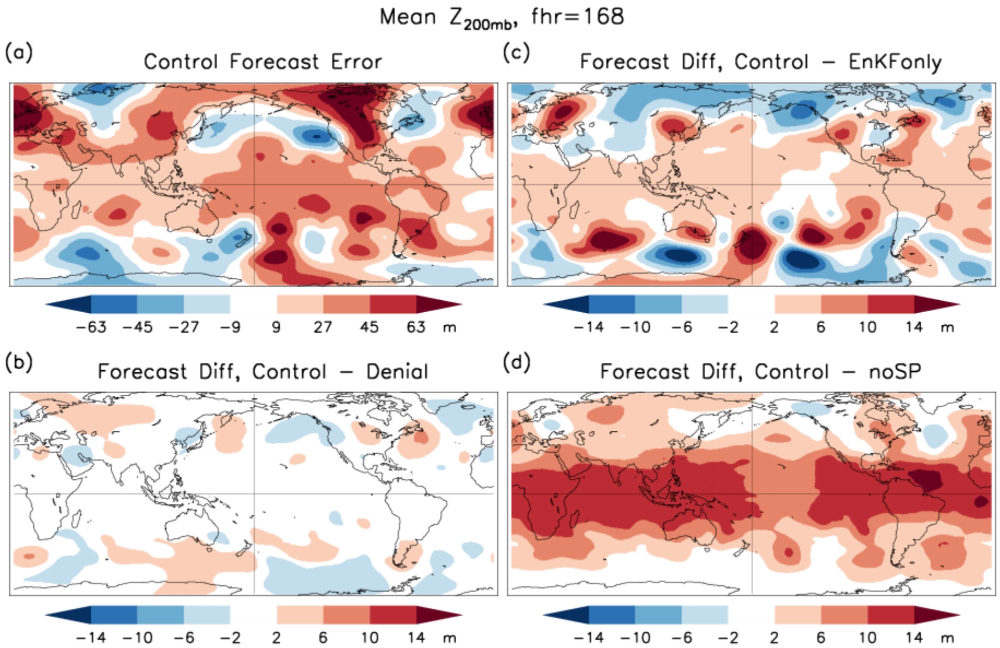

Figure 9. (a) Bias of case-mean ensemble-mean Day-7 Z_{200hPa} Control forecasts with respect to the Control analyses; (b) Difference of case-mean ensemble-mean Control and Denial forecasts; (c) Difference of case-mean ensemble-mean Control and EnKFonly forecasts; (d) Difference of case-mean ensemble-mean Control and noSP forecasts. Note that the contour interval in panel (a) is 4.5 times that in the other panels.

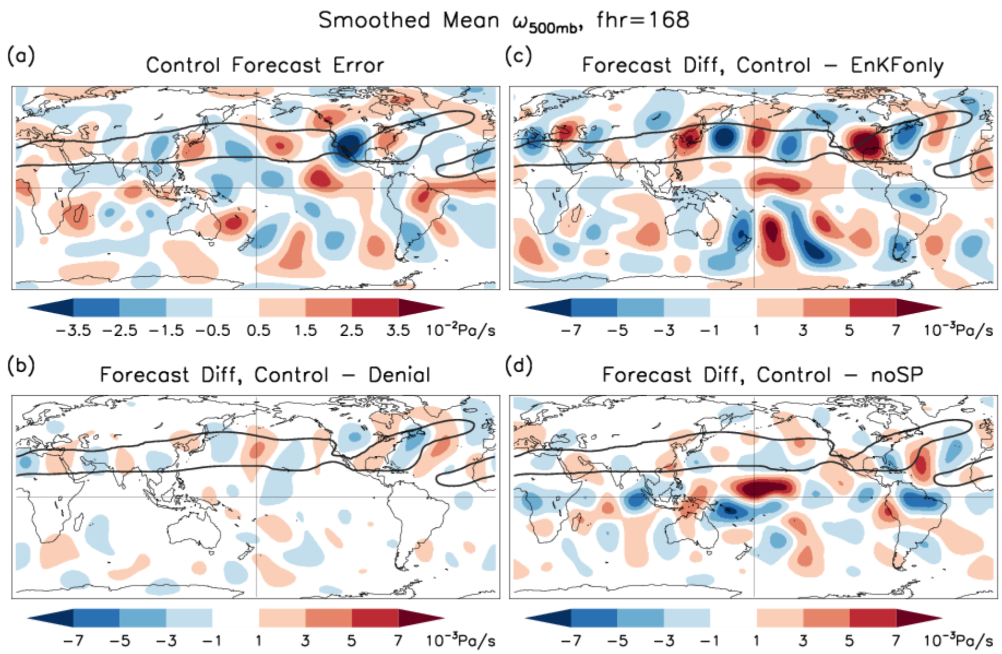

Figure 10. As in Fig. 9, except for ω_{500hPa}. Note that the contour interval in panel (a) is five times that in the other panels. The additional thick black curves in the extratropical Northern Hemisphere enclose the region of 200hPa mean zonal winds stronger than 30m/s in the Control analysis, which is a good proxy of the extratropical baroclinic waveguide.

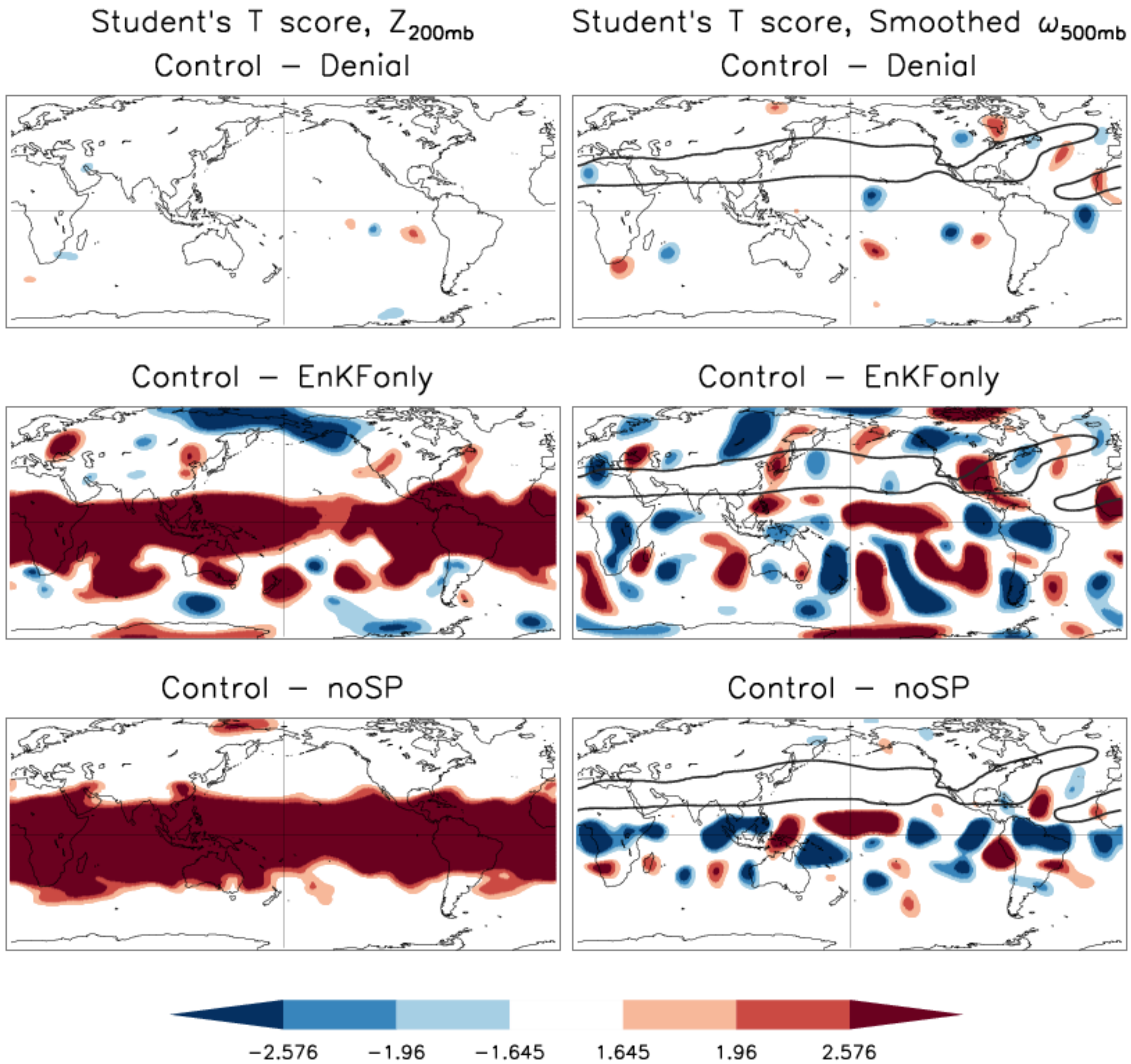

Figure 11. Left panels: The Student's t scores for the Day-7 Z_{200hPa} bias differences between (top) the Control and Denial forecasts, (middle) the Control and EnKFonly forecasts, and (bottom) the Control and noSP forecasts. A value of ±1.645 is 10% significant in two-tailed test, ±1.96 is 5% significant, and ±2.576 is 1% significant. Right Panels: Similar to left panels but for ω_{500hPa} fields. The thick black 30m/s contour of the 200hPa zonal winds in the Northern Hemisphere shows the approximate location of the upper tropospheric jet stream waveguide, as in Fig. 10.

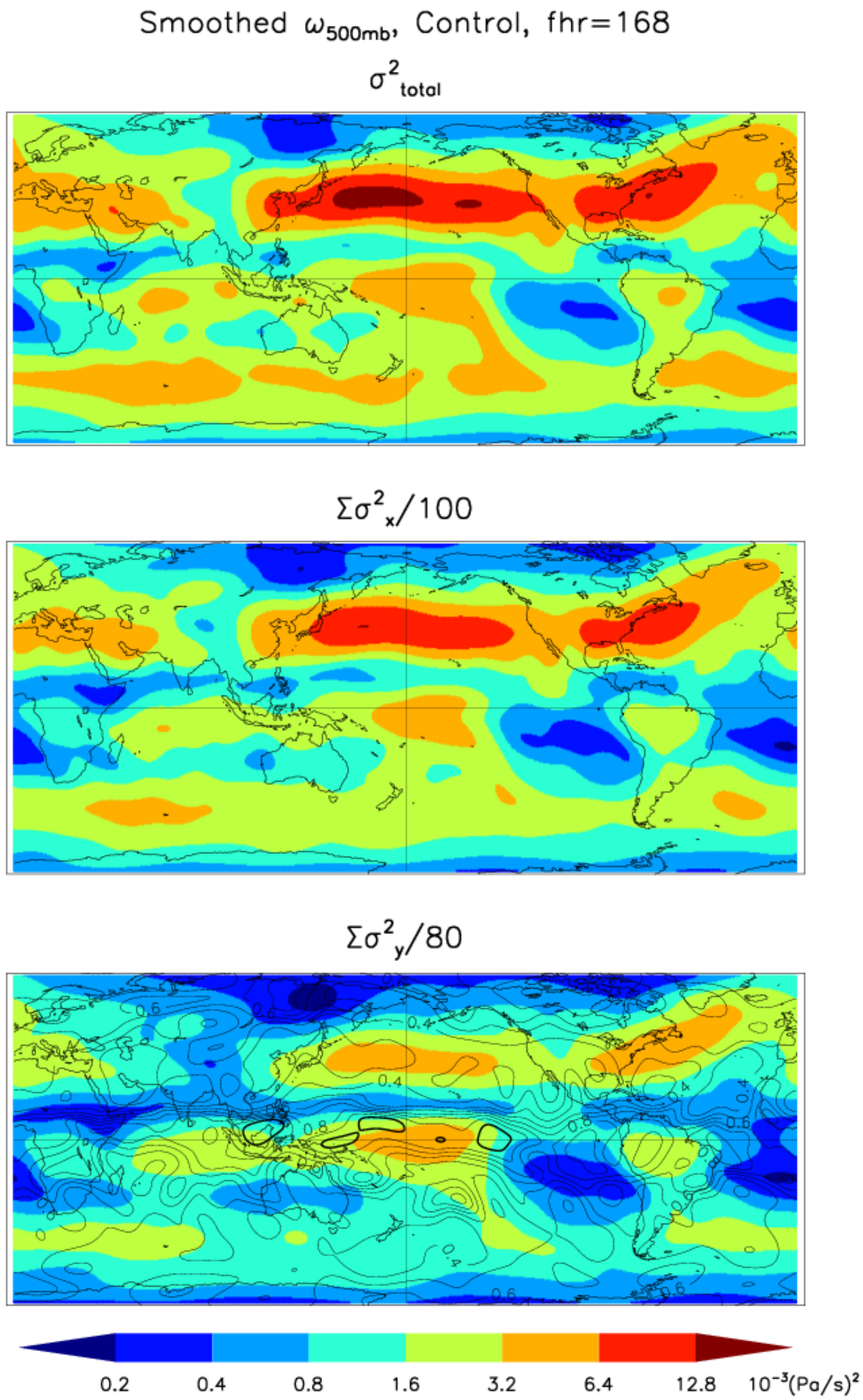

Figure A1. (top) The total variance of the spatially smoothed Day7ω_{500hPa} Control forecasts; (middle) the sum of the variances within the individual ensemble members across the cases, divided by group size 100; (bottom) the sum of the variances within the individual cases across the ensemble members, divided by group size 80 (color shaded), and the ratio of the values of the sum of the variances to the total variance (contours). The contour interval in the bottom panel is 0.1, and the 1 contour is thickened. The variance ratio in the middle panel is ~0.79 almost uniformly over the globe and hence no contour is plotted. Note that if all the forecasts were independent, the values in the middle and bottom panels would be equal to those in the top panel.

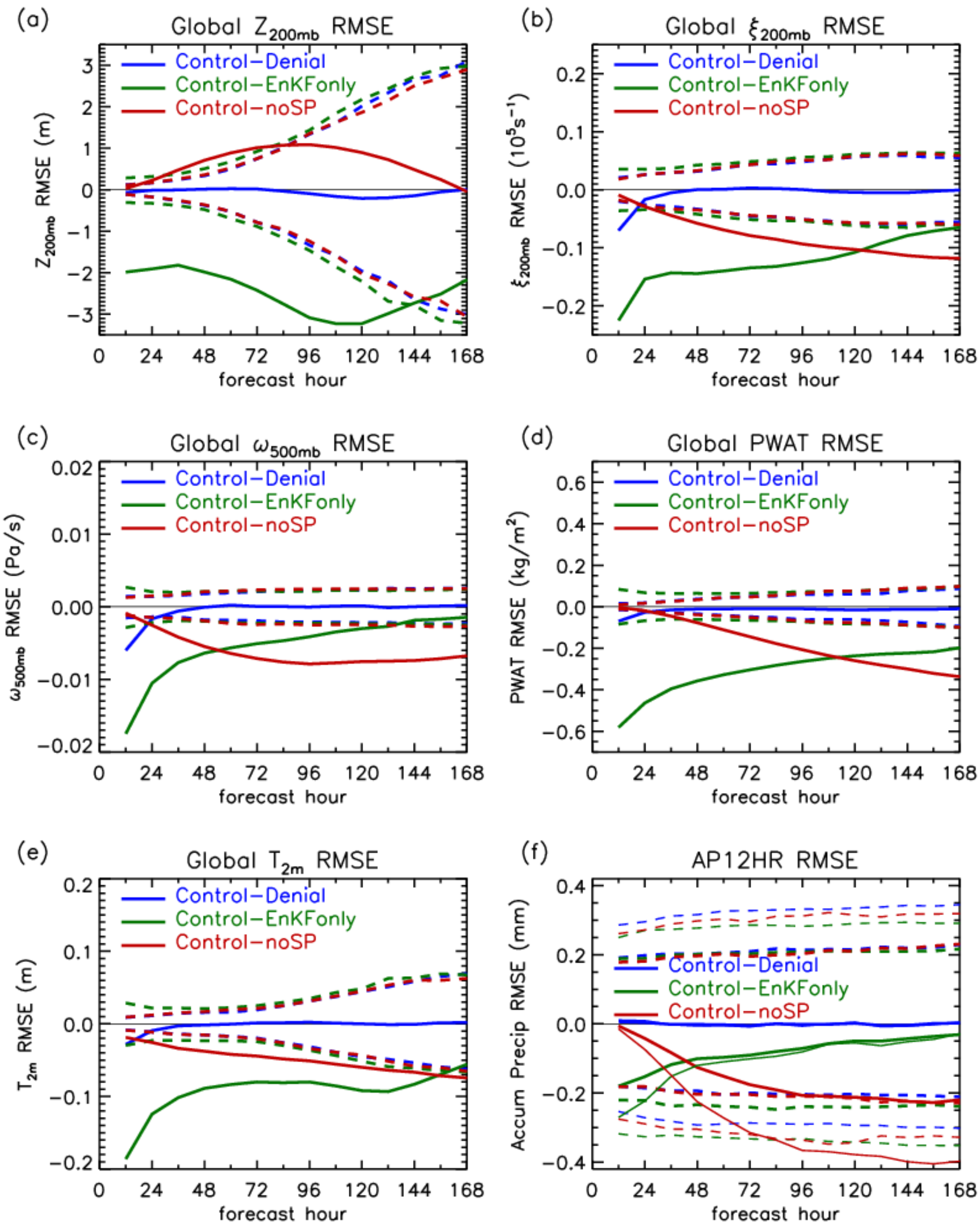

Figure B1. Global RMSE differences between the Control and Denial forecasts (solid blue), between the Control and EnKFonly forecasts (solid green), and between the Control and noSP forecasts (solid red) for (a) 200hPa geopotential heights (Z_{200hPa}), (b) 200hPa vorticity (ξ_{200hPa}), (c) 500hPa vertical p-velocity (ω_{500hPa}), (d) precipitable water (PWAT), and (e) 2-meter air temperature (T_{2m}). (f) Similar to panel (a)-(d), except for 12-hr accumulated precipitation (AP12HR) RMSE differences in the 20^{o}S to 20^{o}N (thin curves) and the 60^{o}S to 60^{o}N (thick curves) latitude domains. The dotted lines represent the 2.5% (below ΔRMSE=0) and 97.5% (above ΔRMSE=0) of the constructed distributions for Control-Denial (blue), Control-EnKFonly (green), and Control-noSP (red), derived from the Bootstrap method.

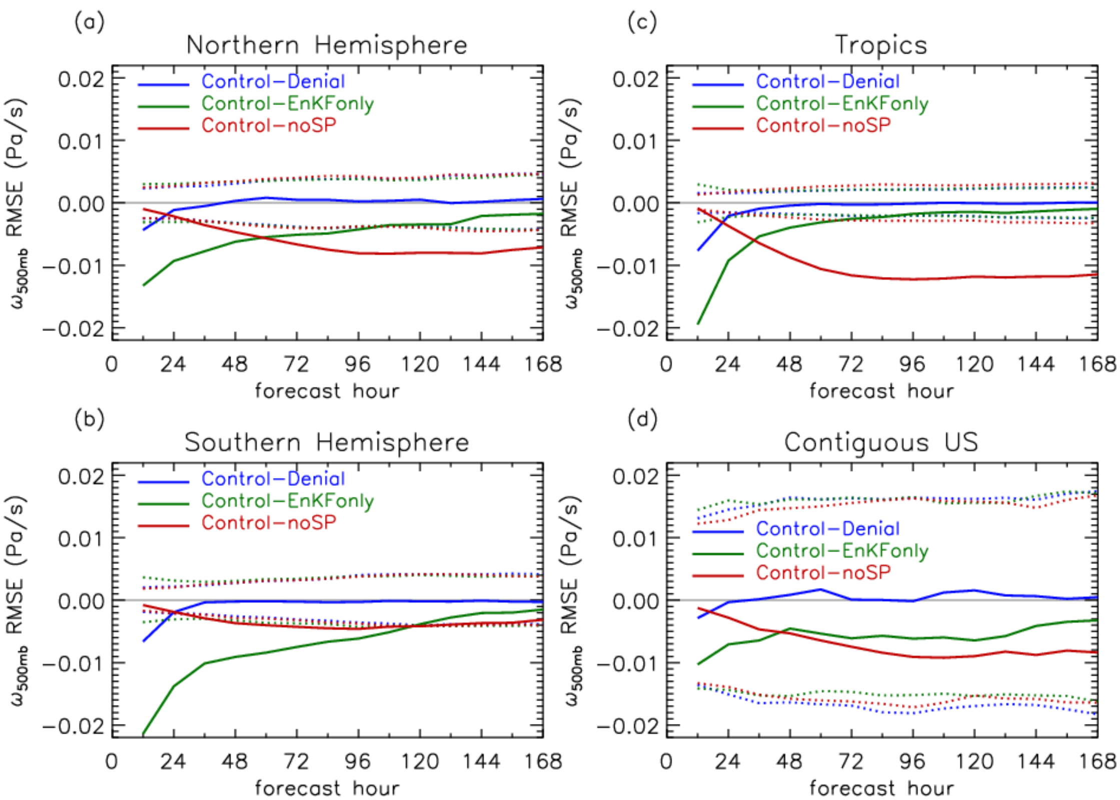

Figure B2. Similar to Fig. B1, except for ω_{500hPa} in (a) Northern Hemisphere, (b) Southern Hemisphere, (c) Tropics, and (d) Contiguous United States. See Fig. 3 and context for domain definitions.

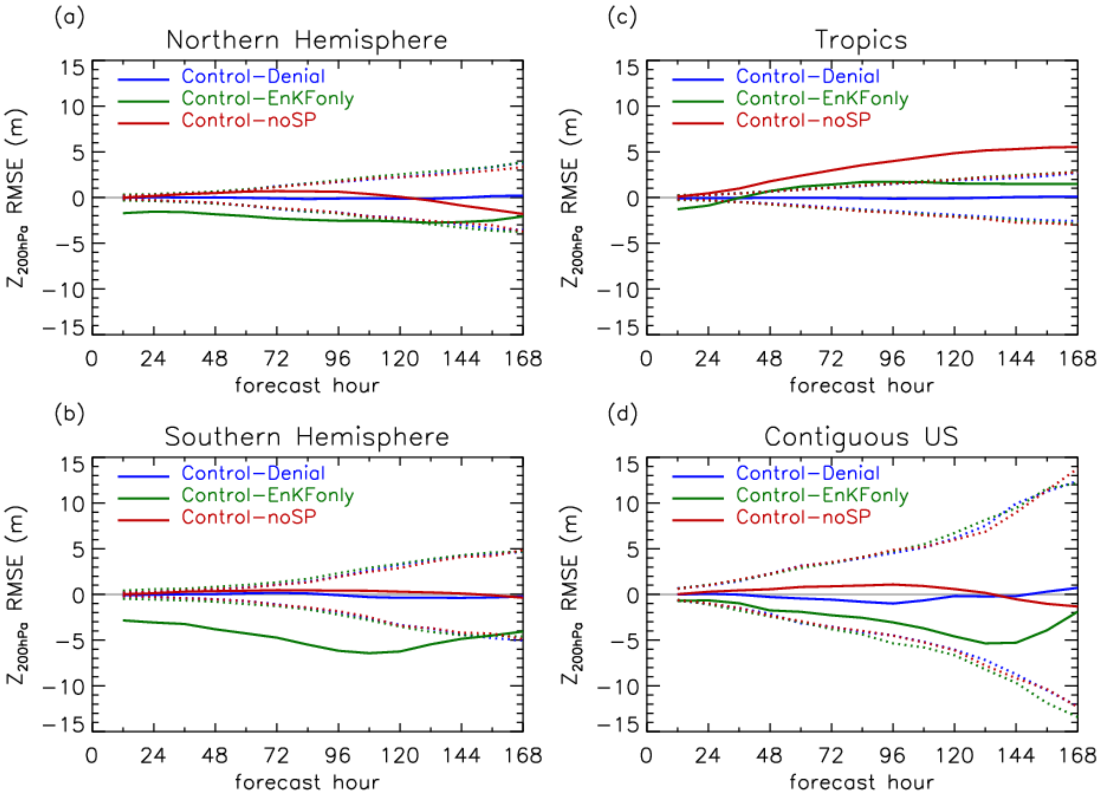

Figure B3. Similar to Fig. B2, except for Z_{200hPa}.