A Comparison of Model Error Representations in Mesoscale Ensemble Data Assimilation

SOYOUNG HA, JUDITH BERNER, AND CHRIS SNYDER

National Center for Atmospheric Research, Boulder, Colorado

(Manuscript received 2 December 2014, in final form 4 June 2015)

ABSTRACT

Mesoscale forecasts are strongly influenced by physical processes that are either poorly resolved or must be parameterized in numerical models. In part because of errors in these parameterizations, mesoscale ensemble data assimilation systems generally suffer from underdispersiveness, which can limit the quality of analyses. Two explicit representations of model error for mesoscale ensemble data assimilation are explored: a multiphysics ensemble in which each member’s forecast is based on a distinct suite of physical parameterization, and stochastic kinetic energy backscatter in which small noise terms are included in the forecast model equations. These two model error techniques are compared with a baseline experiment that includes spatially and temporally adaptive covariance inflation, in a domain over the continental United States using the Weather Research and Forecasting (WRF) Model for mesoscale ensemble forecasts and the Data Assimilation Research Testbed (DART) for the ensemble Kalman filter. Verification against independent observations and Rapid Update Cycle (RUC) 13-km analyses for the month of June 2008 showed that including the model error representation improved not only the analysis ensemble, but also short-range forecasts initialized from these analyses. Explicitly accounting for model uncertainty led to a better-tuned ensemble spread, a more skillful ensemble mean, and higher probabilistic scores, as well as significantly reducing the need for inflation. In particular, the stochastic backscatter scheme consistently outperformed both the multiphysics approach and the control run with adaptive inflation over almost all levels of the atmosphere both deterministically and probabilistically.

1. Introduction

An ensemble Kalman filter (EnKF) is suitable for the mesoscale analysis because it estimates multivariate flow-dependent background error covariance that can capture fast-varying meso- and small-scale features. However, even if “errors of the day” are well described in short-range ensemble forecasts (to be well represented in the background error covariance), the mesoscale analysis is still challenging because of various factors, including the validity of the linear and Gaussian approximation, sampling noise, and unsteadiness in the covariance/correlations between state variables, and error in forecast models, particularly due to the imperfection of the physics parameterization and subgrid-scale uncertainties. That model error can substantially degrade the quality of background error covariance, lead to underestimation of the ensemble spread, and ultimately limit the skill of the subsequent mesoscale forecasts.

To account for errors in the model itself, various model error techniques have been explored in ensemble prediction systems. These include multimodel ensembles, multiphysics ensembles, stochastically perturbed parameterization tendencies (SPPT; Buizza et al. 1999) and stochastic kinetic energy backscatter (SKEB; Shutts 2005; Berner et al. 2009, 2011). Multimodel, multiphysics, or multiparameter ensembles (Charron et al. 2010; Hacker et al. 2011; Berner et al. 2011) use a different version of models or physics parameterizations for each ensemble member to account for model uncertainty due to different underlying assumptions or approximations. In a stochastic approach, SKEB aims at representing an integrated effect of subgrid scales that are not resolved at the specified grid resolution. In its original implementation (Shutts 2005), stochastic perturbations are proportional to an estimated total energy dissipation rate. Berner et al. (2011) use a simplified version, which assumes a constant dissipation rate and relies on the capability of the model to represent the
error growth under less predictable atmospheric conditions. It produces spatially and temporally correlated perturbations and adds them to streamfunction and potential temperature tendencies at each time step. A number of studies demonstrated that the stochastic algorithm improved dispersion and reliability in the ensemble forecast system (Berner et al. 2011, 2015; Charron et al. 2010; Tennant et al. 2011).

In the context of EnKF data assimilation, model error is not limited to errors in the numerical simulation itself but includes all the errors undersampled or underestimated in the background (prior) ensemble. To account for all sources of error, various types of covariance inflation have been explored. They are different in terms of the implementation—multiplicative, additive, or adaptive—and whether or not they use a posteriori information, but all of them artificially increase ensemble spread, not based on an understanding of the characteristics of the error sources (Anderson and Anderson 1999; Anderson 2007, 2009; Hamill et al. 2001; Mitchell and Houtekamer 2000; Mitchell et al. 2002; Whitaker and Hamill 2002, 2012).

Fujita et al. (2007), Meng and Zhang (2007), and Meng and Zhang (2008) found that multiphysics ensembles consistently improved the performance in mesoscale EnKF experiments. Isaksen et al. (2007) investigated the effect of SKEB in their ensemble data assimilation (EnDA) system at European Centre for Medium-Range Weather Forecasts (ECMWF) and found that SKEB introduced realistic perturbations increasing ensemble spread. Houtekamer et al. (2009) demonstrated that additive inflation, the addition of isotropic perturbations to the analysis ensemble, showed a larger positive impact than multiphysics ensemble, SPPT, and SKEB. However, Hamill and Whitaker (2011) found that the additive noise constrained the spread growth and increased the ensemble mean analysis error substantially. Whitaker and Hamill (2012) reported that multiplicative inflation primarily accounted for sampling error associated with the observing network while additive inflation can better deal with errors in the model itself in their perfect model experiments. They also showed that SKEB did not improve upon additive inflation in their idealized case study.

In this study, we continue the model error representation in the mesoscale EnKF context, assimilating real rather than simulated observations. The Advanced Research version of the Weather Research and Forecasting (WRF) Model (Skamarock et al. 2008) coupled with the Data Assimilation Research Testbed (DART; Anderson et al. 2009) system (WRF-DART) is utilized to produce the EnKF analyses and forecasts in a cycling mode. Since a spatially and temporally varying adaptive inflation was introduced in Anderson (2007, 2009) to compensate for a small ensemble size (<100) and to effectively adjust to a heterogeneous observing system, it has been commonly used as a standard configuration for WRF-DART (Torn et al. 2006; Romine et al. 2013; Ha and Snyder 2014). Following the previous studies, we also use the adaptive inflation in all the experiments presented in this study. In addition to the inflation employed in the analysis step, we apply two model error techniques—SKEB and multiphysics ensemble (named “PHYS”)—in the forecast step to explicitly account for model uncertainty. Compared to the baseline experiment (that uses adaptive inflation only), short-term forecasts from the EnKF analyses with the two specific model error representation are evaluated both deterministically and probabilistically.

A couple of key questions we raise in the real data assimilation context are the following: (i) Can the explicit model error representation effectively increase ensemble spread and improve the short-range forecast skills in the mesoscale model? (ii) If the model error techniques can sufficiently increase the spread, would one still need covariance inflation? (iii) To what extent could the inflation and the model error schemes account for different sources of underdispersion in the mesoscale ensemble system? To address these issues, we carefully designed the mesoscale ensemble experiments, as described in section 2. Methodologies for the explicit model error techniques are presented in section 3, and results from different cycling experiments and the forecasts from the ensemble mean analyses for a longer forecast time are compared in section 4, followed by discussion on inflation versus model error representation in section 5, and a summary and discussion in section 6.

2. Experiment design

a. Model configuration

We perform the analysis/forecast cycling experiments over the contiguous United States (CONUS) domain using different model error techniques for the 1-month period of June 2008 at 3-h intervals. The ensemble adjustment Kalman filter, implemented in DART is employed in the analysis step, and the Advanced Research version of the WRF Model (version 3.2) is used in the forecast step for all experiments. Both analysis and forecast ensembles use 50 members, and each member is run on two model domains—123 × 99 at 45-km resolution and 162 × 105 at 15-km mesh—in two-way nesting, as depicted in Fig. 1. Both domains have 41 vertical levels up to 50 hPa and use the same physical parameterizations.
during the analysis/forecast cycle. The physics schemes used are described in section 3b.

b. Observations

During the 3-hourly cycling for June 2008 over the CONUS, we assimilate conventional observations such as radiosonde soundings (raob), Aircraft Communications and Reporting System (ACARS), marine, aviation routine weather report (METAR), and use mesonet surface data as an independent observation for verification, as shown in Fig. 1. All the observations are collected from the Meteorological Assimilation Data Ingest System (MADIS) of the National Oceanic and Atmospheric Administration (NOAA) and preprocessed for the assimilation system for data quality check (QC) or superobing, etc.

c. Adaptive inflation

In WRF-DART cycling experiments, we employ adaptive inflation to maintain prior ensemble spread throughout the period. We refer to the baseline experiment as control (CNTL) ensemble. As the inflation is adjusted to compensate for all sources of error at the analysis step during the cycles, model error is implicitly taken into account in CNTL.

Adaptive covariance inflation implemented in DART varies for various state variables, adapting to different densities of observation network and model error, which may vary in time and space. The inflation for unobserved variables is also updated based on the assumption that prior ensemble covariance between state variables is applicable to the covariance for prior inflation. During cycles, inflation is applied after the forecast step but before the forward operators are computed in the next analysis cycle, and thus the raw prior spread (e.g., the spread computed from ensemble forecast) is different from the inflated spread that is actually used in the analysis step.

The prior adaptive inflation value associated with each state vector is updated by Bayesian algorithms given a set of observations, the prior ensemble estimates of the observations, and observation error variances. It is noted that users can specify the lower bound of the inflation value as well as how quickly the inflation adapts in a densely observed region and a data-sparse region. These options enable users to better control the situation where the observing network or model error rapidly varies with time and the estimated inflation may not be able to keep up. However, it is not straightforward to optimize such options for particular applications. In this study, we follow our previous study that employed adaptive inflation in WRF-DART for the same retrospective case. Readers are referred to Anderson (2009) for details of the implementation, and Ha and Snyder (2014) for the tuning parameters of adaptive inflation used in this study.

As shown in Fig. 2, adaptive inflation seems to effectively increase the ensemble spread and thereby improves the quality of both a priori and a posteriori for a test period of the first 10 days in June 2008, when verified against independent mesonet observations.

The success of an ensemble Kalman filter data assimilation technique relies on its ability to maintain ensemble spread such that it can represent the error of the ensemble mean state. We thus check the spread–error relationship as in the previous studies [e.g., based on Eq. (1) in Ha and Snyder (2014)]. A reliable ensemble will make a good agreement between root-mean-square (rms) error of the ensemble mean forecast (e.g., prior mean) and total spread, the square root of the sum of prior ensemble spread and the observation error variance. As the observation error variance is specified as constant for most observations in this study, the reliability basically differs by ensemble spread in different experiments. We summarize the relationship for surface variables in Table 1, which indicates that the inflation increases total spread by up to 4%, leading to the prior mean better fitted to the corresponding observation (e.g., less rms innovations). The positive impact of inflation is observed in all surface variables, improving the quality of their 3-h ensemble mean forecasts by up to 7%. Note that ensemble spread shown in Fig. 2 (in black dashed line at higher points at each analysis time) and Table 1 is not a raw ensemble forecast spread but the one inflated for the analysis step. As indicated by
the spread–error ratio in Table 1, even after inflation, the predicted mean error is still underestimated in the sample ensemble forecast for most surface variables (e.g., totsprd/rmsi < 1). In other words, the ensemble system is still underdispersive and the EnKF analysis and the following forecasts can be further improved by taking model uncertainty into account. Based on the test, we employ explicit model error techniques on top of the inflation. As we want to examine the actual increase of ensemble spread by the model error representation, we compare the forecast ensemble spread before the inflation is applied unless it is noted differently for the rest of the paper.

3. Model error techniques

To explicitly represent model uncertainty, we employ SKEB and multiple suites of physical parameterization (PHYS) in the forecast ensemble. As we strive to investigate the benefit from the explicit model error representation in the standard (or most common) WRF-DART configuration, we implement those two schemes in addition to the adaptive inflation. Except for the model error representation, all the experiments use the same filter design and assimilate the same observations in the same model configuration as in Ha and Snyder (2014). Thus, only the configurations for the model error representation are described in detail in this section.

a. Stochastic kinetic energy backscatter ensemble

Stochastic parameterizations can efficiently increase ensemble spread by injecting stochastic forcing for each ensemble member. For this purpose, we employ a stochastic kinetic energy backscatter scheme that adds stochastic, small-amplitude perturbations to the rotational component of horizontal wind and potential temperature tendency equations at each time step. A random pattern is created in spectral space and each wavenumber separately evolves as a first-order autoregressive process:

$$
\psi_{k,l}(t + \Delta t) = (1 - \alpha)\psi_{k,l}(t) + g_{k,l}\sqrt{\alpha}\varepsilon_{k,l}(t).
$$

![Fig. 2. A time series of 2-m temperature in terms of root-mean-square (rms) innovations (thick solid lines) and ensemble spread (thin dashed lines) in control simulation with (“Inflation” in black) and without adaptive inflation (“No_inflation” in gray), verified against unassimilated mesonet observations over domain 1. Both posterior (lower points) and prior observations (higher points) were put together at each cycle making a sawtooth pattern for a 10-day test period. The rms innovations were averaged over the cycles to be annotated for both posterior (“po”) and prior (“pr”) in the legend.](image)

<table>
<thead>
<tr>
<th></th>
<th>U10 (m s⁻¹)</th>
<th>V10 (m s⁻¹)</th>
<th>T2 (K)</th>
<th>Td2 (K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infl</td>
<td>2.16</td>
<td>2.18</td>
<td>2.14</td>
<td>1.99</td>
</tr>
<tr>
<td>NoInfl</td>
<td>1.59</td>
<td>1.54</td>
<td>1.60</td>
<td>1.54</td>
</tr>
</tbody>
</table>

Table 1. RMS innovations (“rmsi”), total spread (“totsprd”), and their ratio (=totsprd/rmsi) of 3-h ensemble mean forecasts that are averaged over all common mesonet stations between two experiments with (“Infl”) and without adaptive inflation (“NoInfl”) in domain 1 for a test period of the first 10 days of June 2008.
Here $\psi_{k,l}(t)$ denotes the spectral coefficient of the pattern; $k$ and $l$ are the wavenumber components in the zonal $x$ and meridional $y$ direction in physical space, respectively; $t$ is time; $\Delta t$ is the model time step; $g_{k,l}$ denotes the wavenumber-dependent noise amplitude; and $\epsilon_{k,l}$ is a complex-valued Gaussian white noise process with mean $\langle \epsilon_{k,l}(t) \rangle = 0$ and covariance $\langle \epsilon_{k,l}(s)\epsilon^*_{m,n}(t) \rangle = \sigma^2 \delta_{k,m} \delta_{l,n} \delta_{t,s}$, where $*$ denotes the complex conjugate. Temporal correlations are prescribed by the autoregressive parameter $1 - \alpha$, while the perturbation wavenumber spectrum determines the spatial correlations. Subsequently, perturbations for the horizontal momentum and potential temperature are computed from the appropriate spatial derivatives in spectral space, then transformed back to gridpoint space to be added to momentum and potential temperature tendency equations, respectively. The injected energy is given as the difference in the total kinetic energy between the two consecutive time steps as expressed by the total streamfunction in Fourier space, $\psi_{k,l}(t)$.

For generality, we use the same parameter settings as in Berner et al. (2011), which were tuned to achieve a good spread–error relationship in short-range ensemble forecasts at 45-km grid resolution over the CONUS domain. The forcing spectra follow a $-5/3$ spectrum for kinetic energy and about a $-3$ slope for the potential temperature tendency perturbations. To make sure that the injected energy does not lead to any artifacts in the kinetic energy and potential energy spectra, the spectra of perturbed and unperturbed forecasts were compared (not shown). The spectrum of SKEB matches that of CNTL very well for wavenumbers above the effective resolution of the dynamical core ($\sim 6\Delta x$; Skamarock 2004). Below the effective resolution, the spectrum of SKEB shows a shallower slope than CNTL, better fitted to the Lindborg (1999) spectrum function [his Eq. (71)]. This is an indirect effect, however, since the wavenumbers below the effective resolution were not perturbed. In our experience, results (and the impact of the stochastic forcing) do not heavily depend on the details of the tuning parameters. At each time step, the standard deviation of stochastic perturbations to horizontal wind and temperature tendencies are up to 3% of the corresponding tendencies in the control simulation over the entire domain. In this study, the stochastic perturbations are the same at each model level and the temporal decorrelation time for each wavenumber is chosen as 30 min. A longer decorrelation time might be preferable, but the decorrelation time was not changed to ensure consistency with Berner et al. (2011).

In multidomains like the ones used in this study, perturbations are generated on the parent domain and then interpolated to the nested domain to be coherent across the domains.

This scheme was originally motivated by the notion that upscale and downscale cascading energy resulted in net forcing for the resolved flow from unresolved scales (Shutts 2005) and the perturbation amplitude was proportional to the instantaneous dissipation rate. Houtekamer et al. (2009) applied the stochastic backscatter scheme to their EnKF cycling system by only exciting high wavenumbers and found its impact insignificant. On the contrary, we perturb all wavenumbers available in the parent domain up to its effective resolution and use a spatially and temporally constant dissipation rate. Because we do not include flow-dependent perturbations (e.g., no more weight in regions of large dissipation) or scale-selective forcings in our stochastic scheme, we can interpret this as another way of generically representing model error—in the same way as in the standard Kalman filter equation that includes a model error term like the other additive noise approach introduced in previous studies (Mitchell and Houtekamer 2000; Houtekamer et al. 2005; Hamill and Whitaker 2005, 2011). Our method only differs in that the stochastic forcing (or model error) is applied every time step during the forecast, not just at the end of forecast or once per analysis cycle.

To illustrate the effect of SKEB, we run an ensemble forecast from the same analysis with and without stochastic perturbations for a single convective case. We take the 50-member ensemble analyses from the SKEB cycling experiment at 1800 UTC 8 June 2008 and run 12-h ensemble forecasts from the analyses with and without noise that is different across ensemble members. The ensemble forecast experiment without the stochastic forcing is named NO_SKEB. Note that it is different from CNTL in that it uses the analysis from SKEB and turns off the stochastic forcing only for this particular experiment. At the analysis time, a squall line was initiated along the strong surface cold front extended from western Iowa to the Oklahoma–Texas Panhandle, and developed into a mesoscale convective system (MCS) over the large contiguous area producing massive deep clouds and heavy rainfall associated with the surface front across the central United States 12 h later (Fig. 3). When we compare the horizontal distribution of temperature and horizontal wind speed at 700 hPa in 12-h forecast valid at 0600 UTC 9 June 2008, we find little changes in the ensemble mean analysis between the two experiments (not shown), but a considerable increase of ensemble spread over the entire domain, especially in the convective area and the upstream region due to additive noise (Fig. 4). Since NO_SKEB is also initiated from the SKEB analysis ensemble that
has been spun up with the stochastic forcing for 8 days, it still produces large spread over Kansas and Oklahoma. Note that the same ensemble forecast initialized from CNTL produces much smaller spread (<1) even in the convective area (not shown). All these results indicate that the stochastic forcing can be employed to represent significant model uncertainty associated with the mesoscale features in the severe convection area (such as strong horizontal temperature and wind gradient along the front).

b. Multiphysics ensemble

The WRF Model supports a variety of physics parameterization schemes and each scheme makes different assumptions and approximations. Thus, we can easily produce diversity in the ensemble trajectories by making different combinations of them. From the practical point of view, the biggest challenges in various physics combinations are the maintenance of each physics scheme and the robustness of each combination throughout the cycling period. To represent the forecast uncertainty due to imperfect physics parameterization schemes, we construct a multiphysics ensemble as summarized in Table 2, following Hacker et al. (2011). Each ensemble member uses 1 of 10 suites of physics schemes and each suite is employed 5 times in our 50-member ensemble. The CNTL and SKEB experiments use suite 5.

4. Results

To examine the effect of the explicit model error representation, we compare the relative performance of three different experiments during the cycles. Most results are similar between domain 1 (at 45-km resolution) and 2 (at 15-km resolution).

a. Deterministic forecast verification

Figure 5 shows a time series of rms innovations of the ensemble mean analysis with respect to mesonet observations (which are not assimilated). The observations used for verification in each experiment are subjected to the same quality check as those assimilated, and thus different experiments typically use different numbers of observations in both assimilation and verification. To make a fair comparison, we compute rms errors against observations common in all the experiments used in the comparison at each cycle. The ensemble mean analyses from the three experiments are generally of comparable quality for 10-m zonal wind, but as shown in the legend (inside the parentheses), the cycle-mean rms error for SKEB is slightly smaller than in PHYS, which in turn outperforms CNTL. Results for 10-m meridional wind are similar (not shown), and analyses of 2-m temperature also clearly benefit from the explicit model error representations compared to CNTL. Surface altimeter settings exhibit the largest improvement over CNTL, with 35% reductions of the analysis error (e.g., analysis-minus-observation differences) throughout the period.

Figure 6 illustrates rms innovations for the 3-h ensemble mean forecasts to check if the analysis improvement due to the model error representation results in a better performance in the following forecast. Based on the rms innovations averaged over the cycles, we find that SKEB is still the best in all the surface fields. One might expect that a better analysis leads to a better forecast for such a short lead time. However, while PHYS produces the best surface temperature analysis in Fig. 5b, forecast error quickly grows for 3 h to be slightly worse than SKEB, as shown in Fig. 6b. Bootstrapping over 10000 resamples of the original 240 cycles, the analysis and forecast errors of both SKEB and PHYS are statistically significantly different from those of CTRL at the 95% confidence level in all surface variables. The percentage improvement relative to CNTL is summarized in Fig. 7 in terms of rms innovations of 3-h ensemble mean forecasts for both domains 1 and 2 as follows:

\[
\text{ratio} = \frac{\text{rmsi}_{\text{CNTL}} - \text{rmsi}_X}{\text{rmsi}_{\text{CNTL}}} \times 100, \tag{2}
\]

where \(x\) is SKEB or PHYS. Here the rmsi is the rms innovations averaged over the whole month period in each experiment as shown in the legend in Fig. 6. The positive ratio indicates an improvement over CNTL and the negative means degradation. Overall, the forecast skill relative to CNTL is similar on domains
1 and 2 in most surface fields although the improvements relative to CNTL tend to be slightly larger on the finer mesh. Among the observation types, surface altimeter shows the biggest improvement from model error representation—PHYS improves the surface altimeter simulation over domain 1 by ~8%, while SKEB is better than CNTL more than 20% over domain 2. SKEB outperforms CNTL in all surface fields, improving 2-m temperature more than 10-m winds in 3-h forecast, but PHYS is slightly worse than CNTL in surface wind fields, especially in domain 1. As neither of the model error techniques is tuned to maximize the impact on particular fields, a larger benefit of the model error schemes in surface temperature or altimeter than in horizontal wind can be interpreted as a larger sensitivity (or response) where the model error is more pronounced.

To evaluate how much uncertainty is actually represented in ensemble forecasts, prior ensemble spread from each experiment is compared in Fig. 8. For this particular summer month, PHYS consistently produces the largest ensemble spread in all surface fields except for surface altimeter setting where SKEB produces almost double the spread of PHYS. We show all the time series plots only for common observations between the experiments, but when we check the total number of surface altimeter observations actually used for both assimilation (e.g., METAR) and verification (e.g., mesonet), we find that CNTL rejects a lot of observations because of insufficient ensemble spread. As a result, CNTL uses only ~60% of the observations that are used in the other two experiments, leading to significantly poor forecasts in surface altimeter.

<table>
<thead>
<tr>
<th>Physics suite</th>
<th>Surface</th>
<th>Microphysics</th>
<th>PBL</th>
<th>Cumulus</th>
<th>LW</th>
<th>SW</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Thermal</td>
<td>Kessler</td>
<td>YSU</td>
<td>KF</td>
<td>RRTM</td>
<td>Dudhia</td>
</tr>
<tr>
<td>2</td>
<td>Thermal</td>
<td>WSM6</td>
<td>MYJ</td>
<td>KF</td>
<td>RRTM</td>
<td>CAM</td>
</tr>
<tr>
<td>3</td>
<td>Noah</td>
<td>Kessler</td>
<td>MYJ</td>
<td>BM</td>
<td>CAM</td>
<td>Dudhia</td>
</tr>
<tr>
<td>4</td>
<td>Noah</td>
<td>Lin</td>
<td>MYJ</td>
<td>Grell</td>
<td>CAM</td>
<td>CAM</td>
</tr>
<tr>
<td>5</td>
<td>Noah</td>
<td>WSM5</td>
<td>YSU</td>
<td>KF</td>
<td>RRTM</td>
<td>Dudhia</td>
</tr>
<tr>
<td>6</td>
<td>Noah</td>
<td>WSM5</td>
<td>MYJ</td>
<td>Grell</td>
<td>RRTM</td>
<td>Dudhia</td>
</tr>
<tr>
<td>7</td>
<td>RUC</td>
<td>Lin</td>
<td>YSU</td>
<td>BM</td>
<td>CAM</td>
<td>Dudhia</td>
</tr>
<tr>
<td>8</td>
<td>RUC</td>
<td>Eta</td>
<td>MYJ</td>
<td>KF</td>
<td>RRTM</td>
<td>Dudhia</td>
</tr>
<tr>
<td>9</td>
<td>RUC</td>
<td>Eta</td>
<td>YSU</td>
<td>BM</td>
<td>RRTM</td>
<td>CAM</td>
</tr>
<tr>
<td>10</td>
<td>RUC</td>
<td>Thompson</td>
<td>MYJ</td>
<td>Grell</td>
<td>CAM</td>
<td>CAM</td>
</tr>
</tbody>
</table>
Next we consider the 3-h forecast spread after inflation to understand more clearly how the model error techniques influence the data assimilation. As the performance is consistent throughout the month-long period, we rerun them only for the first 10-day test period. Figure 9 indicates that the model error representation changes reliability more effectively in surface thermodynamic fields and altimeter settings than in surface wind when verified against independent observations. A key result is that PHYS is overdispersive in 2-m temperature and dewpoint. This is consistent with our analysis and forecast results shown in Figs. 5b and 6b, respectively, in that the ensemble mean analysis of PHYS fits observations very closely, perhaps overfitting, then its forecast rapidly loses the information contained in the analysis increments. The large spread in 2-m thermodynamic fields in PHYS partly comes from large uncertainties and/or systematic bias errors in different land surface models and different boundary layer schemes used here (not shown). Either direct improvements to those schemes in the forecast model or bias corrections in the analysis system would be certainly helpful to remedy such overdispersion, which we leave for a future study.

It should also be noted that the observation error specification commonly plays a crucial role in estimating total spread. We use the same observation errors as in Ha and Snyder (2014) where we adjusted surface observation errors and moisture errors based on the spread–error relationship for the same retrospective case. In this study, using the same observation errors across the experiments, PHYS is only slightly better than CNTL producing large spread, while SKEB consistently improves the surface forecast over CNTL and PHYS by increasing the spread only moderately.

To check the performance above the surface, we compute the rms innovations of 3-h ensemble forecast against radiosonde observations for the month-long period. Figure 10 compares the rms innovations of...
each experiment (solid lines) along with the prior spread (dashed lines) in domain 2 for four different variables. For sounding observations, we employ observation error variances taken from the Gridpoint Statistical Interpolation (GSI) analysis system (Kleist et al. 2009), which vary by pressure levels. For example, the observation error standard deviation of horizontal wind ranges from 1.4 to 3.2 m s\(^{-1}\), with a maximum at 250 hPa.

To focus on the differences between experiments (which use the same observation error), we plot ensemble spread instead of total spread in Fig. 10. At all levels, SKEB shows the largest spread and CNTL the least, except in dewpoint where all three experiments are comparable. Wind innovations in SKEB and PHYS are statistically significantly different from CNTL only at a couple of pressure levels at the 95% confidence level. Temperature forecast errors in SKEB and PHYS are also similar to each other, but clearly better than CNTL, while moisture was hard to improve with the specific model error techniques throughout the atmosphere. Ha and Snyder (2014) demonstrated that the quality of the moisture analysis is rather sensitive to the specification of observation error in their mesoscale application. The influence of SKEB on moisture is also indirect since the stochastic forcing is applied only to streamfunction and potential temperature.

b. Probabilistic forecast verification

To determine if the model error representation can result in better probabilistic performance, we compute the Brier score (BS) and the Brier skill score (BSS) of 3-h ensemble forecasts with respect to sounding data for the month of June 2008. For the common observations \( o \) between the experiments, we first compute the time mean \( \mu_o \) and the standard deviation \( \sigma_o \) of the observation at each location for the month-long period. Because the sounding data are available twice daily, every day, the total number of forecast samples or cycles, \( n_{\text{fcst}} \), is 60:

\[
\mu_o = \frac{1}{n_{\text{fcst}}} \sum_{i=1}^{n_{\text{fcst}}} o(i),
\]  

(3)

**Fig. 7.** The improvement (%) of forecast error in SKEB and PHYS over the one in CNTL for both domains (top) 1 and (bottom) 2 in various surface fields. The rms innovations are computed against mesonet observations and averaged over the month-long cycles. Positive means an improvement relative to CNTL in the 3-h ensemble mean forecast.

**Fig. 8.** As in Fig. 6, but for the 3-h ensemble prior spread.
Then we group the ensemble forecasts $f$ into four different observation events as bin 1 ($f_{>\mu_o + \sigma_o}$), bin 2 ($\mu_o < f < \mu_o + \sigma_o$), bin 3 ($\mu_o - \sigma_o < f < \mu_o$), and bin 4 ($f < \mu_o - \sigma_o$). For each event, we compute the BS by comparing the forecast probability from ensemble forecasts $P$ to the observed event $O$ at each forecast cycle as follows:

$$BS = \frac{1}{\text{nstn}} \sum_{n=1}^{\text{nstn}} [P(n) - O(n)]^2,$$

where nstn is the total number of stations available at each sounding level. For a statistical significance test, we bootstrap the BS (with the original 60 forecast samples) over 10,000 resamples as for all other comparisons illustrated in this paper. Because the relative performances between the experiments are similar for all four events, we only show the BS for the first event (e.g., for the strong zonal wind and the high temperature events) in Figs. 11a and 11b. We reverse the x axis to show a better performance to the right. At all levels, SKEB and PHYS significantly outperform CNTL, and SKEB

![Graph showing the ratio of total spread to rms errors of 3-h ensemble mean forecasts in each experiment for different surface variables over domain 2, verified against mesonet (unassimilated) observations.](image)

![Graph showing the vertical profile of 3-h ensemble mean forecast rms innovations (solid lines) against sounding observations and the prior ensemble spread (dashed lines).](image)
mostly offers a small improvement over PHYS. These differences largely come from improvements in the reliability component of the Brier score (not shown).

The BSS gives the skill of an ensemble forecast with respect to a reference:

$$\text{BSS}_x = \frac{\text{BS}_{x} - \text{BS}_{\text{ref}}}{\text{BS}_{\text{ref}}}$$

where $x$ is SKEB or PHYS. For a perfect forecast the BSS will be 1, while zero indicates no improvement over the reference forecast. Since we are interested in the performance of different ensemble experiments compared to that of the baseline experiment, we chose CNTL as a reference. As shown in Figs. 11c and 11d, both SKEB and PHYS show better probabilistic skills than CNTL at all levels thanks to larger ensemble spread. In the same sounding verification for the meridional wind and dewpoint, both model error approaches produce better probabilistic skills than the control run except for dewpoint forecast in SKEB at 925 hPa, which is not statistically significant (not shown).

c. Extended forecast verification

Now we examine how long the positive impact of model error techniques can last. For that, we take the ensemble mean analysis from 0000 and 1200 UTC cycles every other day for the same month and run a deterministic forecast for 72 h to be verified against observations and gridded analyses.

1) VERIFICATION AGAINST OBSERVATIONS

With respect to surface METAR observations, forecast rms errors of each experiment are computed over the whole CONUS domain (e.g., domain 1) and shown for the first 24 h in Fig. 12. In CNTL, surface wind error grows quickly from 1.56 to 2.2 m s$^{-1}$ over the first 12 h, then slowly increases up to 2.4 m s$^{-1}$ by 72 h, while the surface temperature error starts from 2.2 K and gradually increases to $\sim$3 K at the 72-h forecast. At the
analysis time, SKEB starts with the error of 1.4 m s$^{-1}$ in 10-m zonal wind (improving over CNTL by $\sim$9%) and PHYS has the rms innovation of 1.7 K in 2-m temperature (about 22% less than CNTL). The errors quickly grow over the next 9 h in all three experiments to make their forecasts similar to each other, but the improved analysis in both SKEB and PHYS leads to slightly better forecast up to 24 h. These results are consistent with the results shown in Figs. 5 and 6 although they show the verification against mesonet over domain 2, while Fig. 12 shows the verification against METAR, the assimilated observations, which makes the sharper reduction of the impact. Even if the forecast rms differences between the experiments are not huge, they are statistically significant at the 95% confidence level at all lead times.

We also examine the quality of precipitation forecasts in the experiments using the fractional skill score (FSS; Roberts and Lean 2008; Schwartz et al. 2009). Figure 13

![Figure 12](image12.png)

**Fig. 12.** Root-mean-square (rms) innovations verified against METAR observations over domain 1 in (a) 10-m $u$ wind and (b) 2-m temperature. Results are computed from 30 extended forecasts from the ensemble mean analyses for the month of June 2008, twice daily, every other day.

![Figure 13](image13.png)

**Fig. 13.** FSS as a function of time for precipitation thresholds of (left) 1.0 and (right) 10.0 mm (3 h)$^{-1}$ computed over domain 2 for a 105-km radius of influence, aggregated over all forecasts from 0000 UTC initialization time, every other day for the month of June 2008.
shows the FSS for the 3-h accumulated rainfall against NCEP stage-IV data over domain 2 for two different thresholds—1.0 and 10.0 mm—within the radius of influence of 105 km. Overall, the scores are not very sensitive to the radius of influence or to the initialization time of the ensemble forecast, and we show the FSS for precipitation forecasts from the 0000 UTC initialization here. In the light rainfall area (>1 mm), three experiments are comparable over the first 12 h, after which the control experiment becomes less skillful than the other two. For heavy rainfall cases (>10 mm), intensive storms simulated in PHYS are often either mislocated or underestimated to match the observed precipitation for the first 12 h. But, at later forecast times, CNTL tends to develop and intensify convection too quickly, leading to an overestimation of the heavy rain cells along the leading edge of convective systems, while SKEB and PHYS simulate the growth of mesoscale convection moderately (not shown). As a result, both model error schemes produce higher skill scores than CNTL between 12 and 24 h.

2) Verification against the RUC analysis

We also evaluate the extended forecasts on domain 2 against the Rapid Update Cycle (RUC) analysis, which has 13-km horizontal grid resolution and 50 isentropic-sigma hybrid vertical levels, and utilizes substantially more observations than the analyses computed here. Figures 14 and 15 illustrate the verification in terms of zonal wind and temperature, respectively, as a time series at 500 hPa and a vertical profile for the 12-h forecasts. In terms of the zonal wind, SKEB slightly outperforms CNTL and PHYS in the entire atmosphere for the first 24 h (with larger improvements above 700 hPa) while PHYS is worse than CNTL up to 42 h in the middle atmosphere. For temperature, both model error techniques consistently improve the forecast up to 60 h, especially in the midtroposphere. Comparing Fig. 10a with Fig. 14b or Fig. 10c with Fig. 15b indicates that the short-term forecast skills verified against the RUC analyses are consistent with the ones against sounding observations.

5. Inflation versus model error techniques

In the previous sections, we demonstrated that model error schemes effectively increased ensemble spread and thereby accepted more observations to improve short-range forecasts. However, they were applied to the prior states in addition to the inflation that adapted in time and space (and between state variables). Thus, it was not manifest to what extent inflation and the explicit model error representations were redundant and to what extent they accounted for different resources for underdispersion. In this section, we examine those issues briefly. They are challenging problems to untangle in the real data assimilation context, and we admit that not all aspects of those issues are covered here.

We first investigate the systematic behavior of inflation and 3-h ensemble forecast spread by averaging them over the entire model domain at each level for the whole month period (Fig. 16). Here we examine the vertical structure up to level 35 that corresponds to ~100 hPa to be consistent with all other verifications shown in the vertical profile. Comparing inflation versus spread, and between experiments, we find several distinct features:
Both model error schemes—SKEB and PHYS—increase the spread throughout the atmosphere. In general, SKEB produces the largest spread in all state variables (not shown). The larger the ensemble spread gets, the smaller the inflation is needed except where there is a significant systematic bias (e.g., for PHYS near the surface). The vertical structure of inflation for CNTL (black solid line) shows high inflation values near the jet level (around model level 27) and low-level atmosphere (below level 10). As our inflation builds upon the observed information, this can be partly due to a large volume of ACARS, the observation that is known to have a large warm bias around 250 mb (Ballish and Kumar 2008), and a dense surface observing network as shown in Fig. 1. Most importantly, both model error techniques significantly reduce the large inflation shown in CNTL, especially at the levels stated in (iii). In particular, SKEB decreases inflation that is about 2.2 K at level 5 in CNTL to 1.3 K at the same level, and mitigates the large vertical variability in CNTL. The domain-averaged inflation in SKEB only varies between 1.1 and 1.3 for the whole column, while the one in CNTL ranges from 1.2 to more than 2. These results indicate that the adaptive inflation is working as expected, with its magnitude adjusting as other sources of ensemble variance are added. They also suggest that a significant portion of inflation in CNTL compensates for model error that can also be reasonably represented by explicit model error schemes in the forecast step.

In principle, adaptive inflation can also compensate for the systematic underestimation of analysis variance in the EnKF update owing to sampling error. To further understand this potential benefit of adaptive inflation, we conduct an additional experiment, SKEB_noinfl, which includes the stochastic forcing but turns off the adaptive inflation. (We also conducted the same no-inflation experiment for PHYS and obtained similar results, thus we only present and discuss the SKEB without inflation here, for simplicity.) Compared to SKEB_noinfl, SKEB produces larger prior ensemble spread with the maximum increase up to 10% in surface wind. However, the larger spread leads to improving the 3-h surface forecast only by ~1% in both domains (not shown). In other words, SKEB_noinfl is nearly as good as SKEB, and inflation helps only a little.
Next we compare SKEB_noinfl with CNTL. The time series of the rms innovations relative to mesonet temperature illustrates that both the stochastic forcing and the adaptive inflation produce short-range forecasts whose skill is consistent throughout the month-long period and exhibit no sign of filter divergence (Fig. 17a). On average, SKEB_noinfl improves the 3-h ensemble mean forecast over CNTL by 10% in surface temperature. It also makes improvements of 8% and 20% for surface wind and altimeter, respectively (not shown). We also examine the 3-h forecast verification against sounding observations for the two experiments in Fig. 17b. The stochastic perturbations increase ensemble spread more than 40% (red dotted line) and reduce time-mean model-minus-observation bias (red dashed line) and rms errors (red solid line) throughout the entire atmosphere. Thus, while the combination of stochastic backscatter and adaptive inflation performs best, backscatter in isolation (SKEB_noinfl) improves significantly over adaptive inflation in isolation (CNTL). This indicates that in our mesoscale forecast system, the effects of sampling error in causing underdispersion of the ensemble are secondary to those of the neglected model error.

One might ask if the model error representation is beneficial simply because it increases ensemble spread more effectively, leading to assimilating more observations than CNTL (with inflation only). Note that each experiment assimilates different sets of observations because of the quality check based on innovations (e.g., $o - f$) and total spread at each analysis step. To clarify this issue, we perform another control experiment (named "CNTL_SKEBobs") assimilating precisely those observations that were used in SKEB, without any further quality control (based on its own spread). Interestingly, assimilating exactly the same observations as in SKEB, the new control run produces very similar 3-h forecast skills at first, but after the first-day cycles, ensemble mean forecasts are distinctively degraded compared to the ones in SKEB, as shown in Fig. 18. We examine other fields in both domains and find that the new control run with the same observations as in SKEB is consistently worse than SKEB throughout the test period of the first 10 days. This implies that SKEB improves the forecast not only because it assimilates more observations with larger spread but because it can properly simulate the model error that is not fully captured by the covariance inflation.

6. Summary and discussion

Model error is one of the main sources of uncertainty in the mesoscale analysis and the subsequent forecasts. When it is not properly taken into consideration, the ensemble mean analysis tends to deviate from the true state and the ensemble spread fails to reflect the error of the mean state. Through WRF-DART cycling experiments for one month of June 2008, we explore two explicit model error approaches in the mesoscale data assimilation context. A baseline experiment ("CNTL") only uses adaptive inflation to account for the error from all different sources, including (but not limited to) sampling error and the misspecification of observation error. With adaptive inflation applied to prior ensemble states, however, the short-range ensemble forecasting system still suffers from the lack of spread, particularly near the surface. In an effort to achieve more realistic spread, we enhance the representation of model error by using multiphysics ensemble (called "PHYS") and stochastic kinetic energy backscatter ensemble (SKEB). Note that inflation is applied in the analysis step based on the observed information (e.g., a posteriori) while the model error is represented in the forecast step based on the prior states.

When we check the effect of stochastic forcing in a particular convective case, we find that SKEB effectively increases ensemble spread in the entire model domain both in the horizontal and the vertical. Although our stochastic forcing is not scale selective or flow dependent, the maximum dispersion is found in the convective area indicating that the stochastic perturbations can still produce a realistic spread capturing the large model uncertainty associated with severe mesoscale convective systems.

During the cycles, both model error schemes improve the surface analysis with the largest benefit in surface altimeter (~35%) in SKEB, implying that the total air mass integrated through the depth of the entire atmosphere is better represented by accounting for the model uncertainty. In the subsequent forecast, the positive impact remains in most surface fields, again with the biggest improvement in SKEB for surface altimeter (~20%) over the 15-km domain. Interestingly, PHYS is the best in the surface temperature analysis but its error quickly grows for the 3-h forecast to be slightly worse than SKEB. While SKEB shows consistent benefits throughout the period in most variables and metrics, PHYS overestimates dispersion near the surface and overfits 2-m temperature observations, leading to the subsequent surface temperature forecasts that are worse than those in CNTL.

In the vertical profile, SKEB produces the largest ensemble spread for most variables that in turn contributes to the improved reliability of probabilistic forecasts and produces significantly better deterministic forecasts up to 72 h at most levels. However, neither of
FIG. 17. A comparison between CNTL and SKEB_noinfl for 3-h ensemble mean forecast in (a) a time series of rms innovations against mesonet 2-m temperature observations and (b) a vertical profile against sounding temperature observations over domain 2 for the whole month period.
the explicit model error techniques improves moisture. Note that SKEB adds perturbations only in the streamfunction and potential temperature. Although no particular physics parameterization is known to be consistently better than others in mesoscale applications (Wang and Seaman 1997), ensemble members with different physics schemes in PHYS will never be equally likely, particularly for such an active summer month period examined in this study. Distinct systematic bias errors are found in particular physics combinations depending on the variables and levels (not shown), and PHYS can be further optimized for better performances in the short-term forecast, although it is not trivial to find optimal physics combinations for 50–100 members maintaining the diversity and the reliability of the ensemble system.

Despite our potentially suboptimal design of model error techniques and the limitation of a small sample size for a one-month period of summer, both SKEB and PHYS produce promising results in the analysis and short-range forecasts for the entire atmosphere both deterministically and probabilistically. Verified against observations, it is shown that the ensemble mean analysis error is reduced by accounting for model uncertainty up to ~9% and 22% in the surface wind and temperature, respectively. The forecast errors quickly grow for the first 3 h, but the model error representation remains beneficial at least up to 24 h near the surface, leading to a slight improvement of precipitation forecasts. In the verification against the RUC analysis, we confirm that taking model uncertainty into account facilitates the improvement of short-term prediction at most levels. Bootstrapping over 10 000 resamples, most of the verification results showed a statistical significance at the 95% confidence level.

The role of adaptive inflation is also investigated in the context of model error representation. Unlike model error techniques, inflation knows about the observed information (as it is computed based on observation-minus-forecast) and thus accounts for the impact of the inhomogeneous observing network on underdispersion (due to sampling error). During the cycles, it adapts not only to the variations in the observation network but also to model error. When model uncertainty is explicitly taken into account, inflation is reduced by up to 70% in the boundary layer, implying that inflation predominantly responds to model uncertainty [rather than to the observing network or to the \((\sigma - f)\) bias]. When we turn off inflation leaving the stochastic forcing only (in “SKEB_noinfl”) and compare it to CNTL that employs inflation only, we find that the stochastic forcing consistently produces larger spread and better forecasts than the ones in CNTL, significantly reducing the model bias in the boundary layer. An increased ensemble spread affects the quality control, which in turn decides how many observations are assimilated in the ensemble data assimilation system. Therefore, we conduct another control experiment that assimilates exactly the same observations as in SKEB turning off its own quality control, and find that the new control run (with the adaptive inflation only) still gets worse than SKEB with cycles. These additional experiments demonstrate that, although the adaptive inflation mostly accounts for model error, it does not capture model error as efficiently as the explicit model error techniques such as stochastic perturbations. The benefit of model error techniques does not simply come from the increased spread (thereby assimilating more observations) but from the proper representation of the model error that is commonly underestimated even with covariance inflation in our mesoscale ensemble system. However, even if the sampling error is not as critical as the model error, we obtain the best performance when the stochastic forcing is used with adaptive inflation. Given that our model error methods are applied only at the forecast step (where no observation is provided) and are still limited by the finite ensemble size, inflation would be still valuable in assimilating real observations from the heterogeneous observing network in the mesoscale ensemble system.

Although model error techniques used in this study represent model uncertainty in a physically more realistic fashion and produce encouraging results in the mesoscale analysis and the short-range prediction, more research is still needed. Even with the existing stochastic approaches, many of the parameterizations in the numerical models are still deterministic and many of the current approaches have arbitrary or case-dependent factors in the implementation. In many cases, they can greatly increase ensemble spread but cannot improve the forecast skills at the same time. More research and
rigorous tests on different applications are required to produce consistent and robust performances through the stochastic approach. A direct improvement of physics parameterizations would be certainly helpful to further reduce model uncertainty or systematic model errors. Superparameterization that explicitly represents the subgrid-scale dynamics and physics in a high-resolution cloud-resolving model embedded in each model grid column (Grabowski 2001, 2006; Khairoutdinov et al. 2005) could be another approach to improve the representation of subgrid-scale physical processes. From the analysis point of view, a careful treatment of correlated observation error and biased observations is worth more investigation in a sense that observations can be more effectively assimilated in the ensemble data assimilation system. In the mesoscale analysis/prediction system, the large uncertainties in lateral and surface boundary conditions would also play a critical role although they were not considered in the current study.

Acknowledgments. We thank Glen Romine and Jeff Anderson for their helpful discussion on this work and Craig Schwartz for the precipitation verification. We also appreciate Roberto Buizza and two anonymous reviewers for their insightful comments that greatly improved our draft.

REFERENCES


