Key Points:
- Additive noise based on climatological background error covariance partially accounts for model error in convective-scale data assimilation
- Under strong forcing conditions, additive noise performs better than relaxation methods
- Under weak forcing conditions, performance of additive noise relative to combinations degrades a bit but results are still comparable

Correspondence to:
Y. Zeng,
yuefei.zeng@lmu.de

Representation of Model Error in Convective-Scale Data Assimilation: Additive Noise, Relaxation Methods, and Combinations

Yuefei Zeng[1,2] , Tijana Janjić[2] , Alberto de Lozar[3], Ulrich Blahak[3], Hendrik Reich[3], Christian Keil[1], and Axel Seifert[3]

[1]Meteorologisches Institut, Ludwig-Maximilians-Universität (LMU), Munich, Germany, [2]Hans-Ertel-Centre for Weather Research, Deutscher Wetterdienst, Offenbach, Germany, [3]Deutscher Wetterdienst, Offenbach, Germany

Abstract For ensemble data assimilation, background error covariance should account for sampling and model errors. There are a number of approaches that have been developed that try to consider these errors; among them, additive noise and relaxation methods (relaxation to prior perturbation and relaxation to prior spread) are often used. In this work, we compare additive noise, based on random samples from global climatological atmospheric background error covariance, to relaxation methods as well as combinations. Our experiments have been conducted in framework of convective-scale data assimilation with conventional and radar reflectivity observations hourly assimilated for a 2-week convective period over Germany. In the first week under weather conditions characterized by strong large-scale forcing of convection, additive noise performs equally or even better than relaxation methods and combinations during both assimilation and short-range forecasts. In addition, it is shown that the relaxation to prior perturbation may be associated with smoothing of background errors that negatively affect small-scale structures and that the relaxation to prior spread yields more unbalanced model states. For the second week in absence of strong forcing, the performance of additive noise relative to combinations has been degraded a bit but results are still comparable. Overall, additive noise provides a good benchmark for further developments in representation of model error for convective-scale data assimilation.

1. Introduction

Nowadays, convection-permitting numerical weather prediction (NWP) models are used operationally at most national weather forecasting centers (Baldauf et al., 2011; Seity et al., 2011; Skamarock et al., 2005; Tang et al., 2013). As the ensemble Kalman filter (EnKF; Evensen, 1994) has advantages such as easy implementation, a flow-dependent background error covariance, and automatic generation of an optimal ensemble to initialize ensemble forecasts, it has been increasingly used in operational NWP systems in the last decade (Houtekamer et al., 2005). The Local Ensemble Transform Kalman Filter (LETKF; Hunt et al., 2007) has been run operationally in the kilometer-scale ensemble data assimilation (KENDA) system of the Deutscher Wetterdienst (DWD) since March 2017 (Schraff et al., 2016) for the convection permitting COnsortium for Small-scale MOdeling (COSMO; Baldauf et al., 2011) model. As a probabilistic approach, the EnKF should be especially useful for the convective-scale data assimilation because of the stochastic nature of convection. In particular, there have been already a number of studies on performance of the EnKF applied to convective-scale radar data assimilation (Aksoy et al., 2009; Bick et al., 2016; Caya et al., 2005; Dowell & Wicker, 2009; Dowell et al., 2004, 2011; Gao & Xue, 2008; Lange & Craig, 2014; Lange et al., 2017; Snyder & Zhang, 2003; Sobash & Stensrud, 2013; Tong & Xue, 2005; Xue et al., 2006) and promising results have been gained. However, many meteorological centers still rely on the latent heat nudging (LHN) to indirectly assimilate radar reflectivities for the operational usage (Gustafsson et al., 2018).

For the ensemble data assimilation, background error covariances are usually affected by sampling error due to the finite size of ensemble and model error. The model error is defined as a difference between the simulated and the "true" atmospheric states and can be caused by several sources, among others the limited resolution of numerical models and parameterizations of subgrid-scale physical processes. The background error covariance should account for both sampling and model errors. Omitting either error source may result in underdispersive background error covariance of the EnKF and consequently filter collapse in practical data

assimilation system (Dee, 1995). A number of techniques have been proposed to address this problem, for example, (adaptive) multiplicative inflation (Anderson & Anderson, 1999; Houtekamer et al., 2005), relaxation to prior perturbations (RTPP; Zhang et al., 2004) and relaxation to prior spread (RTPS; Whitaker & Hamill, 2012). Bowler et al. (2017) showed with the global 4DEnVar system of the Met Office that the RTPP produces inappropriately balanced and large-scale perturbations and the RTPS requires a relaxation factor larger than 1.0 to achieve sufficient spread. Therefore, Bowler et al. (2017) combined these two relaxation methods to alleviate downsides of each. Kotsuki et al. (2017) developed an adaptive RTPP and compared it with the adaptive RTPS (Ying & Zhang, 2015) and adaptive multiplicative inflation, using the nonhydrostatic icosahedral atmospheric model and the LETKF to assimilate conventional observations and satellite radiances. Kotsuki et al. (2017) found that despite of overdispersion (underdispersion) in sparsely (densely) observed area, both adaptive relaxation methods perform better than the multiplicative inflation. In the convective-scale context with radar data assimilated, the RTPP has been used in Poterjoy et al. (2014) and Zhang et al. (2004), and the RTPS in Johnson et al. (2015) and Bick et al. (2016) to maintain sufficient ensemble spread. Poterjoy et al. (2014) examined how sampling error affects the evolution of storm structure within assimilation cycles and suggested that a properly tuned coefficient of RTPP could improve the structure of the inner-core vortex although it has little impact on model variables far from observations.

As alternatives, several additive noise approaches have been proposed and they were proven to be especially useful in mitigating model error in data assimilation for global NWP models. For instance, Houtekamer et al. (2005) sampled from a scaled 3DVAR background error covariance when using additive noise with EnKF for data assimilation with global NWP model of Enviroment Canada. Hamill and Whitaker (2005) used a primitive equation global model and compared the performance of three different parameterizations of additive noise within assimilation cycles, including scaled samples from a time series of differences between model forecasts at different resolutions, scaled samples of 24-hr forecast tendencies and of model state's anomaly from the model climatology, and they found that the first one produced most accurate analyses. Furthermore, the additive noise based on resolution differences was compared by Whitaker and Hamill (2012) with the RTPS in the same idealized setup. It was shown that the RTPS is favorable for treating the sampling error in densely observed regions whereas the additive noise is especially useful to represent model error. Combination of RTPS with additive noise is usually superior to either method used alone; however, additive noise alone could outperform any combination in case of dominant model error. Bowler et al. (2017) developed another type of additive noise that uses random samples from a long archive of analysis increments collected from 2-month 4DVAR data assimilation experiments and combined it with the relaxation methods and found that the combinations produce a much larger spread and a small reduction in the root-mean-square error (RMSE) of ensemble mean in 72-hr ensemble forecasts. In convective-scale context, additive noise has been tested as well, for instance, Dowell and Wicker (2009) added smoothed Gaussian noise where storms were observed and shown that it provides sufficient spread within the convective storms with the environment relatively undisturbed and it results in less spurious cells than multiplicative inflation. This method is commonly used at University of Oklahoma and NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma, for radar data assimilation.

Due to very short-scale and chaotic property of convection, it is extremely difficult to parameterize associated model error. For instance, most of current convection-permitting models are run at horizontal resolutions 1–5 km, but some important motions within and between convective clouds occur at much smaller scales (Bryan et al., 2003), so parameterization schemes are required to deal with unresolved features of convection. In addition, a complex microphysical scheme is vital, although it is difficult to find a microphysical scheme that is superior in all weather regimes. Considering that there are meanwhile the other sources of uncertainties for convective-permitting models, for example, arising from lateral boundary conditions, surface modeling, radiation scheme, and orographic representation (Clark et al., 2016), model error may be a dominant error source that background error covariance should take into account for convective-scale data assimilation. Based on results of Whitaker and Hamill (2012), in case model error is large, additive noise might outperform relaxation methods.

In a study on predictability of precipitation, Vié et al. (2011) perturbed convective-scale initial conditions and lateral boundary conditions in the AROME model to explore forecast uncertainties of Mediterranean heavy precipitation events. They found that the type of large-scale forcing determines the relative role of both perturbations. If the large-scale forcing is strong, the impact of uncertainties in the lateral boundary conditions is predominant; If the large-scale forcing is weak, the impact of the convective-scale perturbations introduced

at initial time is elevated. In the context of convective-scale data assimilation, uncertainties from the lateral boundary conditions are also an important error source as discussed in Gustafsson et al. (2018). Recently, an additive noise based on random samples of the climatological atmospheric background error covariances from the global EnVar data assimilation system for ICON (ICOsahedral Nonhydrostatic; Zängl et al., 2015) has been newly implemented in the operational KENDA system. Note that the ICON model provides lateral boundary data for the regional COSMO-DE model; therefore, this fashion of additive noise may feed the regional model the information on error arising from the global forcing model and thus may be especially useful in weather situations with strong large-scale forcing. From 27 May to 9 June 2016, Germany was influenced by a large-scale ridge and hit by extraordinarily many severe convective storms, while the prevailing forcing of convection was strong in the first week and weak in the second week. This provides us ideal conditions to study the effectiveness of additive noise in different large-scale forcing patterns. To investigate this, we use the KENDA system to execute a series of data assimilation experiments for the first and second weeks separately, in which radar reflectivity observations are assimilated in addition to conventional data. We compare additive noise with relaxation methods as well as combinations, and we investigate if the relative roles of additive noise and relaxation methods depend on governing weather regimes. We will validate the performance of experiments both during data assimilation and in quality of short-range forecasts, particularly precipitation forecasts.

The paper is organized as follows: In section 2, we give a short introduction to the operational COSMO-DE model and data assimilation system KENDA as well as observations assimilated. In section 3, we present methods implemented in the KENDA system to take into account sampling and model errors. Section 4 describes the weather situation of the study period and experimental setup. Section 5 presents the results of experiments, followed by section 6 for the conclusion and the outlook.

2. The COSMO Model, Operational KENDA System, and Observations Assimilated

The COSMO model is a fully compressible and nonhydrostatic regional NWP model, which has been developed and maintained by the multinational COSMO consortium (Baldauf et al., 2011; Doms & Baldauf, 2015; Doms et al., 2011). The COSMO-DE is a convective-scale configuration operationally run at DWD. The COSMO-DE model domain is horizontally discretized in 421×461 grid points with resolution of 2.8 km and in vertical 50 terrain-following hybrid layers are deployed. Lateral boundary conditions for the ensemble are provided by the Ensemble Prediction System of the operational global model ICON with a resolution of 40 km globally and 13 km over Europe, while the ICON for deterministic system has a resolution of 20 km globally and 6.5 km for Europe. To solve dynamical equations, the time-splitting Runge-Kutta scheme of Wicker and Skamarock (2002) is used. Due to high horizontal resolution, the deep convection can be simulated explicitly, whereas the shallow convection is represented by parametrization based on the Tiedtke (1989) scheme. A convection-allowing Lin-Farley-Orville-type one-moment bulk microphysical scheme is employed, including cloud droplets q_c, cloud ice q_i, rain q_r, snow q_s, and graupel q_g (Lin et al., 1983; Reinhardt & Seifert, 2006). Turbulence parameterization is based on the turbulent kinetic energy (TKE) equation of Raschendorfer (2001). The heating rate caused by radiation is calculated by the scheme of Ritter and Geleyn (1992). The boundary conditions at the ground are provided by the multilayer soil model TERRA (Doms et al., 2011).

In the operational KENDA system, the assimilation window is 1 hr. The LETKF (Hunt et al., 2007) is used to assimilate conventional observations which include upper-air observations such as radiosondes (TEMP), wind profilers (PROF) and aircraft reports (AIREP), and synoptic surface observations (SYNOP). TEMP data measure temperature, horizontal wind and relative humidity, and radiosondes are launched at fixed times, mostly at 00:00 and 12:00 UTC with some also at 06:00 and 18:00 UTC. PROF measurements of horizontal wind are available approximately half-hourly during the whole day. The majority of AIREP horizontal wind and temperature data are collected during the daytime. For SYNOP, surface pressure data are assimilated hourly; 10-m horizontal wind data are only assimilated in northern Germany and mostly rejected in southern Germany because of the high topography; 2-m temperature and humidity data are not assimilated. These conventional observations are assimilated without horizontal or vertical thinning. Localization is done in the observation space, accounting for observations only in the vicinity of the analysis grid point and assigning different weights to observations by the fifth-order Gaspari-Cohn (1999) function. The length of horizontal localization radius is determined adaptively so that the total number N_{loc} of localization-weighted observations for a local analysis is approximately equal to the ensemble size N_e. The reason for this data-density-dependent localization scale is as follows: For a local analysis, the number of degrees of freedom of assimilation is roughly equal to N_e.

Adding more data to an already sufficiently large data set would have little potential to improve the analysis itself but would make the LETKF overconfident by further decreasing the estimated analysis error. For perfect observations, N_{loc} can be chosen equal to N_e, but for imperfect observations, N_{loc} should be chosen somehow larger than N_e. In this study, N_{loc} is set to 100 for $N_e = 40$. To avoid the analysis increments being projected onto too small scales for too small localization radius or being contaminated too much by sampling noise of the background error estimate for too large localization radius, lower and upper bounds r_{loc}^{min} and r_{loc}^{max} should be set to the length of horizontal localization radius. For conventional data used operationally average localization length scale produced by this scheme in horizontal on well-observed levels is less than 100 km as shown in Lange and Janjić (2016). Therefore, in this study, $r_{loc}^{min} = 50$ km and $r_{loc}^{max} = 100$ km. The vertical localization scale varies with height from 0.0075 to 0.5 in logarithm of pressure, but it is not adaptive. More details about localization can be found in Perianez et al. (2014), Schraff et al. (2016), and Lange and Janjić (2016). The observation errors for conventional observations are tuned based on Desroziers, Berre, Chapnik, and Poli (2005), Desroziers, Brousseau, and Chapnik (2005) statistics. The estimated values are given in Figure 3 and are used to specify the observation error variances.

In addition, the DWD operates a network of 17 C-band Doppler radars over Germany and part of neighboring countries (see Figure 1 of Zeng et al., 2016), which can provide reflectivity and radial wind measurements in interval of 5 min. So far, the LHN technique has been applied to each ensemble member to assimilate radar-derived precipitation rates (Schraff et al., 2016; Stephan et al., 2008). In this work, radar reflectivity observations are assimilated by the LETKF, using the efficient radar forward operator EMVORADO (Zeng et al., 2014, 2016). Radial wind observations are not assimilated and serve as an independent data source for verification. Due to high density of radar observations, data thinning technique is applied both temporally and spatially. Temporally, only the latest 5-min radar observations at the analysis time are assimilated. It is found that this temporal thinning performs better than full observations based on all skill scores used in this article (not shown), likely due to removal of temporal observation correlations of radar data; Spatially, superobbing is applied to each station and elevation individually. Superobbing is implemented in a quasi-Cartesian fashion: First, a 2-D horizontal Cartesian grid with a specific resolution d_o is constructed. Second, for each elevation, the radar bin centers (range, azimuth) are projected onto the horizontal Cartesian plane in order to find the closest radar bin for each Cartesian grid point, and this radar bin is regarded as the center of superobbing. Third, radar bins around the center bin are averaged, where the size of the averaging area ("pie-wegde"-shaped area) is given by range and azimuth intervals. The width of the range interval is defined by $d_o \sqrt{2}$ and the number of azimuths at the range r_0 is given by $2 \arctan \left[(d_o \sqrt{2}/2)/r_0 \right]$. Thus, the number of radar bins that contribute to a superobservation decreases with range. Second and third steps are repeated for all elevations. This superobbing technique results in a relatively homogeneous horizontal data distribution. Since the analysis grid is coarsened by a factor of three (see below), d_o is set to 10 km to ensure that model and observations have similar horizontal resolutions. More details and illustrative example for superobbing of radar observations can be found in Bick et al. (2016). A constant horizontal localization with radius of 16 km is applied to radar reflectivity observations. Note that the localization of radar reflectivity observations is done separately from the adaptive localization of conventional observations, but they are assimilated simultaneously. Moreover, all reflectivities ≤5 dBZ are set to 5 dBZ and are assimilated as no-reflectivity information (Aksoy et al., 2009). The observation error for reflectivity is set to 10 dBZ as in Bick et al. (2016).

The operational KENDA system computes the analysis weights on a coarsened grid and interpolates coarse analysis weights onto the high-resolution grid. The analysis increments are then calculated from the full-resolution background ensemble. By doing this, computational costs can be significantly reduced while maintaining a high accuracy of the analysis as shown by Yang et al. (2009). In the operational setup, a coarsening factor of 3 is applied (Bick et al., 2016), that is, analysis weights are computed on around 11% of full model grid points. Furthermore, the prognostic variables of TKE, q_c, q_i, q_r, q_s, and q_g are not updated at the analysis step. Currently, an ensemble of 40 members is employed plus one deterministic run. To deal with issue of imbalanced model state, techniques like saturation adjustment and hydrostatic balancing (Rhodin et al., 2013) are applied after each analysis step.

It should be noted that since 2017 Mode-S data have also been assimilated operationally (Lange & Janjić, 2016) but they are not used in the current study. More details about the KENDA and implementation of the LETKF can be found in Lange and Janjić (2016) and Schraff et al. (2016).

3. Methods to Account for Sampling and Model Errors

In the context of data assimilation, methods such as multiplicative/additive noise and relaxation methods as mentioned in section 1 are often used to represent sampling and model errors. Just a few studies (Hamrud et al., 2015; Houtekamer et al., 2009; Whitaker & Hamill, 2012) exploit stochastic perturbations methods that include multiphysics schemes (Murphy et al., 2004), stochastic backscatter (Shutts, 2005), and stochastic physics (Buizza et al., 1999) since they are not suitable for very short assimilation windows. In this work, we will focus on additive noise and relaxation methods. In this section, we give a short introduction to those methods.

3.1. Additive Noise

Additive noise (or additive covariance inflation) parameterizes model error by adding random perturbations $\boldsymbol{\eta}^{(i)} \in \mathbb{R}^{N_x}$ to the analysis state $\boldsymbol{x}^{a(i)} \in \mathbb{R}^{N_x}$ of ensemble member i:

$$\boldsymbol{x}^{a(i)} \leftarrow \boldsymbol{x}^{a(i)} + \alpha_a \boldsymbol{\eta}^{(i)}, \tag{1}$$

where N_x is the size of state vector, α_a is a tunable scalar and $\boldsymbol{\eta}^{(i)}$ should sample the model error covariance $\boldsymbol{Q} \in \mathbb{R}^{N_x \times N_x}$, that is, $< \boldsymbol{\eta}^{(i)}(\boldsymbol{\eta}^{(i)})^T > = \boldsymbol{Q}$. To ensure that additive noise increases the analysis ensemble covariance without changing the ensemble mean, the mean $\bar{\boldsymbol{\eta}} = \frac{1}{N_e} \sum_{i=1}^{N_e} \boldsymbol{\eta}^{(i)}$ should be removed from each $\boldsymbol{\eta}^{(i)}$ (Whitaker et al., 2008). Since $\boldsymbol{Q}$ is constructed by N_e samples by (1), its rank is at most N_e (see Sommer & Janjić, 2018, for alternative formulation that increases the rank). Therefore, additive noise changes ensemble members but it cannot contribute much to the increase of rank of background covariance matrix and to reduction of sampling error. It is mainly used to deal with model error.

Since March 2017, additive noise based on climatological atmospheric background error covariance matrix, hereafter denoted by $\boldsymbol{B} \in \mathbb{R}^{N'_x \times N'_x}$, from the EnVar data assimilation system for the global ICON model (horizontal resolution 20 km), has been utilized in the operational KENDA system. Here N'_x is the size of state vector from the global model. This is similar to what is done by Houtekamer et al. (2005) who took random draws from background error covariance of 3DVAR and used them as samples for model error for the global EnKF data assimilation, except here it is done in convective scale context. Although additive noise is applied to the whole domain, it will be artificially "damped" closed to lateral boundaries and only have effect on the initial conditions. The model states in the vicinity of the boundaries are "nudged" toward the boundary conditions provided by the ICON ensemble members.

To construct the $\boldsymbol{B}$-matrix, the so-called NMC (National Meteorological Center; Parrish & Derber, 1992) method is used:

$$\boldsymbol{B} \approx \frac{1}{2} < (\boldsymbol{x}^{t_1} - \boldsymbol{x}^{t_2})(\boldsymbol{x}^{t_1} - \boldsymbol{x}^{t_2})^T >, \tag{2}$$

where angled brackets mean an average over a training time, $\boldsymbol{x}^{t_1}$ and $\boldsymbol{x}^{t_2} \in \mathbb{R}^{N'_x}$ are model forecasts of duration t_1 and t_2, respectively, valid at the same time. Therefore, $\boldsymbol{\delta} := \boldsymbol{x}^{t_1} - \boldsymbol{x}^{t_2} \in \mathbb{R}^{N'_x}$ is a sample for forecast error. The error covariance matrix of $\boldsymbol{\delta}$ is twice $\boldsymbol{B}$, so the factor $\frac{1}{2}$ in equation (2) (Bannister, 2008). The $\boldsymbol{B}$-matrix is estimated over 1-year samples with $t_1 = 48$ and $t_2 = 24$ from the global ICON model (20-km resolution). Separability is assumed between the correlations in the vertical and in the horizontal direction and only the vertical part of the NMC derived covariances is used. For a given date, statistics (covariances in the vertical and variances in horizontal) are used from three adjacent months (for instance, May/June/July for a date in June). Variances as well as covariances in the vertical derived from NMC are smoothed using a band-pass filter in the zonal direction, which has the same effect of duplicating each of the samples onto each grid point in zonal direction and then averaging over all samples. In the horizontal, correlations were prescribed as follows: SOAR or TOAR approximated by compactly supported functions (Gneiting, 1999) with a length scale of 400 km for mass, stream-function and velocity potential (wind is derived from that by horizontal differentiation), and Gaussian with a length scale of 150 km for relative humidity. The horizontal correlations are empirically tuned for the regional model COSMO at hand. Temperature and wind covariances are derived under hydrostatic and geostrophic balance conditions. No additional diurnal variability is incorporated in construction of the $\boldsymbol{B}$-matrix. The square root $\boldsymbol{B}^{\frac{1}{2}} \in \mathbb{R}^{N'_x \times N'_x}$ is adapted to the finer grid mesh of the regional COSMO-DE model and stored as $\tilde{\boldsymbol{B}}^{\frac{1}{2}} \in \mathbb{R}^{N_x \times N_x}$ in an economic wavelet format. More implementation details can be found in Rhodin et al. (2013). At each assimilation step and for each ensemble member, a random vector $\boldsymbol{\gamma}$ from normal distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ is drawn. Sample $\boldsymbol{\eta}^{(i)} = \tilde{\boldsymbol{B}}^{\frac{1}{2}} \boldsymbol{\gamma}$ is then added to the ith ensemble member as in (1). Since additive noise samples the background error covariance of the global driving model, it is supposed
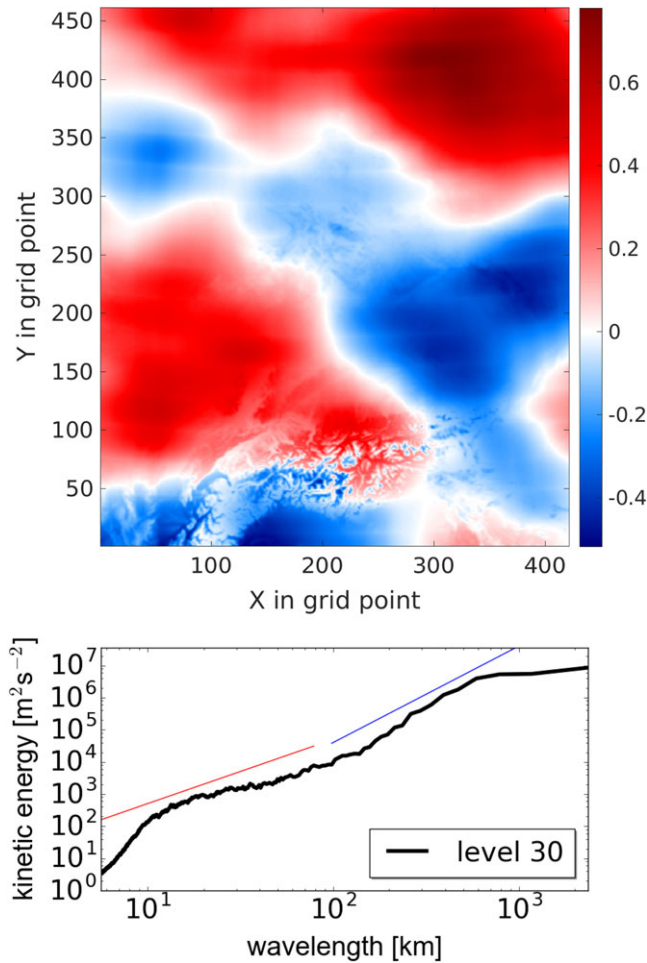
Figure 1. Perturbations of u (upper) and kinetic energy spectrum of perturbations (lower) at model level 30, averaged over 40 random samples. The spectrum is calculated over a square domain by removing 20 grid points in latitude on each side of the COSMO-DE domain. One-dimensional spectrum results from 2-D Fourier transformation into linearly detrended field and subsequent summation of the Fourier coefficients over annuli in wave number space. The blue and red lines show the reference lines for −3 and −5/3 power laws, respectively.

to be greatly representative of large-scale uncertainties in lateral boundary conditions for the regional data assimilation. Hereafter, this method is denoted by "LAN" (Large-scale Additive Noise).

In the operational setup, model variables of horizontal wind u and v, relative humidity q_v, temperature T, and pressure p are perturbed. The other microphysical variables such as q_c, q_i, q_r, q_s, and q_g are not perturbed. For instance, Figure 1 illustrates perturbations of u field and kinetic energy spectrum of perturbations at model level 30 (∼3 km), averaged over 40 random samples. From the field plot it can be seen that perturbations are generally large scale except for some small-scale structures in the Alpine region, which stem from the variations of the model level in pressure coordinates. The spectrum plot shows that the slope at wavelengths between 10 and 100 km is close or even shallower than −5/3, which indicates perturbations are energetic at those scales (Sun & Zhang, 2016); however, perturbations are not energetic at scales smaller than 10 km. α_a has been well tuned for convective-scale data assimilation and operationally set to 0.1. α_a is small because the original NMC-statistics is designed for background error of 3-hr forecasts of global data assimilation system but here it should represent model error of 1-hr forecasts.

3.2. Relaxation Method

One of the relaxation methods is RTPP (Zhang et al., 2004), by which the posterior analysis ensemble perturbations $\boldsymbol{X}^a \in \mathbb{R}^{N_x \times N_e}$ are relaxed toward the background ensemble perturbations $\boldsymbol{X}^b \in \mathbb{R}^{N_x \times N_e}$ independently at each analysis grid point by

$$\boldsymbol{X}^a \leftarrow (1 - \alpha_p)\boldsymbol{X}^a + \alpha_p \boldsymbol{X}^b, \qquad (3)$$

where $\alpha_p \in [0, 1]$ is a tunable parameter. Since α_p is a uniform parameter for each model variable and grid point, the balance among variable will not be strongly disturbed. The RTPP is implemented in the KENDA system and operationally used with α_p tuned to 0.75.

Whitaker and Hamill (2012) introduced another relaxation method called RTPS, which relaxes the analysis ensemble standard deviation σ^a toward the background ensemble standard deviation σ^b by

$$\sigma^a \leftarrow (1 - \alpha_s)\sigma^a + \alpha_s \sigma^b \qquad (4)$$

and this formula can be rewritten as

$$\boldsymbol{X}^a \leftarrow \left(\alpha_s \frac{\sigma^b - \sigma^a}{\sigma^a} + 1 \right) \boldsymbol{X}^a. \qquad (5)$$

It is clear that the RTPS is a purely multiplicative inflation. However, in contrast to the RTPP, the ensemble spread is calculated for each model variable and grid point, which can lead to unbalanced states as mentioned in Whitaker and Hamill (2012).

The RTPS is currently not used in the operational setup, but it has been successfully tested by Bick et al. (2016) and Harnisch and Keil (2015) with $\alpha_s = 0.95$.

4. Experimental Design

4.1. Weather Situation

During a 2-week period from 27 May to 9 June 2016, a great part of Europe was under control of atmospheric blocking, which was characterized by a large-scale ridge (upper-level high-pressure system) spreading over the North Atlantic and the northern Europe that hindered the air mass exchange. The high-pressure block was flanked by two upper-level troughs: On the west side, one headed southward to the Azores; on the east side, the other one over eastern Europe reached out southward to the Black Sea and Turkey. In the first half of the period, the southwestern part of Germany was influenced by the low pressure with moist and warm air that
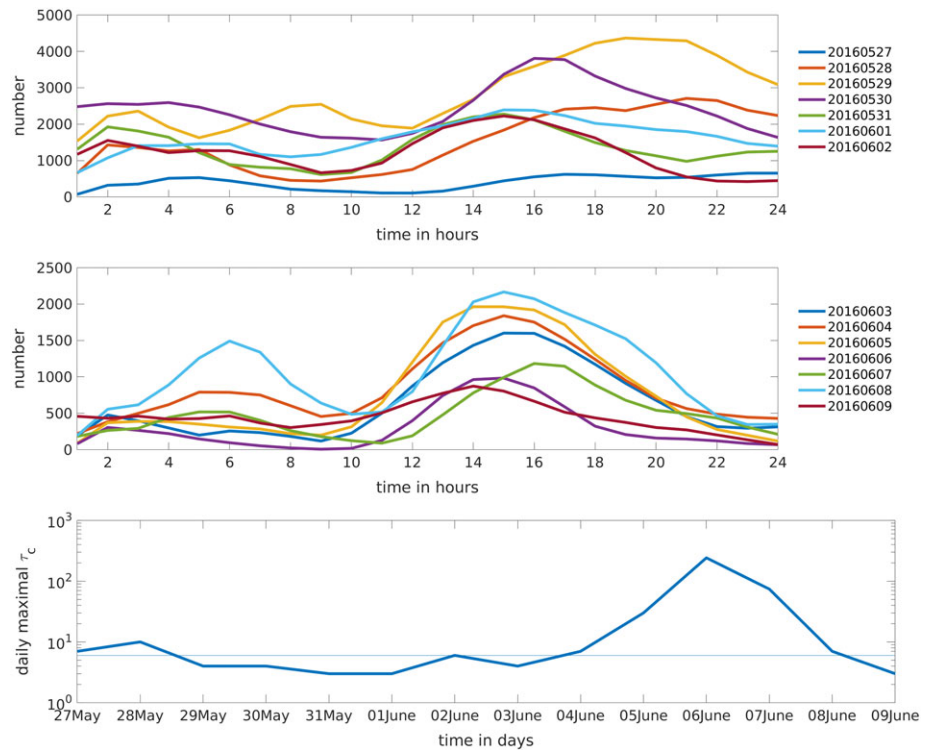
Figure 2. Hourly variations of numbers of model grid points with precipitation rate ≥5.0 mm/hr for different days in the first half (upper panel) and the second half (middle panel) of the investigation period. The lower panel shows the variations of daily maximal τ_c within the investigation period. The *y*-axis is in log scale, and the horizontal line indicates the threshold value of 6 hr. All data are derived from the mean of individual ensemble forecasts of COSMO-DE.

was advected ahead of a deep trough northeastward toward central Europe; in the second half of the period, moisture was maintained mainly by evapotranspiration from local sources and advection from surrounding countries. Piper et al. (2016) found that the wind speeds in the lower troposphere were very low due to the weak pressure gradient, particularly in the second week, which led to nearly stationary thunderstorms and caused large precipitation accumulations in local areas. Overall, Germany was overwhelmed by extraordinarily many severe convective cells which occurred almost every day during the investigation period. The majority of convective cells in systems exhibited a typical diurnal cycle with peaks in the late afternoon or early evening (see Figure 2).

In this work, the convective adjustment time scale τ_c (Keil et al., 2014), which is defined as the ratio of convective available potential energy over the rate of change of the convective available potential energy and describes how fast the conditional instability is discharged by the release of moist convection, is used to distinguish the days with strong large-scale forcing from those with weak forcing. In case of small τ_c, convection is in equilibrium with the large-scale environment and substantially influenced by the large-scale forcing; in case of large τ_c, convection is only triggered by some local mechanism due to lack of large-scale forcing. Here the spatial averaged τ_c is calculated for each hour. If τ_c exceeds the threshold value 6 hr once a day, the day is considered as weak forcing; otherwise, it is considered as strong forcing (Keil et al., 2014; Kühnlein et al., 2014). τ_c is derived from the mean of individual ensemble forecasts of COSMO-DE and illustrated in Figure 2. It can be clearly seen that the first week was predominantly strong forcing and the second week was predominantly weak forcing. This is in good agreement with Piper et al. (2016) that found particularly low atmospheric stability in the second week.

4.2. Experimental Setup

The experiments are conducted in a quasi-operational setting using a basic cycling environment (BACY) developed by the DWD. The experimental configurations are fairly close to the operational ones except: (1) Instead of the LHN, radar reflectivity observations are assimilated by the LETKF; (2) adaptive multiplicative inflation is switched off; (3) TKE and hydrological model variables (q_c, q_i, q_r, q_s, and q_g) are also updated by the LETKF; and (4) Mode-S EHS data are not assimilated.

Table 1
Experimental Configurations of Three Studies

Experiment	Additive noise $\alpha_a = 0.10$	RTPP $\alpha_p = 0.25$	$\alpha_p = 0.75$	RTPS $\alpha_s = 0.45$	$\alpha_s = 0.95$
E_LAN0.10	✓	×	×	×	×
E_RP0.75	×	×	✓	×	×
E_RS0.95	×	×	×	×	✓
E_LAN0.10RP0.25	✓	✓	×	×	×
E_LAN0.10RP0.75	✓	×	✓	×	×
E_LAN0.10RS0.45	✓	×	×	✓	×
E_LAN0.10RS0.95	✓	×	×	×	✓

Note. The symbol "✓" means "on," and the symbol "×" means "off." Study 1 includes E_RP0.75, E_LAN0.10, E_LAN0.10RP0.25, and E_LAN0.10RP0.75; Study 2 includes E_RS0.95, E_LAN0.10, E_LAN0.10RS0.45, and E_LAN0.10RS0.95; Study 3 includes E_LAN0.10, E_LAN0.10RP0.25, and E_LAN0.10RS0.45. RTPP = relaxation to prior perturbation; RTPS = relaxation to prior spread.

We focus on the convective-scale forecasting with the forecast lead time up to 6 hr. Ensemble forecasts are run daily only at 10:00, 11:00, …, 17:00, and 18:00 UTC since precipitation in the study period exhibited a clear diurnal cycle. The predictability of convective events is limited due to the rapid error growth and nonlinearity at small scales (Bei & Zhang, 2007; Hohenegger & Schär, 2007; Selz & Craig, 2015; Sun & Zhang, 2016; Zhang et al., 2003, 2006, 2007). In particular, Sun and Zhang (2016) showed that convective-scale components of the error may saturate in around 6 hours and the predictability of those components may be bounded to this time scale. In addition short range forecasting is challenging for data assimilation as well, since in this short range often nowcasting outperforms NWP.

In this work, three studies are performed. They differ in the investigation period: from 27 May to 2 June under strong forcing conditions for the first and second studies, and from 3 to 9 June under weak forcing conditions for the third study. In each study, a series of experiments are executed as follows:

- Study 1: We compare the LAN ($\alpha_a = 0.10$) to the RTPP as well as combinations with different α_p: E_RP0.75, E_LAN0.10, E_LAN0.10RP0.25, and E_LAN0.10RP0.75.
- Study 2: We compare the LAN ($\alpha_a = 0.10$) to the RTPS as well as combinations with different α_s: E_RS0.95, E_LAN0.10, E_LAN0.10RS0.45, and E_LAN0.10RS0.95.
- Study 3: Under weak-forcing conditions, we compare the LAN ($\alpha_a = 0.10$) to combinations: E_LAN0.10RP0.25 and E_LAN0.10RS0.45.

Table 2
Numbers of Conventional Observations and Radar Reflectivities Assimilated in E_LAN0.10

Obs. type	Variable	Temporal res.	Number/hour
TEMP	Temperature	6 hr	0–122
	Hor. wind		0–121
	Rel. humidity		0–110
PROF	Hor. wind	30 min	501
AIREP	Temperature	1 min	0–494
	Hor. wind		0–478
SYNOP	10 hor. wind	1 hr	242
	Surface pressure		601
RADAR	Reflectivity	5 min	29,937

Note. The approximate numbers of observations collected within 1 hr are given in the third column. These may slightly change for different experiments. The numbers are averaged over the period from 27 May to 2 June 2016.
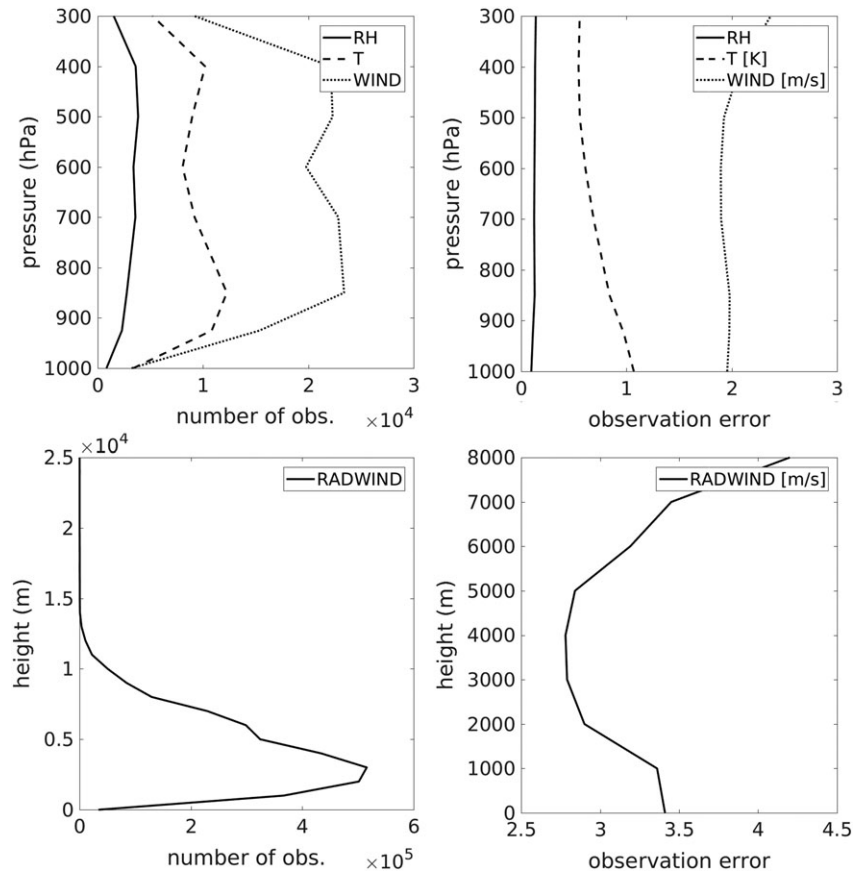
Figure 3. (left) The number of observations used for verification in vertical for relative humidity (RH), temperature (T), and horizontal wind (WIND) from TEMP, PROF, and AIREP (upper) and radial wind (RADWIND) from radars (lower); (right) vertical profile of observation error estimated by Desroziers statistics averaged over all observation types. Vertical profile of radial wind is only up to 8,000 m because there are too few observations in higher levels.

The different configurations of each experiment are given in Table 1. Table 2 summarizes the average numbers of observations assimilated for each type and variable including the number of radar reflectivity data after superobbing and temporal thinning.

5. Results

5.1. Verification Scores

For verification of performance during assimilation, innovation statistics are used, including the RMSE, σ_{f-o} of the ensemble mean and the spread skill ratio r_s (Aksoy et al., 2009), defined as

$$r_s = \frac{\sqrt{\sigma_f^2 + \sigma_o^2}}{\sigma_{f-o}},$$ (6)

where σ_f is the ensemble spread in observation space and σ_o is the observation error estimated based on Desroziers statistics (see Figure 3). The value $r_s = 1.0$ is optimal and $r_s < 1.0$ means that the ensemble spread is insufficient and vice versa.

For verification of performance during free forecasts, several other metrics are used in addition to the RMSE, for instance, the false alarm rate (FAR) and equitable threat score (ETS; Wilks, 2006). The FAR is the number of false alarms divided by the total number of events forecast. With the 2×2 contingency table (see Table 3), the FAR is given by

$$FAR = \frac{\text{False alarms}}{\text{Hits} + \text{False alarms}}.$$ (7)

Table 3
A 2 × 2 Contingency Table Gives the Discrete Joint Distribution of Forecasts and Observations in Terms of Cell Counts

2 × 2 Contingency table		Event observed	
		Yes	No
Event forecast	Yes	Hits	False alarms
	No	Misses	Correct negatives

Note. For dichotomous categorical forecasts, only two outcomes (yes or no) are possible.

It ranges from 0 to 1, and the perfect score is 0. The FAR is often used to detect spurious convection (e.g., Sun et al., 2016).

The ETS is defined as

$$\text{ETS} = \frac{\text{Hits} - \text{Hits}_{\text{random}}}{\text{Hits} + \text{Misses} + \text{False alarms} - \text{Hits}_{\text{random}}}, \tag{8}$$

where $\text{Hits}_{\text{random}} = \frac{(\text{Hits}+\text{Misses})(\text{Hits}+\text{False alarms})}{\text{Total}}$. "Total" is the total number of verification grid points. The ETS measures the fraction of observed and/or forecast events that are correctly predicted, and it is adjusted for hits associated with random chance (e.g., it is easier to correctly forecast precipitation occurrence in a wet climate than in a dry climate). It ranges from $-\frac{1}{3}$ to 1, the perfect score is 1, and 0 indicates no skill. The ETS is often used in the verification of precipitation because its "equitability" allows scores to be compared more fairly across different regimes.

Besides, rather than point comparisons between forecasts and observations, the fraction skill score (FSS; ; Robert & Lean, 2008) is used to make spatial comparisons. The FSS can be calculated as follows:

$$\text{FSS} = 1 - \frac{\text{MSE}}{\text{MSE}_{\text{ref}}}, \tag{9}$$

The MSE (mean square error) is defined as

$$\text{MSE} = \frac{1}{N_i N_j} \sum_{i=1}^{N_i} \sum_{j=1}^{N_j} \left[O(i,j) - M(i,j) \right]^2, \tag{10}$$

where $N_i \times N_j$ is the domain size, and the MSE_{ref} is defined as "largest possible MSE"

$$\text{MSE}_{\text{ref}} = \frac{1}{N_i N_j} \left[\sum_{i=1}^{N_i} \sum_{j=1}^{N_j} O^2(i,j) + \sum_{i=1}^{N_i} \sum_{j=1}^{N_j} M^2(i,j) \right]. \tag{11}$$

$M(i,j)$ is given by the fraction of grid points in model forecast that lie within a neighbourhood of (i,j) defined by a prespecified scale and exceed the threshold value. Accordingly, $O(i,j)$ is computed for observations. Note that the model and observational data are projected onto an identical verification grid, which is the naive model grid in this work. The FSS values vary between 0 and 1, and higher FSS values imply better performance.

To take the uncertainty in above-mentioned verification scores into account, the bootstrap method (Efron & Tibshirani, 1993) is used. E_LAN0.10 is taken as the reference run, the relative differences compared to E_LAN0.10 are calculated, and then 10,000 bootstrap resampling is carried out to examine the statistical significance at 95% confidence intervals. In this work, bootstrapping is applied to the ensemble, which means, for example, false alarms and hits (MSE and MSE_ref) are collected over all ensemble members when calculating the FAR (FSS).
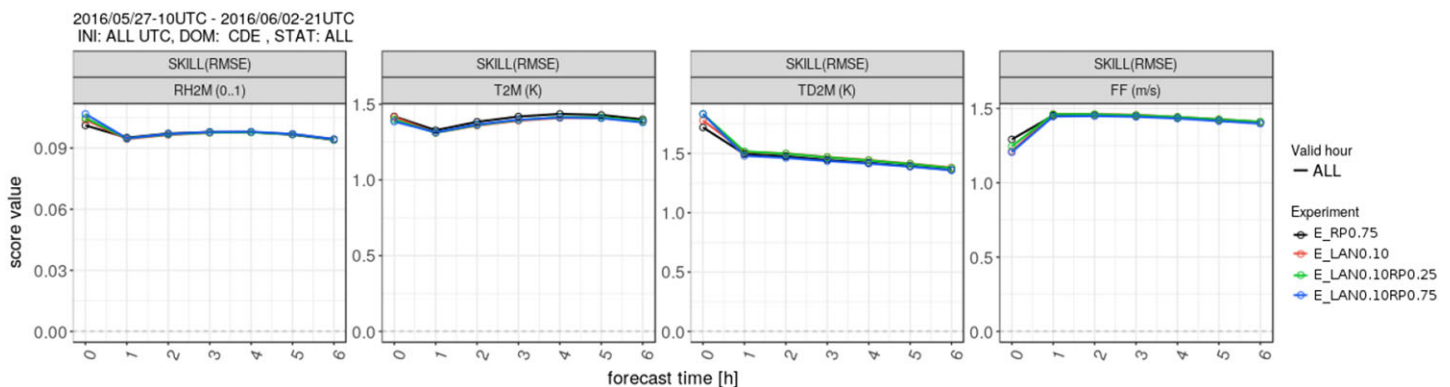
5.2. Study 1

First, we investigate the performance during assimilation cycling by means of innovation statistics (i.e., 1-hr forecasts) including the RMSE σ_{f-o} and the spread skill ratio r_s. The verification is based on upper air observations collected from TEMP, PROF, and AIREP, as well as radial wind observations that are not assimilated. Based on r_s shown in Figure 4, all experiments overestimate the spread up to 600 hPa and underestimate the spread above 600 hPa for relative humidity. Moreover, up to 850 hPa, E_RP0.75 has the best r_s; between 850 and 600
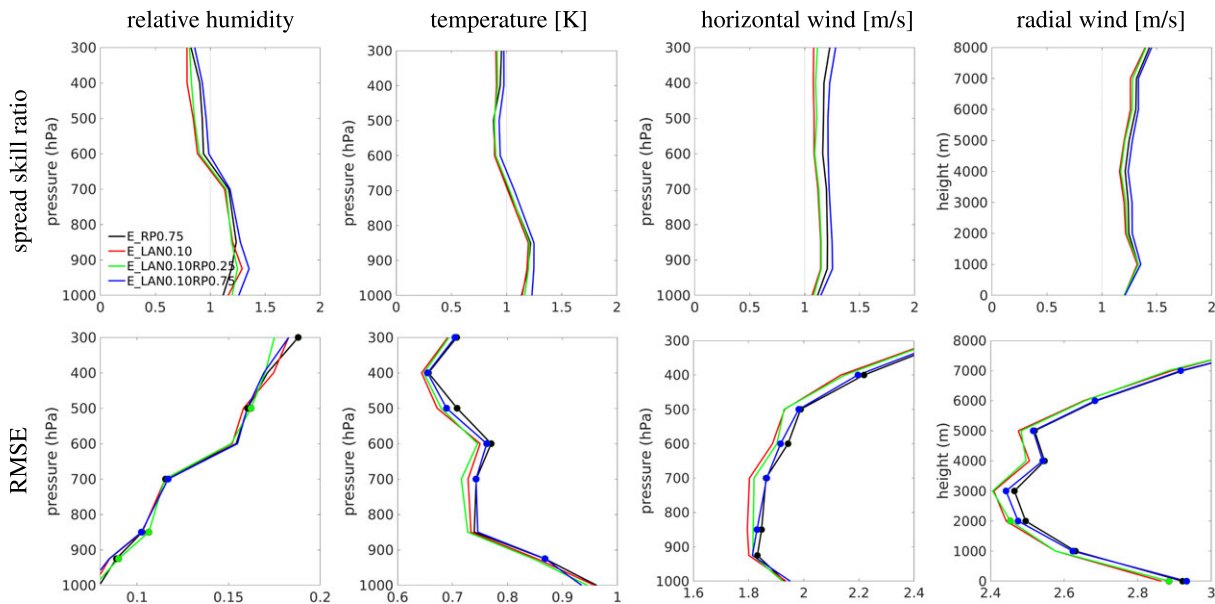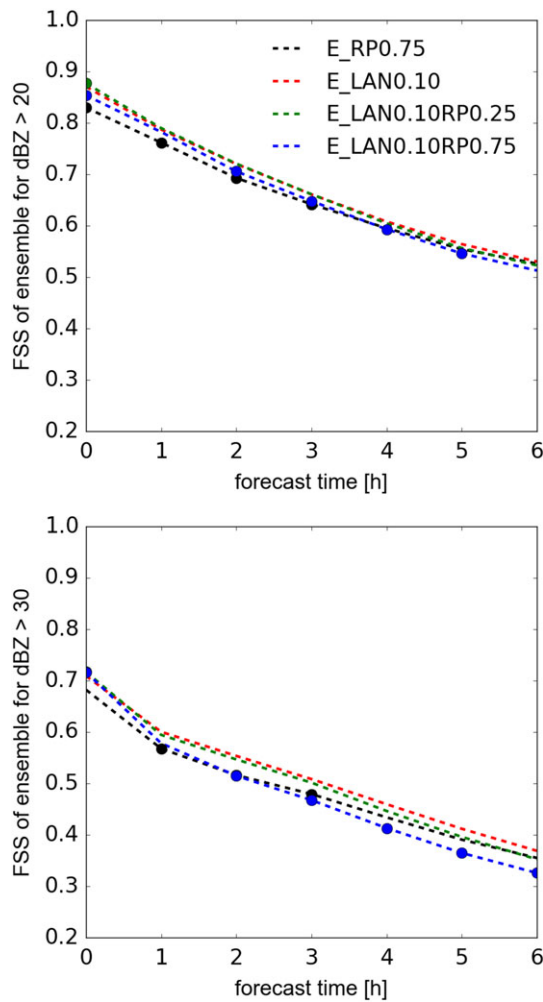
Figure 4. Comparison of E_RP0.75, E_LAN0.10, E_LAN0.10RP0.25, and E_LAN0.10RP0.75 of Study 1 by the vertical profile of spread skill ratio and RMSE of first guess ensemble, verified against observations of relative humidity, temperature, wind speed from TEMP, PROF, and AIREP and radial wind from radars. Results are averaged over all assimilation cycles. The lines of RMSE are marked as filled dots at the levels where the errors are significantly different from the ones in E_LAN0.10 at 95% confidence intervals after 10,000 bootstrap resampling. RMSE = root-mean-square error.

hPa, E_LAN0.10 has the best r_s, followed sequentially by E_LAN0.10RP0.25, E_RP0.75, and E_LAN0.10RP0.75, and it is opposite above 600 hPa. For temperature, similar fashion of overestimated and underestimated spread can be seen, although the spread is much less underestimated above 600 hPa compared to relative humidity. For both horizontal and radial wind, all experiments result in excessive spread for the whole vertical profile, although E_LAN0.10 is associated with the least overestimation, followed by E_LAN0.10RP0.25, E_RP0.75, and E_LAN0.10RP0.75. The vertical profile of radial wind is limited to 8,000 m because there are very few observations available in upper levels as shown in Figure 3. With respect to the RMSE, it increases with height for relative humidity. E_LAN0.10 is slightly better than the others based on statistically significant levels. For temperature, the RMSE decreases with the height. E_LAN0.10RP0.25 is comparable to E_LAN0.10; E_LAN0.10RP0.75 is slightly better than E_RP0.75, and they both are statistically significantly worse than E_LAN0.10 at almost all levels. Similar differences can be seen for horizontal and radial wind, although E_LAN0.10 is slightly better than E_LAN0.10RP0.25 with statistical significance at lower levels for radial wind.

Figure 5. Comparison of E_RP0.75, E_LAN0.10, E_LAN0.10RP0.25, and E_LAN0.10RP0.75 of Study 1 by the RMSE of 6-hr ensemble forecasts (including initial hours) for 2-m relative humidity, 2-m temperature, 2-m dew point temperature, and horizontal wind (from left to right) verified against the surface SYNOP observations. The RMSE is averaged over all assimilation cycles. Filled dots mean that differences are statistically significant; empty dots mean no statistical significance. RMSE = root-mean-square error.
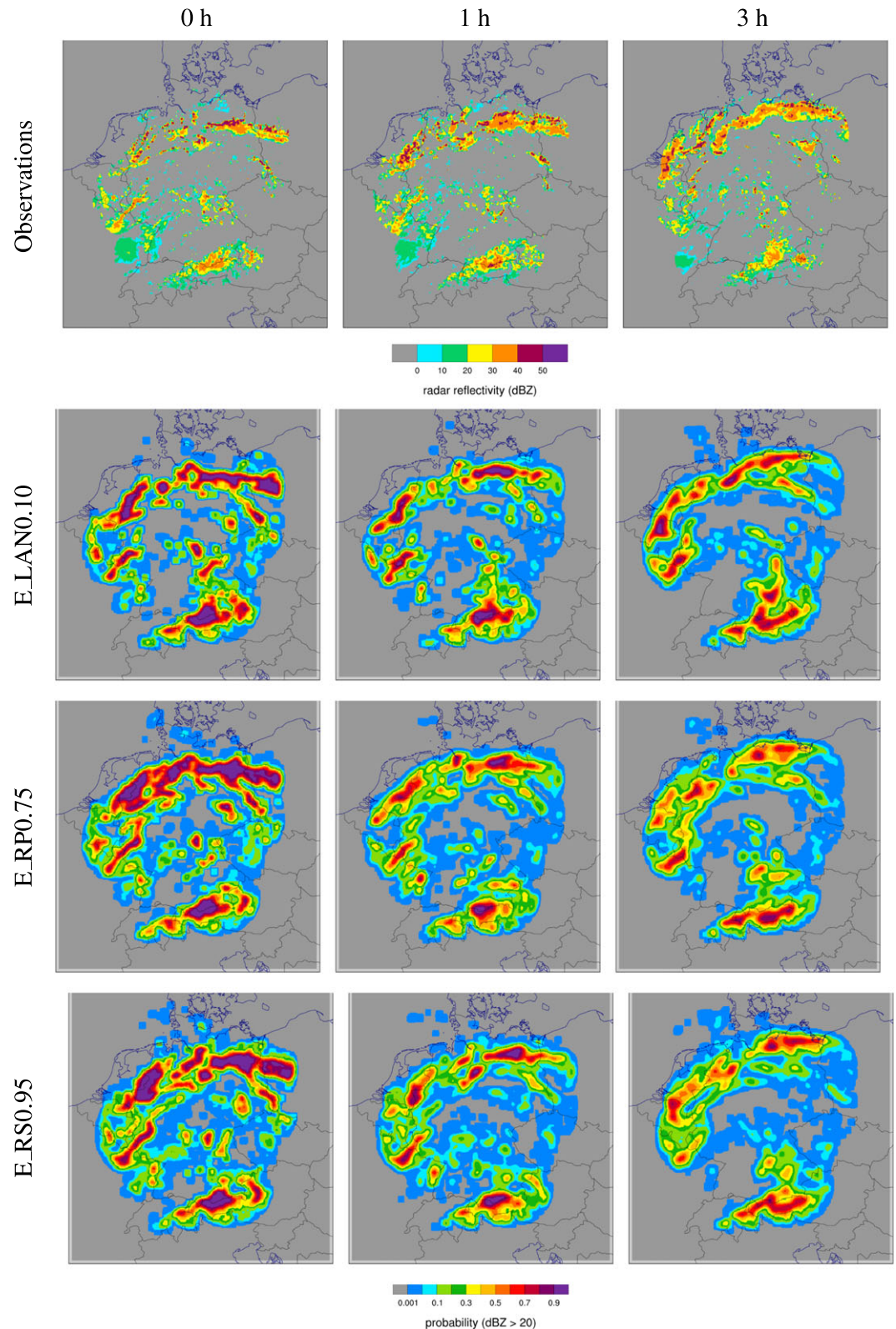
Figure 6. Comparison of E_RP0.75, E_LAN0.10, E_LAN0.10RP0.25, and E_LAN0.10RP0.75 of Study 1 by the verification of 6-hr ensemble forecasts (including initial hours) against radar reflectivity observations from the lowest elevation (0.5°), using the FSS with scale 30 km and threshold values 20 and 30 dBZ. Each FSS value is computed as an average over all 63 forecast runs (the study period contains 7 days, and each day has 9 forecast runs). The lines are marked as filled dots at the forecast lead times where the differences compared to E_LAN0.10 are statistically significant at 95% confidence intervals after 10,000 bootstrap resampling based on 63 difference samples. FSS = fraction skill score.

Figure 5 demonstrates the RMSE of 6-hr ensemble forecasts (including initial hours) for 2-m relative humidity, temperature, dew point temperature, and 10-m horizontal wind verified against the surface SYNOP observations. Since 2-m temperature and humidity data are not assimilated but 10-m horizontal wind data assimilated, the RMSE of wind increases in the first hour. Despite some small (statistically insignificant) discrepancies at initial hours, all experiments performs quite the same in the 6-hr forecast period. It seems that the short-range forecasts of surface model variables are not very sensitive to details in the atmospheric assimilation. It may be because the low level profiles will quickly adapt (back) to the soil conditions that have not been significantly changed in short time. More differences can be expected for a longer time scale since the upper air assimilation through instabilities and precipitation will start to influence the soil.

Figure 6 shows the verification of 6-hr ensemble forecasts (including initial hours) against radar reflectivity observations from the lowest elevation (0.5°), using the FSS. For threshold value 20 dBZ, E_LAN0.10 and E_LAN0.10RP0.25 are fairly close; E_LAN0.10 is better than E_LAN0.10RP0.75 and E_RP0.75 with statistical significance and the advantage over E_LAN0.10RP0.75 (E_RP0.75) increases (decreases) with increasing forecast lead time. Similar features can be seen for 30 dBZ with even larger advantage of E_LAN0.10 and E_LAN0.10RP0.25 (up to 5%), and E_LAN0.10 becomes gradually better than E_LAN0.10RP0.25 although the differences are not statistically significant.

Figure 7 illustrates 3-hr ensemble forecasts of E_LAN0.10 and E_RP0.75 initiated at 14:00 UTC 30 May 2016 and compares them with the reflectivity observation from the lowest elevation (0.5°) in a composite form. The simulated reflectivities in ensemble are represented by probability, defined as ratio of the number of ensemble members that exceed the given threshold value (here 20 dBZ) over the ensemble size. At 14:00 UTC (i.e., initial time), observations show multiple isolated convective cells in the northwest and a more organized squall line in the northeast. The locations and intensity of observed reflectivities are fairly captured by E_LAN0.10 and most of locations where reflectivities ≥ 20 dBZ are observed are reproduced by a great fraction of ensemble members (≥ 50%). The major precipitation system is also present in E_RP0.75; however, the areal coverage is much broader than in E_LAN0.10 and than in observations, especially in the northern part of the domain and for high probabilities (≥50%). This indicates that a considerable number of ensemble members in E_RP0.75 are associated with spurious cells that do not exist in observations. At 15:00 UTC (i.e., the first forecast hour), the observed cells have hardly moved but the intensity has strengthened in most places. Although the overall distribution of observed reflectivities are also well reproduced by E_LAN0.10, the area with high probabilities (≥50%) at places where reflectivities ≥20 dBZ are actually observed has considerably shrunk instead of extension compared to 14:00 UTC, which indicates that a number of valid cells in ensemble members are dissipated during the model spin-up. E_RP0.75 exhibits a much smoother field with larger areal coverage. At 17:00 UTC (i.e., 3-hr forecast), the observed precipitation structure is still well captured with high probabilities in E_LAN0.10 but poorly present in E_RP0.75 with increasing smoothed field. The same comparison has been done for threshold value 30 dBZ, and same fashion of differences between E_LAN0.10 and E_RP0.75 can be seen (not shown). This explains the advantage of E_LAN0.10 over E_RP0.75 in Figure 6. In total, it can be stated that the ensemble members in E_RP0.75 are associated with more widespread and smoother reflectivity field than in E_LAN0.10. It is consistent to some extent with results in Bowler et al. (2017). They found that the RTPP (with $\alpha_p = 0.85$) can effectively weaken the assimilation effect that decreases both the power and the ensemble spread at large scales but such a large α_p tends to strengthen multivariate

Figure 7. Three-hour evolution of reflectivities from the lowest elevation 0.5°, starting at 14:00 30 May. (first row) Observations; (second row) probabilities of E_LAN0.10, defined as the number of ensemble members exceeding the threshold value 20 dBZ divided by ensemble size; (third and fourth rows) as second row but for E_RP0.75 and E_RS0.95.
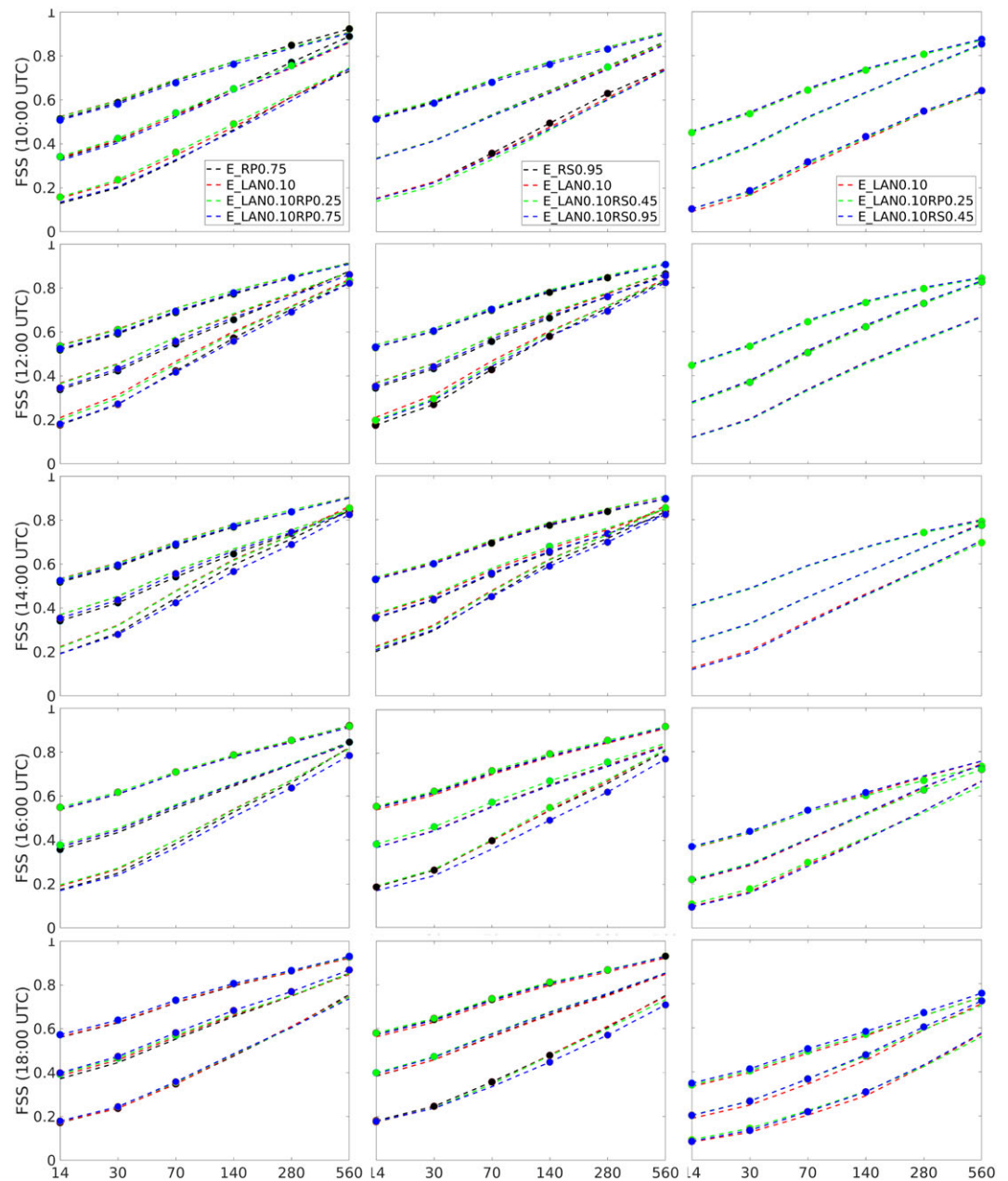
Figure 8. Verification against radar-derived precipitation for Study 1 (left), Study 2 (middle) and Study 3 (right): the FSS values for precipitation rate as a function of scale (from 14 to 560 km). The labels for the abscissa denote the scales for which the FSS is computed. In each panel, the upper, middle, and lower groups of lines are for threshold values 0.1, 1.0, and 5.0 mm/hr, respectively. Each group of lines shows scores for experiments in the ensemble forecast runs. The five rows are forecasts, initiated at 10:00, 12:00, 14:00, 16:00, and 18:00 UTC, respectively. Each FSS value is computed as an average over 42 FSS values (the study period contains 7 days, and the forecast lead time is 6 hr). The lines are marked as filled dots if the differences compared to E_LAN0.10 are statistically significant at 95% confidence intervals after 10,000 bootstrap resampling based on 42 difference samples.

coupling between wind and pressure, resulting in perturbations which are very balanced, at a level which may be not appropriate for perturbations.

Figure 8 depicts the FSS values of precipitation rate (derived from terrain-following precipitation scans that differ from radar scans used in assimilation) for different threshold values (0.1 mm/hr ≈ 7 dBZ, 1.0 mm/hr ≈ 23 dBZ, and 5.0 mm/hr ≈ 34 dBZ) as a function of scale. The panels correspond to different initial
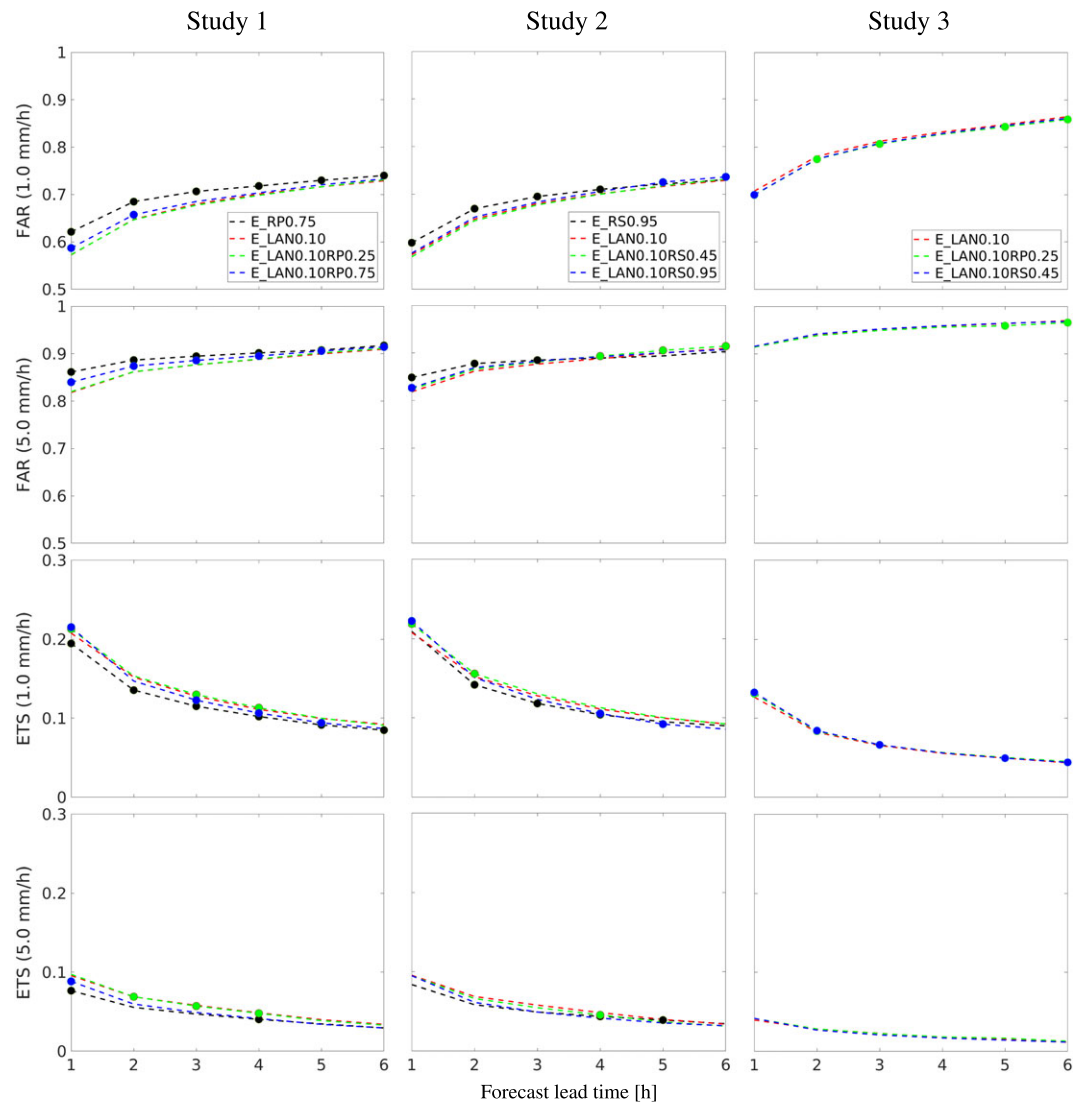
Figure 9. Verification of 6-hr ensemble forecasts against radar-derived precipitation for Study 1 (left), Study 2 (middle), and Study 3 (right), based on the FAR (first and second rows for threshold values 1.0 and 5.0 mm/hr, respectively) and ETS (third and fourth rows for threshold values 1.0 and 5.0 mm/hr, respectively). Each FAR/ETS value is computed as an average over all 63 forecast runs (the study period contains 7 days, and each day has 9 forecast runs). The lines are marked as filled dots at the forecast lead times where the differences compared to E_LAN0.10 are statistically significant at 95% confidence intervals after 10,000 bootstrap resampling based on 63 difference samples. ETS = equitable threat score; FAR = false alarm rate.

forecast times. First of all, there are statistically significant differences for 1.0 mm/hr at all initial times but they are very small. At 10:00 UTC, E_LAN0.10RP0.25 is slightly better than E_LAN0.10 and the both are moderately better than E_LAN0.10RP0.75 and E_RP0.75 for 5.0 mm/hr. At 12:00 and 14:00 UTC, E_LAN0.10 and E_LAN0.10RP0.25 are moderately better than E_LAN0.10RP0.75 and E_RP0.75 for 1.0 mm/hr and much better for 5.0 mm/hr. At 16:00 UTC, E_LAN0.10 and E_LAN0.10RP0.25 are slightly better than E_RP0.75 and moderately better than E_LAN0.10RP0.75 for 5.0 mm/hr. At 18:00 UTC, E_LAN0.10RP0.75 is slightly better than the others for 1.0 mm/hr. Furthermore, it is noticed by comparison of E_RP0.75 and E_LAN0.10 that the advantage of E_LAN0.10 is more clear at smaller scales and it decreases with the increasing scale, likely due to the smoothing effect of the RTTP.

Figure 9 compares 6-hr ensemble forecasts of precipitation rate, using the FAR and ETS for threshold values 1.0 and 5.0 mm/hr. With respect to the FAR, for both 1.0 and 5.0 mm/hr, E_LAN0.10 and E_LAN0.10RP0.25 are very close, and they are slightly better than E_LAN0.10RP0.75 and much better than E_RP0.75, which indi-
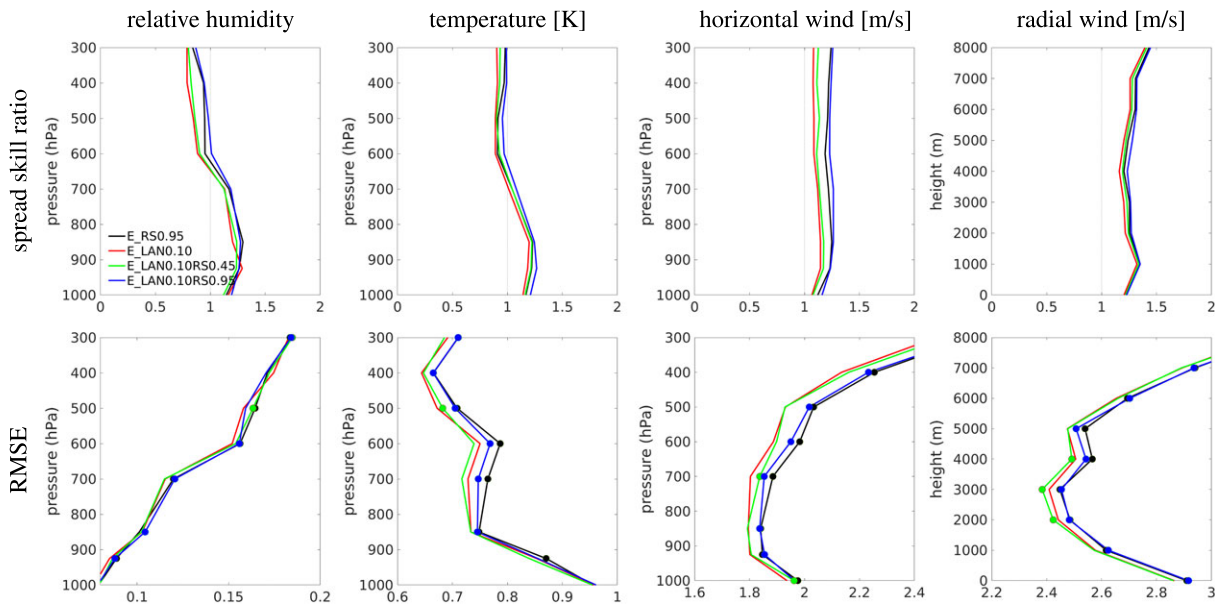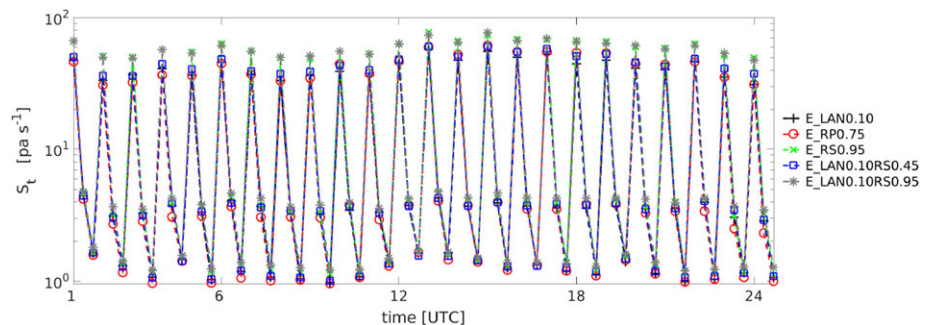
Figure 10. Same as Figure 4 but for E_RS0.95, E_LAN0.10, E_LAN0.10RS0.45, and E_LAN0.10RS0.95 of Study 2. RMSE = root-mean-square error.

cates much less spurious convection produced in E_LAN0.10 and E_LAN0.10RP0.25, and this is consistent with Figure 7. With respect to the ETS, E_LAN0.10 and E_LAN0.10RP0.25 are also equally skillful for 1.0 and 5.0 mm/hr, and they are better than E_LAN0.10RP0.75 and much better than E_LAN0.10. It can be also seen that the advantage of E_LAN0.10 decrease gradually with increasing forecast lead time and may not be statistically significant for longer lead times at some occasions.

In summary, the LAN alone is able to provide sufficient spread for horizontal and radial winds, although for relative humidity and temperature there is some insufficiency above 600 hPa. The RMSE decreases as α_p decreases in combinations, and the LAN alone actually performs the best. In case of 6-hr ensemble forecasts verified against SYNOP observations, no distinguishable differences can be found. In case of 6-hr ensemble forecasts verified against radar reflectivity observations as well as radar-derived precipitation rates, the scores decrease with the increasing α_p in combinations and the LAN alone performs equally or better than combinations.
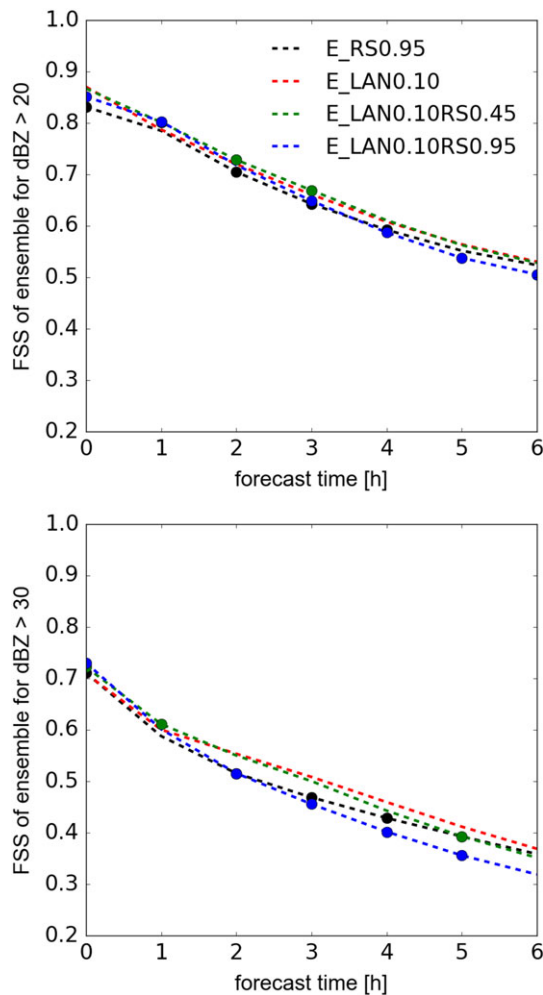
5.3. Study 2

Similar study has been done with the RTPS instead of RTPP. Again, we compare performance of experiments E_RS0.95, E_LAN0.10, E_LAN0.10RS0.45, and E_LAN0.10RS0.95 (see Table 1) based on r_s and RMSE as illustrated in Figure 10. In general, the whole pattern of differences is quite close to Figure 4. With respect to r_s, all experiments overestimate the spread up to 600 hPa and underestimate the spread above 600 hPa for relative humidity and temperature. For horizontal and radial wind, all experiments are overall associated with overestimated spread and E_LAN0.10 has the best r_s, followed by E_LAN0.10RS0.45, E_RS0.95, and E_LAN0.10RS0.95.

Figure 11. Surface pressure tendency S_t averaged over all ensemble members for the period 01:00 UTC 1 June to 00:00 UTC 2 June for E_LAN0.10, E_RP0.75, E_RS0.95, E_LAN0.10RS0.45, and E_LAN0.10RS0.95.

Figure 12. Same as Figure 6 but for E_RS0.95, E_LAN0.10, E_LAN0.10RS0.45, and E_LAN0.10RS0.95 of Study 2. RMSE = root-mean-square error.

With respect to the RMSE, for relative humidity, E_LAN0.10 and E_LAN0.10RS0.45 are similar although there is one level (i.e., 500 hPa) with statistical significance in favor of E_LAN0.10; for temperature, horizontal and radial wind, E_LAN0.10 is considerably better than E_RS0.95 and E_LAN0.10RS0.95 with statistical significance almost at all levels. E_LAN0.10 is comparable to E_LAN0.10RS0.45 in general, while for temperature and horizontal wind E_LAN0.10 is slightly better than E_LAN0.10RS0.45 with statistical significance at few levels and for radial wind E_LAN0.10RS0.45 is slightly better between heights 2,000 and 4,000 m.

It should be emphasized that the RTPS actually does purely multiplicative inflation of the analysis ensemble perturbation $\mathbf{X}^a$, which is calculated for each variable and grid point separately. This discontinuous fashion of manipulation on the analysis ensemble may deteriorate the existing covariances between model variables (e.g., temperature and wind fields; Whitaker & Hamill, 2012), which may cause unbalanced model states and consequently unphysical noise. An often used metric to quantify noise is surface pressure tendency, hereafter denoted by S_t, usually defined as the sum of absolute values of first time derivative of surface pressure at each grid point, averaged over the model domain (Stuffer & Seaman, 1990). Figure 11 shows the half-hourly evolution of S_t during the assimilation cycles (recall that the assimilation window is 1 hr), averaged over all the ensemble members, for 1 day of 1 June 2016. It is evident that E_RS0.95 introduces much more noise into analyses than the other experiments. One hour model spin-up time is able to reduce a great part of noise, but the noise in forecasts of E_RS0.95 is higher than the others. Moreover, the comparison of E_LAN0.10, E_LAN0.10RS0.45 and E_LAN0.10RS0.95 reveals that less noise arises with smaller α_s.

For verification of 6-hr ensemble forecasts of 2-m relative humidity, temperature, dew point temperature, and wind speed against the SYNOP data, all experiments also performs equally (not shown), in spite of small differences at the initial time as seen in Figure 5.

As Figure 6, Figure 12 shows the FSS values of 6-hr ensemble forecasts (including initial hours) verified against radar reflectivity observations. For threshold value 20 dBZ, E_LAN0.10 and E_LAN0.10RS0.45 are slightly better than E_LAN0.10RS0.95 and much better than E_RS0.95 initially. E_LAN0.10 degrades a bit in the first few hours but ends up close to E_LAN0.10RS0.45. For 30 dBZ, all experiments have similar starts but their performance diverges clearly during the forecast period: E_LAN0.10 becomes gradually superior, up to 3% better than E_LAN0.10RS0.45, up to 5% better than E_RS0.95 and more than 5% better than E_LAN0.10RS0.95 with statistical significance.

The fourth row of Figure 7 shows probabilities of simulated reflectivities of ensemble in E_RS0.95. At 14:00 UTC, the whole precipitation pattern is also well represented by high probabilities ($\geq$ 50%) although the areal coverage is considerably larger than E_LAN0.10 and comparable to E_RP0.75. At 15:00 and 16:00 UTC, the distribution of high probabilities is close to E_LAN0.10 and less smooth than E_RP0.75, but with some cells missing (e.g., in the middle of domain).

The verification of precipitation rates of 6-hr ensemble forecasts by the FSS is given in the right column of Figure 8. At 10:00 UTC, E_RS0.95 is slightly better than the others for 5.0 mm/hr. At 12:00 UTC, E_LAN0.10 and E_LAN0.10RS0.45 are slightly better than the others for 1.0 mm/hr. E_LAN0.10 is slightly better than E_LAN0.10RS0.45 and moderately better than E_RS0.95 and E_LAN0.10RS0.95 for 5.0 mm/hr. Similar differences can be also distinguished at 14:00 UTC. At 16:00 UTC, E_LAN0.10RS0.45 is slightly better than the others for 1.0 mm/hr; E_RS0.95, E_LAN0.10, and E_LAN0.10RS0.45 are comparable and much better than E_LAN0.10RS0.95 for 5.0 mm/hr, which can be also seen for 18:00 UTC. Regarding the scale dependence, the RTPS seems to work well rather at small scales than at large scales (e.g., 5.0 mm/hr at 16:00 and 18:00 UTC), which is opposite to the RTPP.
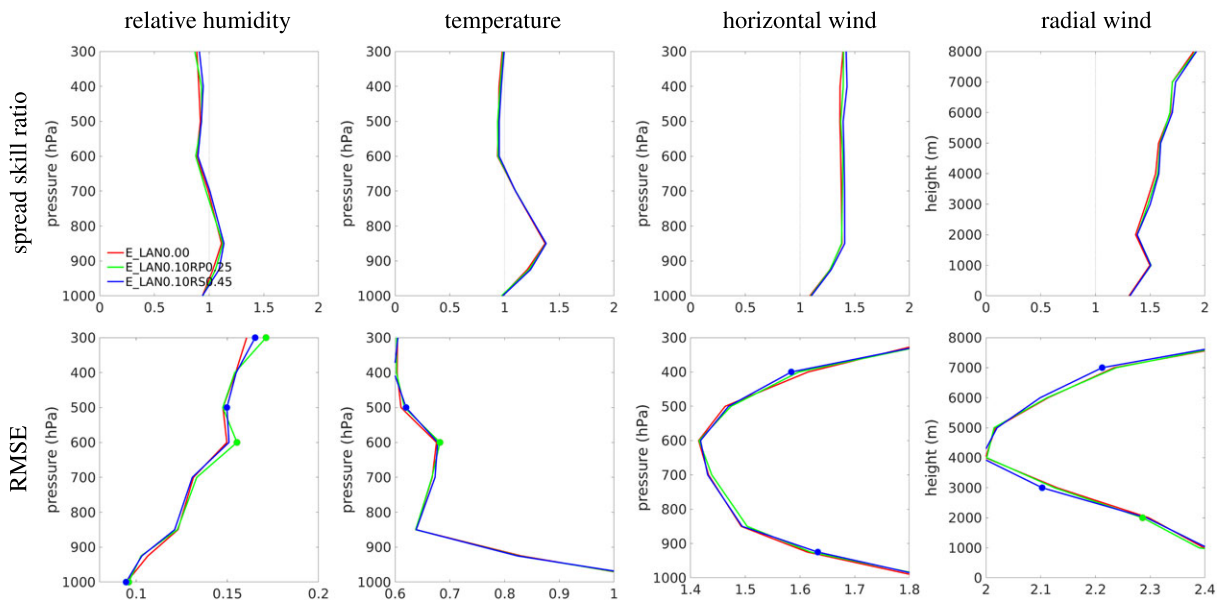
Figure 13. Same as Figure 4 but for E_LAN0.10, E_LAN0.10RP0.25, and E_LAN0.10RS0.45 of Study 3. RMSE = root-mean-square error.

In Figure 9, based on the FAR for both 1.0 and 5.0 mm/hr, E_LAN0.10, E_LAN0.10RS0.45 and E_LAN0.10RS0.95 are comparable and they are slightly better than E_RS0.95 up to four hours, after which E_LAN0.10RS0.95 (E_LAN0.10RS0.45) becomes slightly worse than E_RS0.95 for 1.0 (5.0) mm/hr. Based on the ETS for 1.0 mm/hr, E_RS0.95 is slightly worse than E_LAN0.10; E_LAN0.10RS0.45 is slightly better than E_LAN0.10 in the first two hours and then they approach. E_LAN0.10RS0.95 is also slightly better than E_LAN0.10 in the first two hours but it becomes inferior afterward. For 5.0 mm/hr, E_LAN0.10 is marginally better than E_LAN0.10RS0.45, followed sequentially by E_LAN0.10RS0.95 and E_RS0.95.

To conclude, the LAN alone generally results in the smallest RMSE, compared to combinations with the RTPS, and the RMSE decreases as α_s decreases in combinations. This is also valid for verification of 6-hr ensemble forecasts against reflectivity and precipitation rate. The RTPS may not have smoothing issues at small scales, but it is prone to producing more noise in analyses (Zeng & Janjić, 2016; Zeng et al., 2017), which may contribute to the degraded quality of ensemble forecasts as shown in Bick et al. (2016).
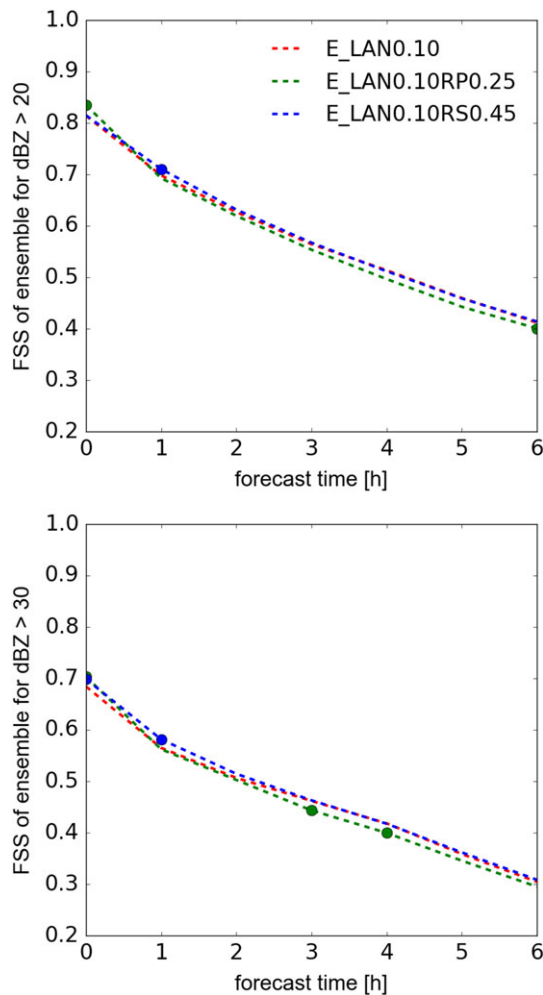
5.4. Study 3

In this study, we choose three experimental setups, which have performed well in previous studies: E_LAN0.10, E_LAN0.10RP0.25, and E_LAN0.10RS0.45 and focus on weakly forced case. As shown in Figure 13, E_LAN0.10RS0.45 has overall the largest r_s, followed by E_LAN0.10 that is E_LAN0.10RP0.25 for all observed variables, while all experiments are associated with overestimated (underestimated) spread in the lower (higher) atmosphere for relative humidity and temperature and overestimated spread for horizontal and radial wind. With respect to the RMSE, for relative humidity and temperture, E_LAN0.10 is marginally better than the others at some levels with statistical significance, while for horizontal and radial wind E_LAN0.10RS0.45 is marginally better than the others at few levels.

For the verification of 6-hr ensemble forecasts against reflectivity (see Figure 14), the FSS values are generally lower than in Figure 12 due to the weak forcing and less predictability (Barthlott et al., 2011; Trentmann et al., 2009). The FSS values of E_LAN0.10 and E_LAN0.10RS0.45 are similar and slightly higher than those of E_LAN0.10RP0.25 for both threshold values (20 and 30 dBZ). This is a bit different from Figure 12, where E_LAN0.10 is somewhat better than E_LAN0.10RS0.45 for 30 dBZ.

In terms of the precipitation verification (see third column of Figure 8), E_LAN0.10RS0.45 is close to E_LAN0.10RP0.25 and marginally better than E_LAN0.10 for 5.0 mm/hr at 10:00 UTC. At 18:00 UTC, E_LAN0.10RP0.25 and E_LAN0.10RS0.45 are slightly better than E_LAN0.10 for 1.0 and 5.0 mm/hr.

According to the FAR for 1.0 mm/hr in Figure 9, E_LAN0.10RP0.25 and E_LAN0.10RS0.45 are marginally better than E_LAN0.10. For 5.0 mm/hr, E_LAN0.10RP0.25 may have a slight advantage over the others at the end.

Figure 14. Same as Figure 6 but for E_LAN0.10, E_LAN0.10RP0.25, and E_LAN0.10RS0.45 of Study 3. FSS = fraction skill score.

With respect to the ETS for 1.0 mm/hr, E_LAN0.10RP0.25 and E_LAN0.10RS0.45 are marginally better than E_LAN0.10 for the first 2 hr and then they become undistinguishable. For 5.0 mm/hr, no statistically significant differences can be seen.

In general, the performance of the LAN alone degrades a bit under weather regimes with weak large-scale forcing, especially in comparison to E_LAN0.10RS0.45. For instance, the RMSE of E_LAN0.10 is smaller than E_LAN0.10RS0.45 for horizontal wind in Study 2, while it is comparable for horizontal wind in this study. E_LAN0.10RS0.45 is also slightly better based on verification of 6-hr ensemble forecasts against precipitation rate. Nevertheless, the performance of the LAN alone is still very good.

6. Conclusion and Outlook

Background error covariance should incorporate both sampling and model errors for ensemble data assimilation. To address this issue, additive noise and relaxation methods (RTPP and RTPS) are often used. In this work, we studied the performance of those methods in the framework of convective-scale data assimilation. This is done using the operational convection-permitting COSMO-DE model and data assimilation system KENDA of DWD, for a 2-week convective period in May 2016 over Germany. Conventional and radar reflectivity observations were assimilated hourly by the LETKF. Our focus was prediction of short range ensemble forecasts.

To take into account in regional data assimilation the information about uncertainties arising from the large scale forcing model, we use additive noise with the samples from the climatological atmospheric background error covariance of the global EnVar data assimilation system for the ICON model that provides lateral boundary conditions for the COSMO-DE model. Therefore, this additive noise scheme is expected to be particularly effective for weather regimes characterized by strong large-scale forcing of convection. Contrary to additive noise, RTPP and RTPS use information that is already available to data assimilation about the scales that are resolved by model. Both errors on the larger scale as well as errors in unresolved scales remain unrepresented during data assimilation, but their effect is mitigated by trusting less the analysis ensemble (RTPP) or by not reducing ensemble spread (RTPS). To investigate this, a test period is chosen, with strong forcing in the first week and weak forcing in the second week, and additive noise is compared to relaxation methods as well as combinations (i.e., additive noise + RTPP or RTPS) in a series of experiments, which have been conducted for the first and second week separately.

In studies on days with strong forcing, innovation statistics of assimilation cycles show that, based on upper air verification, additive noise alone results overall in best spread skill ratio except for relative humidity and temperature in the upper atmosphere. Moreover, it also results in the smallest RMSE and the RMSE tendentiously decreases as relaxation methods are less used in combinations. With respect to 6-hr ensemble forecasts, the performance of all experiments is similar in terms of verification against surface observations. However, they are quite distinguishable if verified against reflectivity composite and precipitation rate, where additive noise alone has statistically significant advantage. Also here the benefits of decreasing the influence of relaxation methods in combinations are noticeable. In addition, it is shown that the RTPP may be associated with smoothing side effects. This disadvantage for short range precipitation forecasts is validated by verification using scale-depending metric (e.g., fraction skill score), which shows that the RTPP can harm especially smallest resolvable scales. These results are consistent with findings of Bowler et al. (2017) that the RTPP may produce overbalanced perturbations. The RTPS may not have problems at smallest resolvable scales, but it usually results in unbalanced model states, which could deteriorate the quality of forecasts as well. Generally speaking, additive noise performs equally or even better than relaxation methods and combinations during both assimilation and short-range forecasts in those experiments. On days with weak forcing, the performance of additive noise relative to combinations degrades a bit but results are still comparable. Overall, it can be

concluded that additive noise can partially account for model error and its performance is not very sensitive to weather regimes. It should be pointed out that our results are based on the specific regions over 2-week convective period under specific configurations of the data assimilation system, including the ensemble size, localization strategy, and observation network. However, our results do imply that it is surprisingly difficult to outperform additive noise, which may provide a good benchmark for further development in representation of model error for data assimilation.

Since additive noise introduced in this work samples background error covariance used by the global data assimilation, it may be mainly representative for large-scale error and does not explicitly represent small-scale uncertainties of the regional model. For a further improvement, mechanisms that address small-scale uncertainties are needed. The relaxation methods are not designed particularly for this either. However, the multimodel/multiphysics method presented in Meng and Zhang (2007), which has been used successfully in Houtekamer et al. (2009) and Meng and Zhang (2008a, 2008b) might be an option. Further especially, the simultaneously state and parameter estimation method presented in Hu et al. (2010) that directly perturbs the physical parameterization uncertainties and updates the uncertain parameters with the EnKF could represent boundary layer uncertainties. A review of alternative methods can be found in Houtekamer and Zhang (2016) and Meng and Zhang (2011). Also one straightforward idea may be to build a climatological background error covariance for the regional model and use it in analogous way to additive noise. Another idea is based on the fact that the refinement of the horizontal resolution could improve the short-range convective-scale precipitation forecasts (Buzzi et al., 2014; Clark et al., 2016; Raynaud & Bouttier, 2017). It motivates us to incorporate the model truncation error into convective-scale data assimilation in the form of additive noise as suggested in Hamill and Whitaker (2005). This implementation is ongoing; the first results are very promising, especially for the weak forcing period, and will be published soon in another article.

References

Aksoy, A., Dowell, D. C., & Snyder, C. (2009). A multiscale comparative assessment of the ensemble Kalman filter for assimilation of radar observations. Part I: Storm-scale analyses. *Monthly Weather Review*, *137*, 1805–1824.

Anderson, J. L., & Anderson, S. L. (1999). A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Monthly Weather Review*, *127*, 2741–2758.

Baldauf, M., Seifert, A., Förstner, J., Majewski, D., Raschendorfer, M., & Reinhardt, T. (2011). Operational convective-scale numerical weather prediction with the COSMO model: Description and sensitivities. *Monthly Weather Review*, *139*, 3887–3905.

Bannister, R. N. (2008). A review of forecast error covariance statistics in atmospheric variational data assimilation. I: Characteristics and measurements of forecast error covariances. *Quarterly Journal of the Royal Meteorological Society*, *134*, 1951–1970.

Barthlott, C., Burton, R., Kirshbaum, D., Hanley, K., Richard, E., Chaboureau, J. P., et al. (2011). Initiation of deep convection at marginal instability in an ensemble of mesoscale models: A case-study from cops. *Quarterly Journal of the Royal Meteorological Society*, *137*(S1), 118–136.

Bei, N., & Zhang, F. (2007). Impacts of initial condition errors on mesoscale predictability of heavy precipitation along the Mei-Yu front of china. *Quarterly Journal of the Royal Meteorological Society*, *133*, 83–99.

Bick, T., Simmer, C., Trömel, S., Wapler, K., Stephan, K., Blahak, U., et al. (2016). Assimilation of 3D-radar reflectivities with an ensemble Kalman filter on the convective scale. *Quarterly Journal of the Royal Meteorological Society*, *142*, 1490–1504.

Bowler, N. E., Clayton, A. M., Jardak, M., Lee, E., Lorenc, A. C., Piccolo, C., et al. (2017). Inflation and localization tests in the development of an ensemble of 4D-ensemble variational assimilations. *Quarterly Journal of the Royal Meteorological Society*, *143*, 1280–1302.

Bryan, G. H., Wyngaard, J. C., & Fritsch, J. M. (2003). Resolution requirements for the simulation of deep moist convection. *Monthly Weather Review*, *131*, 2394–2416.

Buizza, R., Milleer, M., & Palmer, T. (1999). Stochastic representation of model uncentainties in the ECMWF ensemble prediciton system. *Quarterly Journal of the Royal Meteorological Society*, *125*(560), 2887–2908.

Buzzi, A., Davolio, S., Malguzzi, P., Drofa, O., & Mastrangelo, D. (2014). Heavy rainfall episodes over Liguria in autumn 2011: Numerical forecasting experiments. *Natural Hazards and Earth System Sciences*, *14*, 1325–1340.

Caya, A., Sun, J., & Snyder, C. (2005). A comparison between the 4D-VAR and the ensemble Kalman filter techniques for radar data assimilation. *Monthly Weather Review*, *133*, 3081–3094.

Clark, P., Roberts, N., Lean, H., Ballard, S. P., & Charlton-Perez, C. (2016). Convection-permitting models: A step-change in rainfall forecasting. *Meteorological Applications*, *23*(2), 165–181.

Dee, D. P. (1995). On-line estimation of error covariance parameters for atmospheric data assimilation. *Monthly Weather Review*, *123*, 1128–1145.

Desroziers, G., Berre, L., Chapnik, B., & Poli, P. (2005). Diagnosis of observation, background and analysis-error statistics in observation space. *Quarterly Journal of the Royal Meteorological Society*, *131*, 3385–3396.

Desroziers, G., Brousseau, P., & Chapnik, B. (2005). Use of randomization to diagnose the impact of observations on analyses and forecasts. *Quarterly Journal of the Royal Meteorological Society*, *131*, 2821–2837.

Doms, G., & Baldauf, M. (2015). A description of the nonhydrostatic regional cosmo-model LM. Part I: Dynamics and numerics (Tech. rep.): Consortium for Smallscale Modeling (COSMO) (p. 158). Retrieved from http://www.cosmo-model.org/content/model/documentation/core/cosmoDyncsNumcs.pdf

Doms, G., Förstner, J., Heise, E., Herzog, H.-J., Mirononv, D., Raschendorfer, M., et al. (2011). A description of the nonhydrostatic regional cosmo-model. Part II: Physical parameterization (Tech. rep.): Consortium for Smallscale Modeling (COSMO) (p. 154). Retrieved from http://www.cosmo-model.org/content/model/documentation/core/cosmoPhysParamtr.pdf

Dowell, D. C., & Wicker, L. J. (2009). Additive noise for storm-scale ensemble data assimilation. *Monthly Weather Review*, *26*, 911–927.

Dowell, D. C., Wicker, L. J., & Snyder, C. (2011). Ensemble Kalman filter assimilation of radar observations of the 8 May 2003 Oklahoma City supercell: Influences of reflectivity observations on storm-scale analyses. *Monthly Weather Review*, *139*, 272–294.

Dowell, D. C., Zhang, F., Wicker, L. J., Snyder, C., & Crook, N. A. (2004). Wind and temperature retrievals in the 17 May 1981 Arcadia, Oklahoma, supercell: Ensemble Kalman filter experiments. *Monthly Weather Review*, *132*, 1982–2005.

Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Abingdon, UK: Chapman & Hall/CRC.

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, *99*(10), 143–162.

Gao, J., & Xue, M. (2008). An efficient dual-resolution approach for ensemble data assimilation and tests with simulated Doppler radar data. *Advances in Atmospheric sciences*, *136*, 945–963.

Gaspari, G., & Cohn, S. E. (1999). Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, *125*, 723–757.

Gneiting, T. (1999). Correlation functions for atmospheric data analysis. *Quarterly Journal of the Royal Meteorological Society*, *125*, 2449–2464.

Gustafsson, N., Janjić, T., Schraff, C., Leuenberger, D., Weissman, M., Reich, H., et al. (2018). Survey of data assimilation methods for convective-scale numerical weather prediction at operational centres. *Quarterly Journal of the Royal Meteorological Society*, *144*, 1218–1256. https://doi.org/10.1002/qj.3179

Hamill, T. M., & Whitaker, J. S. (2005). Accounting for the error due to unresolved scales in ensemble data assimilation. *Monthly Weather Review*, *133*, 3132–3147.

Hamrud, M., Bonavita, M., & Isaksen, L. (2015). EnKF and hybrid gain ensemble data assimilation. Part I: EnKF implementation. *Monthly Weather Review*, *143*(12), 4847–4864.

Harnisch, F., & Keil, C. (2015). Initial conditions for convective-scale ensemble forecasting provided by ensemble data assimilation. *Monthly Weather Review*, *143*(5), 1583–1600.

Hohenegger, C., & Schär, C. (2007). Predictability and error growth dynamics in cloud-resolving models. *Journal of Atmospheric Sciences*, *64*, 4467–4478.

Houtekamer, P. L., Mitchell, H. L., Buehner, M., Charron, M., Spacek, L., & Hansen, M. (2005). Atmospheric data assimilation with an ensemble Kalman filter: Results with real observations. *Monthly Weather Review*, *133*, 604–620.

Houtekamer, P. L., Mitchell, H. L., & Deng, X. (2009). Model error representation in an operational ensemble Kalman filter. *Monthly Weather Review*, *137*, 2126–2143.

Houtekamer, P. L., & Zhang, F. (2016). Review of the ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review*, *144*, 4489–4532.

Hu, X.-M., Zhang, F., & Nielsen-Gammon, J. W. (2010). Ensemble-based simultaneous state and parameter estimation for treatment of mesoscale model error: A real-data study. *Geophysical Research Letters*, *37*, L08802. https://doi.org/10.1029/2010GL043017

Hunt, B. R., Kostelich, E. J., & Szunyogh, I. (2007). Efficient data assimilation for spatiotemporal chaos: A Local Ensemble Transform Kalman Filter. *Physica D: Nonlinear Phenomena*, *230*, 112–126.

Johnson, A., Wang, X., Carley, J. R., Wicker, L., & Karstens, C. (2015). A comparison of multiscale GSI-based EnKF and 3DVar data assimilation using radar and conventional observations for midlatitude convective-scale precipitation forecasts. *Monthly Weather Review*, *143*, 3087–3108.

Keil, C., Heinlein, F., & Craig, G. C. (2014). The convective adjustment time-scale as indicator of predictability of convective precipitation. *Quarterly Journal of the Royal Meteorological Society*, *140*, 480–490.

Kotsuki, S., Ota, Y., & Miyoshi, T. (2017). Adaptive covariance relaxation methods for ensemble data assimilation: Experiments in the real atmosphere. *Quarterly Journal of the Royal Meteorological Society*, *143*, 2001–2015.

Kühnlein, C., Keil, C., Craig, G. C., & Gebhardt, C. (2014). The impact of downscaled initial condition perturbations on convective-scale ensemble forecasts of precipitation. *Quarterly Journal of the Royal Meteorological Society*, *140*, 1552–1562.

Lange, H., & Craig, G. C. (2014). The impact of data assimilation length scales on analysis and prediciton of convective storms. *Monthly Weather Review*, *142*, 3781–3808.

Lange, H., Craig, G. C., & Janjić, T. (2017). Characterizing noise and spurious convection in convective data assimilation. *Quarterly Journal of the Royal Meteorological Society*, *143*, 3060–3069.

Lange, H., & Janjić, T. (2016). Assimilation of Mode-S EHS aircraft observations in COSMO-KENDA. *Monthly Weather Review*, *144*, 1697–1711.

Lin, Y., Farley, R. D., & Orville, H. D. (1983). Bulk parameterization of the snow field in a cloud model. *Journal of Applied Meteorology and Climatology*, *22*, 1065–1092.

Meng, Z., & Zhang, F. (2007). Tests of an ensemble Kalman filter for mesoscale and regional-scale data assimilation. Part II: Imperfect model experiments. *Monthly Weather Review*, *135*(4), 1403–1423.

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

Meng, Z., & Zhang, F. (2008b). Tests of an ensemble Kalman filter for mesoscale and regional-scale data assimilation. Part IV: Comparison with 3DVAR in a month-long experiment. *Monthly Weather Review*, *136*, 3671–3682.

Meng, Z., & Zhang, F. (2011). Limited-area ensemble-based data assimilation. *Monthly Weather Review*, *139*(7), 2025–2045.

Murphy, J. M., Sexton, D. M., Barnett, G. S., Jones, D. N., Webb, M. J., Collins, M., & Sainforth, D. (2004). Quantification of modelling uncentainties in a large ensemble of climate change simulations. *Nature*, *430*(7001), 768–772.

Parrish, D. F., & Derber, J. C. (1992). The National Meteorological Center's spectral statistical-interpolation analysis system. *Monthly Weather Review*, *120*, 1747–1763.

Perianez, A., Reich, H., & Potthast, R. (2014). Optimal localization for ensemble Kalman filter systems. *Journal of the Meteorological Society of Japan*, *92*, 585–597.

Piper, D., Kunz, M., Ehmele, F., Mohr, S., Mühr, B., Kron, A., & Daniell, J. (2016). Exceptional sequence of severe thunderstorms and related flash floods in May and June 2016 in Germany—Part 1: Meteorological background. *Natural Hazards and Earth System Sciences*, *16*, 2835–2850.

Poterjoy, J., Zhang, F., & Weng, Y. (2014). The effects of sampling errors on the ENKF assimilation of inner-core hurricane observations. *Monthly Weather Review*, *42*, 1609–1630.

Raschendorfer, M. (2001). The new turbulence parametrization of LM. *COSMO-Newsletter*, *1*, 89–97.

Raynaud, L., & Bouttier, F. (2017). The impact of horizontal resolution and ensemble size for convective-scale probabilistic forecasts. *Quarterly Journal of the Royal Meteorological Society*, *143*, 3037–3047.

Reinhardt, T., & Seifert, A. (2006). A three-category ice scheme for LMK. *COSMO News Letter*, *6*, 115–120.

Rhodin, A., Lange, H., Potthast, R., & Janjić, T. (2013). Documentation of the DWD data assimilation system (Tech. Rep.): Deutscher Wetterdienst (DWD).

Ritter, B., & Geleyn, J. F. (1992). A comprehensive radiation scheme for numerical weather prediction models with potential applications in climate simulations. *Monthly Weather Review*, *120*, 303–325.

Robert, N., & Lean, H. (2008). Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, *136*, 78–96.

Schraff, C., Reich, H., Rhodin, A., Schomburg, A., Stephan, K., Periáñez, A., & Potthast, R. (2016). Kilometre-scale ensemble data assimilation for the COSMO model (KENDA). *Quarterly Journal of the Royal Meteorological Society*, *142*, 1453–1472.

Seity, Y., Brosseau, P., Malardel, S., Hello, G., Bénard, P., Bouttier, F., et al. (2011). The AROME-France convective-scale operational model. *Monthly Weather Review*, *139*, 976–991.

Selz, T., & Craig, G. C. (2015). Upscale error growth in a high-resolution simulation of a summertime weather event over Europe. *Monthly Weather Review*, *143*, 813–827.

Shutts, G. (2005). A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Quarterly Journal of the Royal Meteorological Society*, *131*(612), 3079–3102.

Simmer, C., Adrian, G., Jones, S., Wirth, V., Göber, M., Hohenegger, C., et al. (2016). HErZ—The German Hans-Ertel Centre for weather research. *Bulletin of the American Meteorological Society*, *97*, 1057–1068. https://doi.org/10.1175/BAMS-D-13-00227.1

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Wang, W., & Powers, J. G. (2005). A description of the Advanced Research WRF version 2 (Tech. Rep.): NCAR Technical Note NCAR/TN-468+STR.

Snyder, C., & Zhang, F. (2003). Assimilation of simulated Doppler radar observations with an ensemble Kalman filter. *Monthly Weather Review*, *131*, 1663–1677.

Sobash, R. A., & Stensrud, D. (2013). The impact of covariance localization for radar data on ENKF analyses of a developing MCS: Observing system simulation experiment. *Monthly Weather Review*, *141*, 3691–3709.

Sommer, M., & Janjić, T. (2018). A flexible additive inflation scheme for treating model error in ensemble Kalman filters. *Quarterly Journal of the Royal Meteorological Society*. https://doi.org/10.1002/qj.3254

Stephan, K., Klink, S., & Schraff, C. (2008). Assimilation of radar-derived rain rates into convective-scale model COSMO-DE at DWD. *Quarterly Journal of the Royal Meteorological Society*, *134*, 1315–1326.

Stuffer, D., & Seaman, N. (1990). Use of four-dimensional data assimilation in a limited-area mesoscale model. Part I: Experiments with synoptic data. *Monthly Weather Review*, *118*, 1250–1277.

Sun, J., Wang, H., Tong, W., Zhang, Y., Lin, C.-Y., & Xu, D. (2016). Comparison of the impacts of momentum control variables on high-resolution variational data assimilation and precipitation forecasting. *Monthly Weather Review*, *144*, 149–169.

Sun, Y. Q., & Zhang, F. (2016). Intrinsic versus practical limits of atmospheric predictability and the significance of the butterfly effect. *Journal of Atmospheric Sciences*, *73*, 1419–1438.

Tang, Y., Lean, H. W., & Bornemann, J. (2013). The benefits of the Met Office variable resolution NWP model for forecasting convection. *Applied Meteorology*, *20*, 417–426.

Tiedtke, M. (1989). A comprehensive mass flux scheme for cumulus parameterization in large-scale models. *Monthly Weather Review*, *117*, 1779–1799.

Tong, M., & Xue, M. (2005). Ensemble Kalman filter assimilation of Doppler radar data with a compressible nonhydrostatic model: OSS experiments. *Monthly Weather Review*, *133*, 1789–1807.

Trentmann, J., Keil, C., Salzmann, M., Barthlott, C., Bauer, H., Schwitalla, T., et al. (2009). MCS rainfall forecast accuracy as a function of large-scale forcing. *Meteorology and Atmospheric Physics*, *103*, 95–103.

Vié, B., Nuissier, O., & Ducrocq, V. (2011). Cloud-resolving ensemble simulations of Mediterranean heavy precipitating events: Uncertainty on initial conditions and lateral boundary conditions. *Monthly Weather Review*, *139*, 403–423.

Weissmann, M., Göber, M., Hohenegger, C., Janjić, T., Keller, J., Ohlwein, C., et al. (2014). The Hans-Ertel Centre for Weather Research—Research objectives and highlights from its first three years. *Meteorologische Zeitschrift*, *23*(3), 193–208.

Whitaker, J. S., & Hamill, T. M. (2012). Evaluating methods to account for system errors in ensemble data assimilation. *Monthly Weather Review*, *140*(9), 3078–3089.

Whitaker, J. S., Hamill, T. M., Wei, X., Song, Y., & Toth, Z. (2008). Ensemble data assimilation with the NCEP Global Forecasting System. *Monthly Weather Review*, *136*, 463–482.

Wicker, L., & Skamarock, W. (2002). Time-splitting methods for elastic models using forward time schemes. *Monthly Weather Review*, *130*, 2088–2097.

Wilks, D. S. (2006). *Statistical Methods in the Atmospheric Sciences*. New York: Academic Press.

Xue, M., Tong, M., & Brewster, K. (2006). An OSSE framework based on the ensemble square root Kalman filter for evaluating the impact of data from radar network on thunderstorm analysis and forecasting. *Journal of Atmospheric and Oceanic Technology*, *23*, 46–66.

Yang, S. C., Kalnay, E., Hunt, B., & Bowler, N. E. (2009). Weights interpolation for efficient data assimilation with the Local Ensemble Transform Kalman Filter. *Quarterly Journal of the Royal Meteorological Society*, *135*, 251–262.

Ying, Y., & Zhang, F. (2015). An adaptive convariance relaxation method for ensemble data assimilation. *Quarterly Journal of the Royal Meteorological Society*, *140*, 2898–2906.

Zängl, G., Reinert, D., Rípodas, P., & Baldauf, M. (2015). The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core. *Quarterly Journal of the Royal Meteorological Society*, *141*, 563–579.

Zeng, Y., Blahak, U., & Jerger, D. (2016). An efficient modular volume-scanning radar forward operator for NWP models: Description and coupling to the COSMO model. *Quarterly Journal of the Royal Meteorological Society*, *142*, 3234–3256.

Zeng, Y., Blahak, U., Neuper, M., & Jerger, D. (2014). Radar beam tracing methods based on atmospheric refractive index. *Journal of Atmospheric and Oceanic Technology*, *31*, 2650–2670.

Zeng, Y., & Janjić, T. (2016). Study of conservation laws with the Local Ensemble Transform Kalman Filter. *Quarterly Journal of the Royal Meteorological Society*, *699*, 2359–2372.

Zeng, Y., Janjić, T., Ruckstuhle, Y., & Verlaan, M. (2017). Ensemble-type Kalman filter algorithm conserving mass, total energy and enstrophy. *Quarterly Journal of the Royal Meteorological Society*, *143*, 2902–2914.

Zhang, F., Bei, N., Rotunno, R., Snyder, C., & Epifanio, C. C. (2007). Mesoscale predictability of moist baroclinic waves: Convection-permitting experiments and multistage error growth dynamics. *Journal of the Atmospheric Sciences*, *64*, 3579–3594.

Zhang, F., Odins, A. M., & Nielsen-Gammon, J. W. (2006). Mesoscale predictability of an extreme warm-season precipitation event. *Weather Forecasting*, *21*, 149–166.

Zhang, F., Snyder, C., & Rotunno, R. (2003). Effects of moist convection on mesoscale predictability. *Journal of the Atmospheric Sciences*, *60*, 1173–1185.

Zhang, F., Snyder, C., & Sun, J. (2004). Impacts of initial estimate and observation availability on convective-scale data assimilation with an ensemble Kalman filter. *Monthly Weather Review*, *132*(5), 1238–1253.