Improving Large-Domain Convection-Allowing Forecasts with High-Resolution Analyses and Ensemble Data Assimilation

CRAIG S. SCHWARTZ
National Center for Atmospheric Research,* Boulder, Colorado

(Manuscript received 20 August 2015, in final form 16 November 2015)

ABSTRACT
Analyses with 20-km horizontal grid spacing were produced from continuously cycling three-dimensional variational (3DVAR), ensemble square root Kalman filter (EnSRF), and “hybrid” variational–ensemble data assimilation (DA) systems over a domain spanning the conterminous United States. These analyses initialized 36-h Weather Research and Forecasting Model forecasts containing a large convection-allowing 4-km nested domain, where downscaled 20-km 3DVAR, EnSRF, and hybrid analyses initialized the 4-km forecasts. Overall, hybrid analyses initialized the best 4-km precipitation forecasts.
Furthermore, whether 4-km precipitation forecasts could be improved by initializing them with true 4-km analyses was assessed. As it was computationally infeasible to produce 4-km continuously cycling ensembles over the large 4-km domain, several “dual-resolution” hybrid DA configurations were adopted where 4-km backgrounds were combined with 20-km ensembles to produce 4-km hybrid analyses. Additionally, 4-km 3DVAR analyses were produced.
In both hybrid and 3DVAR frameworks, initializing 4-km forecasts with true 4-km analyses, rather than downscaled 20-km analyses, yielded superior precipitation forecasts over the first 12 h. Differences between forecasts initialized from 4-km and downscaled 20-km hybrid analyses were smaller for 18–36-h forecasts, but there were occasionally meaningful differences. Continuously cycling the 4-km backgrounds and using static background error covariances with larger horizontal length scales in the hybrid led to better forecasts. All hybrid-initialized forecasts, including those initialized from downscaled 20-km analyses, were more skillful than forecasts initialized from 4-km 3DVAR analyses, suggesting the analysis method was more important than analysis resolution.

1. Introduction
Numerical weather prediction (NWP) models with grid spacing fine enough to obviate the need for convective parameterization have been shown to produce better precipitation forecasts than NWP models with parameterized convection (Done et al. 2004; Kain et al. 2006; Lean et al. 2008; Schwartz et al. 2009; Clark et al. 2010). Traditionally, large-domain convection-allowing forecasts have been initialized by interpolating analyses from convection-parameterizing models onto the high-resolution grid (e.g., Kain et al. 2008; Hohenegger et al. 2008; Kong et al. 2008; Weisman et al. 2008; Hanley et al. 2011; Duc et al. 2013; Romine et al. 2013, 2014; Schumacher and Clark 2014). While this approach has had some success, convection-allowing forecasts may be improved if they are initialized by convection-allowing analysis systems that update a convection-allowing background.
Although any data assimilation (DA) method can produce convection-allowing analyses, ensemble Kalman filters (EnKFs; Evensen 1994; Burgers et al. 1998; Houtekamer and Mitchell 1998) have become popular for convection-allowing applications because they incorporate time-evolving, multivariate forecast error statistics [or background error covariances (BECs)], contrasting the static, time-invariant BECs typically used in three-dimensional variational DA (3DVAR; e.g., Wu et al. 2002; Barker et al. 2004) algorithms. Numerous studies have assimilated radar observations...
with convection-allowing EnKFs for selected severe weather cases (e.g., Snyder and Zhang 2003; Zhang et al. 2004; Dowell at al. 2004; Yussouf et al. 2013 and references therein), and convection-allowing EnKF analyses can initialize better forecasts than convection-allowing 3DVAR analyses (Zhang et al. 2009a; Gao et al. 2013; Johnson et al. 2015). Moreover, Ancell (2012) found that 4-km EnKF analyses initialized better 12-h surface wind and temperature forecasts over the U.S. Pacific Northwest than 36-km EnKF analyses, even though the 4-km EnKF had 40 members and the 36-km EnKF had 80 members. Similarly, Johnson et al. (2015) noted that downscaled 12-km EnKF analyses initialized worse 4-km precipitation forecasts than true 4-km EnKF analyses through ~6 h for 10 cases over the central United States, although their 4-km analyses assimilated radar observations whereas their 12-km analyses did not.

Thus, given the demonstrated benefits of high-resolution ensemble-based DA systems, it is desirable to develop convection-allowing analysis systems over domains large enough to resolve meso-α to synoptic-scale motions to minimize impacts of lateral boundaries (e.g., Warner et al. 1997; Davies 2014). Unfortunately, it is very expensive to produce high-resolution ensemble forecasts necessary for large-domain, convection-allowing EnKF DA. Therefore, current operational convection-allowing DA systems employ either 3DVAR (e.g., Hirahara et al. 2011; Seity et al. 2011; Tang et al. 2013) or nudging techniques (Stephan et al. 2008; Baldauf et al. 2011) over domains no larger than approximately 1500 $\times$ 1500 km², and these analysis systems may be improved by incorporation of ensemble BECs.

However, flow-dependent BECs may be incorporated within large-domain convection-allowing analysis systems by leveraging a “hybrid” variational–ensemble DA approach (e.g., Hamill and Snyder 2000; Lorenc 2003; Buehner 2005; Wang et al. 2008a; Zhang et al. 2009b; Wang 2010; Clayton et al. 2013; Kuhl et al. 2013; Wang and Lei 2014). Hybrid DA systems traditionally produce deterministic analyses within variational frameworks, like 3DVAR, but ingest ensembles to derive flow-dependent BECs, as in EnKFs, and typically initialize better forecasts than DA methods using purely static BECs (e.g., Wang et al. 2008b, 2013; Buehner et al. 2010; Hamill et al. 2011b; Wang 2011; Zhang and Zhang 2012; Zhang et al. 2013; Pan et al. 2014; Schwartz and Liu 2014), including when the analysis has convection-allowing resolution (Li et al. 2012; Gao et al. 2013; Li et al. 2015).

Hybrid systems can be implemented in a “dual-resolution” (DR) configuration, where a single deterministic high-resolution (HR) background is combined with a low-resolution (LR) ensemble providing flow-dependent BECs to produce a HR analysis, eliminating the need for a costly HR ensemble. Given these computational savings, DR hybrid systems have been used in several global DA systems (e.g., Buehner et al. 2010; Hamill et al. 2011b; Kuhl et al. 2013), including operationally in the United States and United Kingdom (e.g., Clayton et al. 2013; Kleist and Ide 2015a,b). Furthermore, Schwartz et al. (2015a) described a DR implementation for a limited-area NWP model where both the HR and LR components had convection-parameterizing resolution.

Therefore, a computationally affordable DR hybrid system can be developed where the deterministic background has convection-allowing resolution while the ensemble resolution is substantially coarser and requires convective parameterization. To examine whether a DR hybrid configuration combining a convection-allowing background and convection-parameterizing ensemble can be successful, and in hopes of improving convection-allowing forecasts by increasing analysis resolution, this study builds on Schwartz and Liu (2014, hereafter SL14), who found that downscaled 20-km hybrid analyses initialized better 4-km precipitation forecasts over the central and eastern regions of the conterminous United States (CONUS) than those initialized by similarly configured 20-km 3DVAR and EnKF DA systems during portions of May and June 2011. First, SL14’s 20-km 3DVAR, EnKF, and SR hybrid DA experiments were repeated, but with a larger sample size spanning May and June 2013, which featured many significant severe weather events over the CONUS (Weisman et al. 2015). These 20-km DA systems initialized 36-h 4-km forecasts, where the 4-km initial conditions (ICs) were provided by downscaled 20-km analyses. These 4-km forecasts were then compared to those initialized by various 4-km hybrid DA systems, most of which employed DR configurations that combined 4-km deterministic backgrounds with 20-km ensembles to produce 4-km analyses. Forecasts initialized from 4-km 3DVAR systems were also examined.

The next section introduces the NWP model and DA configurations, and the assimilated observations are detailed in section 3. The experimental design is described in section 4. Section 5 examines the quality of the ensembles ingested by hybrid analyses and section 6 presents verification of the 4-km forecasts, with a focus on precipitation. A discussion and conclusion are provided in section 7.
2. Model and data assimilation configurations

a. Forecast model

The NWP model and DA settings were similar to SL14’s. Specifically, all weather forecasts were produced by version 3.3.1 of the nonhydrostatic Advanced Research version of the Weather Research and Forecasting (ARW; Skamarock et al. 2008) model over a nested computational domain centered over the CONUS (Fig. 1). The horizontal grid spacing was 20 km (300 × 200 grid points) in the outer domain and 4 km (801 × 616 grid points) in the inner nest. Both domains had 57 vertical levels and a 10-hPa top. The time step was 80 s in the 20-km domain and 16 s in the 4-km nest. Global Forecast System (GFS) forecasts provided lateral boundary conditions (LBCs) for the 20-km domain every 3 h and the 20-km domain provided LBCs for the 4-km nest every 20-km time step (80 s).

SL14 employed Morrison microphysics (Morrison et al. 2009), but further testing revealed that high precipitation biases, particularly during the first few hours of model integration, could be somewhat ameliorated by using Thompson microphysics (Thompson et al. 2008). Thus, this work used Thompson microphysics along with a full suite of other physical parameterizations (Table 1). The same physics and positive definite moisture advection (Skamarock and Weisman 2009) were used on both domains, except cumulus parameterization was turned off on the 4-km grid.

b. Data assimilation systems

As in SL14, the hybrid and 3DVAR algorithms from the National Centers for Environmental Prediction’s (NCEP) operational Gridpoint Statistical Interpolation analysis system (GSI; Kleist et al. 2009) were used. GSI’s 3DVAR formulation is described by Wu et al. (2002), while GSI’s hybrid algorithm employs extended control variables (Lorenc 2003) and is detailed by Wang (2010).

The GSI-hybrid system uses an ensemble of short-term forecasts to incorporate flow-dependent BECs within a variational cost function and it was necessary to update the ensemble when new observations were available. As in SL14, an ensemble square root Kalman filter (EnSRF; Whitaker and Hamill 2002) was used to update 50-member ARW model prior (before assimilation) ensembles. To reduce spurious correlations due to sampling error, covariance localization was applied following SL14: EnSRF analysis increments were forced to zero 1280 km from an observation in the horizontal and 1 scale height (in log pressure coordinates) in the vertical using a Gaspari and Cohn (1999) polynomial piecewise function. Further, as in SL14, multiplicative inflation was applied to posterior (after assimilation) perturbations about ensemble mean analyses following Whitaker and Hamill (2012)’s “relaxation-to-prior spread” approach with an inflation parameter \( \alpha = 1.12 \) [see Whitaker and Hamill (2012), SL14, and Harnisch and Keil (2015) for...
more details about the inflation algorithm]. SL14 also used $\alpha = 1.12$ and described the rationale for this choice, which helped engender proper ensemble spread for the assumed observation errors (section 5).

The GSI-hybrid, like most hybrid algorithms, contains adjustable parameters that determine how much the total BECs are weighted toward static (e.g., 3DVAR) and ensemble contributions. The static BECs were constructed via the “NMC method” (Parrish and Derber 1992) by taking differences between forecasts of different lengths valid at common times. Following SL14, differences of 24- and 12-h ARW model forecasts valid at 120 common times over May and June 2010 were used to compute static BECs for both the 20- and 4-km domains using the model configurations described in section 2a. The initially computed static BECs were empirically tuned as described in SL14, and these tunings were fixed across all experiments.

For all hybrid experiments, the BECs were weighted 75% toward the ensemble contribution and 25% toward the static (i.e., NMC generated) component, following SL14. These weightings are also used operationally at NCEP (Kleist and Ide 2015a) for the GFS DA system.

Horizontal and vertical localization were applied in the hybrid to limit the spatial extent of the ensemble contribution to the analysis increments, and the length scales were identical to those in the EnSRF. As in Wang et al. (2013), Wang and Lei (2014), and SL14, two outer loops (Courtier et al. 1994) were employed in GSI’s variational minimization.

3. Observations

All experiments assimilated a variety of conventional meteorological observations as described in SL14 that were assumed valid at the analysis time (Table 2). Observations taken within $\pm 1.5$ h of each analysis were assimilated (i.e., a 3-h total time window), except the time window was shortened to $\pm 0.5$ h for METAR and SYNOP surface pressure observations. The initial observation error standard deviations ($\sigma_o$) were similar to SL14’s, except the specific humidity observation error was approximately doubled to improve ensemble spread–skill relationships (section 5). SL14 discussed how GSI alters initially specified observation errors to produce a “final” observation error ($\sigma_o^*$) actually used in the assimilation (Developmental

<table>
<thead>
<tr>
<th>Physical parameterization</th>
<th>ARW model option</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microphysics</td>
<td>Thompson</td>
<td>Thompson et al. (2008)</td>
</tr>
<tr>
<td>Longwave and shortwave radiation</td>
<td>Rapid Radiative Transfer Model for Global Climate Models (RRTMG) with ozone and aerosol climatologies</td>
<td>Mlawer et al. (1997); Iacono et al. (2008); Tegen et al. (1997)</td>
</tr>
<tr>
<td>Land surface model</td>
<td>Noah</td>
<td>Chen and Dudhia (2001)</td>
</tr>
<tr>
<td>Cumulus parameterization</td>
<td>Tiedtke</td>
<td>Tiedtke (1989); Zhang et al. (2011)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Observing platform</th>
<th>Observation type</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radiosonde</td>
<td>Surface pressure, Temperature, Specific humidity, Wind</td>
<td>Thinned to 40 km and 40 hPa</td>
</tr>
<tr>
<td>Aircraft</td>
<td>Temperature, Wind</td>
<td>Thinned to 40 km and 40 hPa</td>
</tr>
<tr>
<td>Wind profiler</td>
<td>Temperature, Wind, Refractivity, Wind</td>
<td>Thinned to 40 km and 40 hPa</td>
</tr>
<tr>
<td>Global positioning system radio occultation (GPSRO)</td>
<td>Surface pressure, Temperature, Specific humidity, Wind</td>
<td>Observations above 200 hPa and over land rejected</td>
</tr>
<tr>
<td>Infrared and water vapor channel satellite-tracked wind</td>
<td>Surface pressure, Temperature, Specific humidity, Wind</td>
<td></td>
</tr>
<tr>
<td>Ship and buoy</td>
<td>Surface pressure, Temperature, Specific humidity, Wind</td>
<td></td>
</tr>
<tr>
<td>SYNOP and METAR</td>
<td>Surface pressure</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Physical parameterizations used in all ARW model forecasts. These parameterizations were used on both the 20- and 4-km grids, except no cumulus scheme was used on the 4-km grid.

Table 2. Observations assimilated by all experiments.
Testbed Center 2014), and an identical procedure was used here. A typical distribution of conventional observations available for assimilation at 0000 UTC is shown in Fig. 2.

Observations were quality controlled and preprocessed as described by SL14 and noted in Table 2. Furthermore, an “outlier check” was applied independently for each experiment, whereby an observation was rejected if its innovation (difference between the observation and model-simulated observation based on the model guess) exceeded $a o_r$. For the EnSRF, this outlier check was based on the ensemble mean prior. Values of $a$ varied between 4 and 10 depending on the observation type and were identical to those in SL14’s Table 1, which were loosely based on default GSI settings for the GFS DA system.

All observation processing was performed by GSI, and for the EnSRF, GSI generated prior model-simulated observations for each ensemble member, assigned observation errors, and performed observation thinning and quality control decisions [as in Hamill et al. (2011a), SL14, Pan et al. (2014), Wang and Lei (2014), Zhu et al. (2013), and Kleist and Ide (2015a,b)]. The observations, prior model-simulated observations, and final observation errors were then ingested into the EnSRF. This complete dependence on GSI for all observation-related issues ensured consistency of the forward operators and observation errors in all DA systems.

4. Experimental design

Many experiments employing hybrid DA systems were performed, where hybrid analyses incorporated flow-dependent BECs provided by prior EnSRF ensembles. In the continuously cycling configurations, each EnSRF (hybrid) analysis initialized a 6-h ensemble (deterministic) forecast that served as the background for the next EnSRF (hybrid) analysis. EnSRF analyses always had 20-km horizontal grid spacing, while hybrid analyses had either 4- or 20-km horizontal grid spacing, depending on the experiment.

In a coupled EnKF–hybrid system, EnKF analysis ensembles can optionally be recentered about hybrid analyses, which shifts the ensemble mean but preserves perturbations about the mean (e.g., Zhang et al. 2013; Schwartz et al. 2015a). However, as noted by Kleist and
Ide (2015a), several studies (e.g., Clayton et al. 2013; Wang et al. 2013; Pan et al. 2014; Schwartz et al. 2015a) revealed little sensitivity to whether this recentering was performed, which suggests that a single ensemble can safely provide flow-dependent BECs for multiple hybrid experiments (Clayton et al. 2013). Thus, based on these collective previous results, EnSRF analyses were not recentered about hybrid analyses, and unless otherwise stated, the same 20-km ensembles provided BECs for all hybrid configurations, which permitted a large number of hybrid experiments to be performed that would have otherwise been impossible (given finite computing resources) had it been necessary to produce new EnSRF analyses for each hybrid experiment.

a. EnSRF experimental design

The initial ensemble was produced by interpolating the 0000 UTC 4 May 2013 deterministic 0.5° × 0.5° GFS analysis onto the 20-km computational domain and adding Gaussian random draws with zero mean and static BECs (Barker 2005; Torn et al. 2006) to the wind, surface pressure, temperature, water vapor, and geopotential height fields. The randomly generated ensemble member, and sea surface temperature (SST) was updated each 0000 UTC analysis from 1/12° NCEP SST analyses (http://polar.ncep.noaa.gov/sst/rtg_high_res), as in SL14.

From 7 May through 30 June, 0000 UTC ensemble mean EnSRF analyses initialized 36-h ARW forecasts on both the 20- and 4-km domains (55 forecasts). The 4-km nest was initialized by downsampling 20-km ensemble mean analyses using a monotone interpolation scheme (Smolarkiewicz and Grell 1992; Skamarock et al. 2008). Thus, the EnSRF was utilized for two purposes: to provide perturbations for hybrid analyses and as a standalone DA system to initialize ARW forecasts.

Although 36-h EnSRF forecasts were initialized from ensemble mean analyses, which are smooth compared to individual members’ and hybrid analyses, Schwartz et al. (2014) showed that forecasts initialized from EnKF mean analyses had comparable or better skill than forecasts initialized from individual members’ analyses. Thus, it is unlikely that the relative performance of 36-h EnSRF- and hybrid-initialized forecasts was due to issues related to effective resolution of EnSRF mean analyses.

b. Continuously cycling hybrid experiments

Several continuously cycling experiments employing hybrid DA were performed, where the various experimental configurations differed slightly to study particular sensitivities. For all continuously cycling hybrid experiments, as with the EnSRF, the background for the first hybrid analysis was the 0000 UTC 4 May 2013 GFS analysis interpolated onto the computational domain. Procedures related to cycling of hydrometeors and soil states, SST updates, cycling interval (i.e., 6 h), and beginning and ending dates of the cycling period were identical to those described for the EnSRF (section 4a) and the same for all continuously cycling hybrid experiments.

Both “single-resolution” (SR) and DR continuously cycling hybrid configurations were employed. In the SR configuration, deterministic hybrid backgrounds and prior ensembles had identical resolutions, while in DR systems, the ensembles had coarser resolution than the backgrounds. The SR and DR experiments are now described.

1) SINGLE-RESOLUTION HYBRID EXPERIMENT

As in SL14, continuously cycling SR hybrid analyses were performed where 20-km deterministic backgrounds were combined with 20-km ensemble BECs and observations to produce 20-km hybrid analyses (“Hyb_SR_20km”). The static BECs were provided via the NMC method operating on 20-km ARW model forecasts (section 2b). As with the EnSRF, beginning 7 May, each 0000 UTC analysis initialized 36-h ARW model forecasts on both the 20- and 4-km domains. As only 20-km hybrid analyses were produced, down-scaled 20-km hybrid analyses initialized the 4-km nest.

2) DUAL-RESOLUTION HYBRID EXPERIMENTS

To investigate whether precipitation forecasts could be improved by increasing analysis resolution, two experiments were performed where continuously cycling 4-km hybrid analyses were produced along with 20-km analyses. Since it was computationally infeasible to produce 4-km continuously cycling ensembles over the 4-km computational domain (Fig. 1), a DR configuration was adopted that combined deterministic 4-km
hybrid backgrounds with 20-km ensemble BECs provided by the cycling EnSRF (section 4a) to produce 4-km hybrid analyses. The 4-km DR analyses only assimilated conventional observations (section 3) located over the 4-km domain. To permit a fair comparison with forecasts initialized from downscaled 20-km continuously cycling hybrid analyses, 4-km analyses did not assimilate radar observations.

All continuously cycling DR hybrid experiments produced separate, independent 20- and 4-km analyses each DA cycle, where the same 20-km ensembles provided flow-dependent BECs for both 20- and 4-km hybrid analyses (Fig. 3). Then, to advance the ARW model between hybrid analyses, 6-h one-way nested ARW model forecasts were produced with the 4-km nest embedded within the 20-km domain. One-way nesting was used during model advances between DA cycles to maintain independence of the 20- and 4-km analysis systems (i.e., the continuously cycling DR hybrid systems produced identical 20-km analyses as Hyb_SR_20km; Table 3).

Production of 4-km hybrid analyses meant choosing whether to use static BECs produced via the NMC method operating on 20- or 4-km ARW model forecasts (section 2b). Some differences were noted regarding the 4- and 20-km static BEC vertical length scales, and, as expected, 4-km static BEC standard deviations were larger than 20-km static BEC standard deviations at most locations (not shown). However, differences between the 4- and 20-km static BEC standard deviations and vertical length scales were small compared to those of the static BEC horizontal length scales [as defined by Wu et al. (2002)], as the 4-km static BEC horizontal length scales were substantially smaller than the 20-km static BEC horizontal length scales (Fig. 4). Horizontal BEC length scales computed directly from 20-km 0000 UTC EnSRF prior ensemble perturbations\(^1\) were closer to the 20-km static BEC horizontal length scales than 4-km static BEC horizontal length scales (Fig. 4).

\(^1\) The NMC method can be deconstructed as two broad steps. In the first, differences between deterministic forecasts of different lengths (e.g., 24- and 12-h forecasts) valid at common times are computed. In the second step, these differences, or, perturbations, are used as input to algorithms that compute BEC length scales and variances. Alternatively, the first step can be replaced by computing perturbations about an ensemble mean, which was performed for the prior 20-km 0000 UTC ensembles. These ensemble-based perturbations were then input to the second step.
<table>
<thead>
<tr>
<th>Experiment name</th>
<th>20-km analysis</th>
<th>4-km analysis</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyb_SR_20km</td>
<td>Control hybrid experiment that produced continuously cycling 20-km hybrid analyses</td>
<td>None</td>
<td>Hybrid control</td>
</tr>
<tr>
<td>Hyb_SR_4km_Cool_start</td>
<td>Identical to Hyb_SR_20km</td>
<td>SR 4-km hybrid analyses where 4-km backgrounds were spun up over 6-h forecasts initialized from 1800 UTC Hyb_SR_20km analyses, 4-km ensembles were spun up over 6-h forecasts initialized from 1800 UTC EnSRF analysis ensembles, and static BECs were derived from 4-km forecasts</td>
<td>Compare to Hyb_DR_Cool_start_4km_static to assess sensitivity to ensemble perturbation resolution</td>
</tr>
<tr>
<td>Hyb_DR_Cool_start_4km_static</td>
<td>Identical to Hyb_SR_20km</td>
<td>DR 4-km hybrid analyses incorporating 20-km BECs from the EnSRF, where 4-km backgrounds were spun up over 6-h forecasts initialized from 1800 UTC Hyb_SR_20km analyses and static BECs were derived from 4-km forecasts</td>
<td>Compare to Hyb_DR_Cool_start_20km_static to assess sensitivity to static BECs</td>
</tr>
<tr>
<td>Hyb_DR_Cool_start_20km_static</td>
<td>Identical to Hyb_SR_20km</td>
<td>Identical to Hyb_DR_Cool_start_4km_static except static BECs were derived from 20-km forecasts</td>
<td>Compare to Hyb_DR_Cool_start_20km_static to assess sensitivity to 4-km background source (continuously cycled or “spun up”)</td>
</tr>
<tr>
<td>Hyb_DR_4km_static</td>
<td>Identical to Hyb_SR_20km</td>
<td>Identical to Hyb_DR_4km_static, except using static BECs derived from 20-km forecasts</td>
<td>Compare to Hyb_DR_20km static to assess sensitivity to static BECs</td>
</tr>
<tr>
<td>Hyb_DR_20km_static</td>
<td>Identical to Hyb_SR_20km</td>
<td>Identical to Hyb_DR_4km_static, except using static BECs derived from 20-km forecasts</td>
<td>Compare to Hyb_SR_20km to assess utility of performing high-resolution analyses</td>
</tr>
<tr>
<td>3DVAR_20km</td>
<td>Control 3DVAR experiment that produced continuously cycling 20-km 3DVAR analyses</td>
<td>None</td>
<td>3DVAR control</td>
</tr>
<tr>
<td>3DVAR_4km_BEC20</td>
<td>Identical to 3DVAR_20km</td>
<td>4-km continuously cycling 3DVAR analyses with BECs derived from 20-km forecasts</td>
<td>Compare to 3DVAR_20km to assess utility of performing high-resolution analyses</td>
</tr>
<tr>
<td>3DVAR_4km_BEC4</td>
<td>Identical to 3DVAR_20km</td>
<td>Identical to 3DVAR_4km_BEC20 except BECs derived from 4-km forecasts</td>
<td>Compare to 3DVAR_4km_BEC20 to assess sensitivity to static BECs</td>
</tr>
</tbody>
</table>
While reduction of static BEC horizontal length scales with increased resolution was expected (Wu et al. 2002), the different static BEC length scales could engender large analysis impacts. Given uncertainty about which static BECs would yield the best results, continuously cycling 4-km DR hybrid analyses were performed that differed solely in the specification of the static BECs (see descriptions of “Hyb_DR_4km_static” and “Hyb_DR_20km_static” in Table 3). As static BECs can be tuned, relative performances of the experiments differing solely by static BECs should not be interpreted narrowly as a comparison of the inherent goodness of 4- versus 20-km static BECs, but, rather, as assessing general sensitivity to different static BEC options that primarily differ by their horizontal length scales.

Beginning 0000 UTC 7 May, continuously cycling DR hybrid analyses initialized 36-h ARW model forecasts on both the 20- and 4-km domains. For these DR experiments, true 4-km analyses always initialized the 4-km nest, contrasting the SR 20-km hybrid experiment (i.e., Hyb_SR_20km), where ICs for 4-km forecasts were downscaled 20-km analyses. Moreover, 4-km forecast differences between Hyb_SR_20km and Hyb_DR_20km_static can be attributed entirely to the presence, or absence, of 4-km DA and subsequent initialization of the 4-km ARW model domain.

**Fig. 4.** Background error covariance horizontal length scales (km) derived from the NMC method applied to 24 – 12 h deterministic 20-km forecasts (solid lines) and 4-km forecasts (short dashed lines) and 0000 UTC prior ensembles (long dashed lines) for (a) streamfunction, (b) unbalanced velocity potential, (c) unbalanced virtual temperature, and (d) pseudo–relative humidity. These statistics were averaged over the geographic area encompassed by the 4-km domain (Fig. 1).
c. “Cool-start” hybrid experiments with spunup high-resolution backgrounds

Although it was unaffordable to produce continuously cycling 4-km ensembles over the 4-km domain, it was desirable to examine whether 4-km hybrid analyses ingesting 20-km BECs were disadvantaged by the relatively coarse ensemble. Furthermore, as 36-h forecasts were only initialized at 0000 UTC, could HR analyses be produced just once daily at 0000 UTC to eliminate the need for continuously cycling HR DA?

To examine both the necessity of HR continuous cycling and sensitivity of ensemble perturbation resolution, a pragmatic “cool-start” approach was employed to produce 4-km ensembles and deterministic backgrounds for DA purposes. Specifically, beginning 6 May 2013, each 1800 UTC 20-km EnSRF analysis (i.e., section 4a) initialized a 50-member ensemble of 6-h one-way nested ARW model forecasts containing the 4-km nest to “spin up” 4-km ensembles valid at 0000 UTC. The 4-km ensemble ICs were simply downscaled 20-km 1800 UTC EnSRF analysis ensembles. Similarly, 1800 UTC Hyb_SR_20km analyses initialized 6-h one-way nested deterministic ARW model forecasts on both the 4- and 20-km grids, where the 4-km ICs were downscaled 1800 UTC Hyb_SR_20km analyses. Then, at 0000 UTC, SR 4-km hybrid analyses were produced where spunup 4-km deterministic backgrounds were combined with spunup 4-km prior ensembles and 4-km static BECs (“Hyb_SR_4km_Cool_start”; Table 3). Additionally, two sets of DR analyses were produced that combined spunup 0000 UTC 4-km deterministic backgrounds with 20-km ensembles produced by the continuously cycling EnSRF (section 4a) and solely differed regarding the static BECs (“Hyb_DR_Cool_start_4km_static” and “Hyb_DR_Cool_start_20km_static”; see Table 3).

These three cool-start experiments’ 0000 UTC analyses also initialized 36-h ARW model forecasts on both the 4- and 20-km grids, where the 4-km ICs were true 4-km analyses. Because all cool-start experiments used identical deterministic 4-km backgrounds (i.e., 4-km fields spun up over 6-h from 20-km ICs), analysis and forecast differences between Hyb_SR_4km_Cool_start and Hyb_DR_Cool_start_4km_static can be attributed entirely to the resolution of the ensemble BECs (4 or 20 km). Additionally, experiments Hyb_DR_Cool_start_20km_static and Hyb_DR_20km_static differed solely regarding the source of the 4-km background (i.e., continuously cycled or spun up), permitting an assessment of the need to continuously cycle HR deterministic backgrounds.

d. Continuously cycling 3DVAR experiments

1) 20-KM 3DVAR SYSTEM

Another experiment configured exactly as Hyb_SR_20km was performed (“3DVAR_20km”), except pure 3DVAR analyses were produced (i.e., BECs were 100% static and generated via the NMC method described in section 2b). Beginning 7 May, each 0000 UTC 3DVAR analysis initialized 36-h ARW model forecasts on both the 20- and 4-km domains, where 4-km ICs were provided by downscaled 20-km 3DVAR analyses (analogously to Hyb_SR_20km).

2) 4-KM 3DVAR SYSTEMS

To again assess the impact of producing high-resolution analyses and sensitivity to static BECs, two experiments configured exactly as Hyb_DR_20km_static and Hyb_DR_4km_static were performed, except 3DVAR analyses were produced (“3DVAR_4km_BEC20” and “3DVAR_4km_BEC4”; Table 3). Similar to Hyb_DR_20km_static and Hyb_DR_4km_static, these 3DVAR systems differed solely by their static BECs, produced independent 4- and 20-km analyses each DA cycle, and initialized 36-h ARW model forecasts on both the 20- and 4-km domains from 0000 UTC analyses, where true 4-km 3DVAR analyses initialized the 4-km nest.

e. GFS-initialized forecasts

Finally, operational 0000 UTC GFS analyses were interpolated onto the ARW model domains and initialized 36-h ARW model forecasts on both the 4- and 20-km grids between 7 May and 30 June to serve as a benchmark for the limited-area DA experiments. During this period (2013), the GFS analysis system produced continuously cycling global DR hybrid analyses at ~0.2° horizontal grid spacing (Wang et al. 2013) and assimilated many additional observations, including satellite radiances, that were not assimilated by any of the DA experiments described above. As described by Schwartz et al. (2015b), following standard ARW model pre-processing procedures, GFS hydrometeor analyses were discarded, such that GFS-initialized ARW model forecasts began from empty (i.e., zero value) hydrometeor fields, in contrast to all other experiments, where non-zero hydrometeor ICs were present.

5. Prior 20-km ensemble characteristics

High-quality prior ensembles are necessary for successful hybrid analyses. In a well-calibrated EnKF analysis/forecast system, when compared to observations, the
prior ensemble mean root-mean-square error (RMSE) will equal the prior “total spread,” defined as the square root of the sum of the observation error variance and prior ensemble variance of the simulated observations (Houtekamer et al. 2005). Thus, the ratio of the total spread to RMSE [called the consistency ratio (CR; Dowell and Wicker 2009)] should be near 1.0, with a CR = 1.0 meaning insufficient spread.

The CR, ensemble mean RMSE, and ensemble mean additive bias, aggregated over all 20-km prior ensembles (6-h forecasts) every 6 h between 0000 UTC 7 May and 0000 UTC 30 June, are shown in Fig. 5. For aircraft winds (Figs. 5a,b), CRs were usually near 1.0, but for aircraft temperature (Fig. 5c), CRs were near 1.1, indicating slightly too much spread. Absolute values of aircraft temperature (wind) biases were less than 0.2 K (m s⁻¹) except for wind at some levels at or below 850 hPa. Radiosonde wind CRs were usually near 1.0, and absolute values of biases were typically <0.2 m s⁻¹, except for 200-hPa zonal wind, near jet stream level (Fig. 5d). For radiosonde temperature, there was too much (little) spread above (below) 700 hPa and warm biases for most levels above 700 hPa (Fig. 5e). This warm bias with respect to radiosondes has also been noted in other continuously cycled EnKF DA systems over the CONUS (e.g., Romine et al. 2013; Schumacher and Clark 2014; SL14). CRs and biases for moisture observations at and below 700 hPa were near 1.0 and <0.2 g kg⁻¹, respectively (Fig. 5f), with CRs > 1.0 at and above 500 hPa, suggesting radiosonde specific humidity observation errors could be decreased above 500 hPa.

These statistics indicate the EnSRF was reasonably well tuned and usually had small biases. Furthermore, CRs typically near 1.0 suggest the inflation parameter (α) was acceptable for the assumed observation errors. Thus, the prior EnSRF ensembles appeared appropriate for use within hybrid analyses.

6. Verification of 4-km forecasts

a. Precipitation verification methodology

Forecasts of hourly accumulated precipitation were directly compared to gridded stage-IV (ST4) observations produced at NCEP (Lin and Mitchell 2005) by
interpolating rainfall fields to the ST4 grid (~4.7-km grid spacing; referred to here as the verification grid) using a budget interpolation approach (e.g., Accadia et al. 2003). As in SL14, objective verification was performed over a fixed domain encompassing most of the central CONUS (Fig. 1). Assessments of skill focused on the first 12 forecast hours and 18–36-h forecasts to evaluate initial representation of convection and next-day convective initiation and evolution, respectively.

The fractions skill score (FSS; Roberts and Lean 2008), a “neighborhood approach” commonly employed to assess convection-allowing NWP model performance, was used to evaluate spatial skill of precipitation forecasts. To compute the FSS, a threshold \( q \) is chosen, either as an absolute value (e.g., 1.0 mm h\(^{-1}\)) or percentile (e.g., top 0.5% of the precipitation amounts), and a radius of influence \( r \) (e.g., \( r = 50 \) km) is selected. Then, a fractional value for the \( i \)th grid point is determined by dividing the number of points within \( r \) km of \( i \) containing accumulated precipitation \( \geq q \) by the total number of points within \( r \) km of \( i \). This procedure is performed for both the observations (e.g., ST4 analyses) and forecasts (after the model data have been interpolated to the ST4 grid) to produce observed \( (o_i) \) and forecast \( (f_i) \) fractions (e.g., Theis et al. 2005). Then, the FSS for a single forecast is computed by comparing the fractions Brier score (FBS) to the worst possible FBS (FBS\(_{\text{worst}}\)) at each of \( N_v \) verification points:

\[
FSS = 1 - \frac{\text{FBS}}{\text{FBS}_{\text{worst}}} = 1 - \frac{\sum_{i=1}^{N_v} (f_i - o_i)^2}{\sum_{i=1}^{N_v} (f_i^2 + o_i^2)}.
\]

The FSS ranges from 0 to 1, with a perfect forecast attaining a score of 1 and FSS = 0 indicating no skill. Aggregate FSSs over a \( p \)-h period and all 55 forecasts were obtained by summing over \( i = 1, \ldots, N_v \times 55 \times 5 \) grid points in Eq. (1).

As in SL14, FSSs were calculated for \( r = 5, 25, 50, 75, 100, 125, \) and 150 km and computed with percentile thresholds. Percentiles were computed separately for each forecast hour for all experiments and the ST4 observations based on all rainfall amounts, including zeros, in the verification domain on the verification grid over all 55 forecasts. This method effectively removed bias between the forecasts and observations and permitted a robust assessment of spatial skill (e.g., Roberts and Lean 2008; Mittermaier and Roberts 2010; Dey et al. 2014). These hourly varying percentiles are shown in Fig. 6 for selected experiments to illustrate the absolute thresholds corresponding to the percentiles.

In general, after \( \sim 18 \) h, all experiments (including those not shown in Fig. 6) had similar percentiles, but larger differences related to spinup occurred at earlier times and are discussed further in section 6d.

Additionally, after using the neighborhood approach to transform each deterministic forecast into a fractional field, the \( f \) values were used to compute areas under the relative operating characteristic (ROC) curve (Mason and Graham 2002). The ROC areas provided similar conclusions as FSSs and are not discussed further.

b. Statistical significance testing

Statistical significance was assessed by a bootstrap resampling technique (e.g., Hamill 1999) based on differences between pairs of experiments’ forecasts using 1000 iterations (see SL14 for more specifics). The forecasts, which were only initialized at 0000 UTC (i.e., 24 h apart), were assumed to be uncorrelated, following Hamill (1999). As in Davis et al. (2010), the significance level for differences between two experiments was interpreted as the percentile where the distribution of resampled differences crossed zero, with significance levels \( \geq 90\% \) considered meaningful. This approach was applied to compute bootstrap confidence intervals for the FSS, additive bias, and RMSE. Even though there was some serial correlation between adjacent hours of a single, particular forecast, bootstrapping was nevertheless applied to FSSs aggregated over \( p \)-h periods. A block-bootstrapping methodology, which attempts to account for serial correlation (e.g., Wilks 1997), was also used to calculate significance levels for FSSs aggregated over \( p \)-h periods but did not yield different interpretations regarding statistical significance.

c. Results for 4-km forecasts initialized from downscaled 20-km analyses

FSSs aggregated hourly over the first 12-h and all 55 4-km forecasts were highest for hybrid-initialized forecasts (Fig. 7), followed by those initialized from EnSRF analyses. Both the hybrid- and EnSRF-initialized rainfall forecasts were better than GFS- and 3DVAR-initialized precipitation forecasts, which were similar. All FSS differences between experiments were statistically significant (SS) at the 98th percentile except those between the 3DVAR- and GFS-initialized forecasts, which were usually not SS at the 90% level (not shown). As GFS-initialized forecasts had no initial hydrometeors (section 4e), they were disadvantaged over early forecast hours and clearly had the lowest 1–12-h FSSs when they were computed using absolute thresholds (not
shown). These 1–12-h forecast results were nearly identical to SL14’s.

For 18–36-h forecasts (Fig. 8), the GFS-, hybrid-, and EnSRF-initialized 4-km precipitation forecasts were statistically significantly better at the 95th percentile compared to 3DVAR-initialized forecasts for all \( q \) and \( r \). Hybrid- and EnSRF-initialized forecasts were also usually statistically significantly better than GFS-initialized forecasts at the 90% level. Although hybrid-initialized forecasts typically had higher FSSs than EnSRF-initialized forecasts and differences were SS at the 95% level for the 97th and 98th percentile thresholds, significance levels were <90% for the 99th percentile and above. These findings differ somewhat from SL14, who found that 18–36-h hybrid-initialized forecasts were statistically significantly better at the 95% level compared to EnSRF-initialized forecasts at all thresholds, suggesting that static BECs may have been less important in this period (2013) than SL14’s period (2011). SL14 also found that GFS-initialized forecasts were usually poorest, but often statistically indistinguishable from 3DVAR-initialized forecasts. However, during this period, the GFS was initialized with hybrid DA, whereas during SL14’s period, the GFS was initialized with 3DVAR DA. Both here and in SL14, the analysis systems not incorporating flow-dependent BECs initialized the worst 18–36-h precipitation forecasts, strongly demonstrating the merits of DA systems incorporating flow-dependent BECs.

Fig. 6. Absolute thresholds (mm h\(^{-1}\)) corresponding to the (a) 97th, (b) 98th, (c) 99th, (d) 99.25th, (e) 99.5th, and (f) 99.75th percentile thresholds computed from a climatological perspective each forecast hour.
d. Results for 4-km forecasts initialized from true 4-km analyses

1) PRECIPITATION VERIFICATION

The 4-km precipitation forecasts initialized from continuously cycling 4-km DR hybrid analyses (both Hyb_DR_4km_static and Hyb_DR_20km_static) had higher aggregate 1–12-h FSSs than forecasts initialized from Hyb_SR_20km for all thresholds and radii, with significance levels compared to Hyb_SR_20km FSSs regularly exceeding 99% (Figs. 9 and 10). Differences between Hyb_DR_4km_static and Hyb_DR_20km_static with Hyb_SR_20km generally increased slightly with $r$. Consistent with the hybrid results, 1–12-h rainfall forecasts were also better when true 4 km, rather than downscaled 20-km 3DVAR analyses, served as 4-km ICs (Fig. 9). However, all hybrid-initialized forecasts, including cool-start-initialized forecasts (not shown in Fig. 9) and forecasts initialized from Hyb_SR_20km, were better than all 3DVAR-initialized forecast sets. That downscaled 20-km hybrid analyses (i.e., Hyb_SR_20km) initialized better 4-km precipitation forecasts than true 4-km 3DVAR analyses echoes Johnson et al. (2015)’s findings and indicates that incorporation of flow-dependent BECs was more important than analysis resolution.

The 1–12-h precipitation forecasts were better when 4-km DR hybrid and 3DVAR systems used 20-km, rather than 4-km, static BECs (Figs. 9 and 10), with smaller differences between the hybrid experiments, since static BECs only contributed 25% to hybrid analyses. Moreover, while 1–12-h 4-km rainfall forecasts initialized from cool-start 4-km hybrid analyses were better than those initialized from Hyb_SR_20km, they were not statistically significantly better...

![Fig. 7](image-url)
at the 90% level as regularly as 4-km forecasts initialized from the two continuously cycling DR hybrid systems (Fig. 10). Within the cool-start framework, incorporation of 4-km ensembles in the hybrid sometimes led to slight degradation (cf. Hyb_DR_Cool_start_4km_static and Hyb_SR_4km_Cool_start in Fig. 10).

Within the first 12 h, the largest improvement of the continuously cycling DR hybrid-initialized forecasts compared to forecasts initialized from Hyb_SR_20km occurred over the first ~6 h (Fig. 11). Interestingly, when 20-km static BECs were used in the 4-km 3DVAR DA system, FSS differences between forecasts initialized from 4-km and downscaled 20-km 3DVAR analyses persisted to ~12 h, longer than in the corresponding hybrid systems, possibly because the 3DVAR analyses were poorer and analysis resolution assumed more importance (Fig. 11). Although 3DVAR_4km_BEC20 had higher FSSs than Hyb_SR_20km for the first 1–2 h, 3DVAR_4km_BEC20 FSSs fell below Hyb_SR_20km FSSs by ~6 h, indicating any benefit of HR 3DVAR DA relative to coarser-resolution hybrid DA was quickly lost (Fig. 11).

The higher FSSs for the continuously cycling DR hybrid experiments over the first ~6 h compared to Hyb_SR_20km were consistent with improved biases compared to Hyb_SR_20km over this period, as illustrated by Fig. 12, which shows fractional occurrence of precipitation exceeding certain thresholds ($q$) over the verification domain, calculated on the verification grid, and aggregated hourly over all 55 forecasts. At all thresholds, forecasts initialized from downscaled 20-km hybrid analyses (purple line in Fig. 12) took at least 2 h to reach maximum coverages and grossly overpredicted coverages at 2 h for $q \geq 5.0 \text{ mm h}^{-1}$ (this behavior was also noted and discussed by SL14). This

---

**FIG. 8.** As in Fig. 7, but FSSs aggregated hourly for 18–36-h forecasts.

2 When digital filter initialization (e.g., Huang and Lynch 1993) was applied to forecasts initialized from Hyb_SR_20km, the spurious peak at 2 h for thresholds $\geq 5.0 \text{ mm h}^{-1}$ was removed but the spinup was noticeably slower and not as desirable as those from the continuously cycling DR systems.
overshoot was ameliorated some by initializing 4-km forecasts from cool-start 4-km DR hybrid analyses (blue lines), although 1–2-h coverages were always higher than those observed. Coverages from forecasts initialized with continuously cycling 4-km DR hybrid analyses (green lines) usually agreed best with observations at 1 h for all thresholds and for $q \geq 5.0$ mm h$^{-1}$ were closest to ST4 coverages through at least $\sim 6$ h (Figs. 12d–f). Yet, the continuously cycling DR coverages became relatively small compared to ST4 observations at $\sim 3$ h for $q \leq 1.0$ mm h$^{-1}$ (Figs. 12a–c), suggesting difficulty maintaining areas of light rainfall after initialization. The 3DVAR forecasts initialized from downscaled 20-km and true 4-km analyses displayed qualitatively similar areal coverage characteristics as the corresponding hybrid experiments (not shown).

By 18 h, all experiments’ areal coverages were comparable and remained similar thereafter. Although, generally, 18–36-h FSS differences between experiments were smaller than those over the first 12 h, possibly owing to increasing influence of LBCs with forecast time (e.g., Romine et al. 2014), some meaningful differences emerged. Compared to forecasts initialized from Hyb_SR_20km, forecasts initialized from continuously cycling DR hybrid analyses were improved or similar when 20-km static BECs were used but degraded when 4-km static BECs were used in the hybrid algorithm (Figs. 13 and 14), with the largest Hyb_DR_20km_static improvements relative to Hyb_SR_20km for bigger $q$ and $r$. Conversely, even when 20-km static BECs were used, 18–36-h forecasts initialized from cool-start DR hybrid analyses were poorer compared to
Hyb_SR_20km (Fig. 14). Resolution of the ensemble perturbations mattered little within the cool-start experiments (cf. Hyb_DR_Cool_start_4km_static and Hyb_SR_4km_Cool_start in Fig. 14).

The 18–36-h 3DVAR forecasts were usually collectively poorest, but when 20-km static BECs were used, differences between forecasts initialized with true 4-km and downscaled 20-km 3DVAR analyses were larger than those between the corresponding hybrid experiments, especially for larger $r$. Consistent with the hybrid experiments, forecasts initialized from 3DVAR analyses incorporating 4-km static BECs were worse than those using 20-km static BECs.

Within the 18–36-h range, FSSs varied substantially (Fig. 15). The 3DVAR experiments were usually worst, and Hyb_DR_4km_static always had comparable or lower FSSs than Hyb_DR_20km_static. Most hourly differences between the continuously cycling DR hybrid experiments and Hyb_SR_20km were not SS at the 90% level. However, at most thresholds, other than between 24–27 and 30–33 h, Hyb_DR_20km_static had the highest FSSs, including between 18 and 24 h, the main convective initiation period.

Overall, higher-resolution analyses improved 4-km forecasts the most through ~12 h, consistent with Seity et al. (2011), who obtained similar results with high- and low-resolution 3DVAR DA systems. However, some benefit of the continuously cycled DR hybrid system using 20-km static BECs (i.e., Hyb_DR_20km_static) compared to Hyb_SR_20km persisted to later times, suggesting information from initial finescale structures endured throughout the 36-h forecast. Additionally, comparing forecasts initialized from Hyb_DR_20km_static and Hyb_DR_Cool_start_20km_static indicates that continuous cycling of 4-km backgrounds was directly responsible for forecast improvements. The similar performances of Hyb_DR_Cool_start_4km_static and Hyb_SR_4km_Cool_start suggest little sensitivity to ensemble BEC resolution, which agrees with some of Schwartz et al. (2015a)’s findings at convective-parameterizing resolution.

2) QUALITY OF THE 4-KM DATA ASSIMILATION SYSTEMS

Differences regarding 4-km precipitation forecast skill were associated with varying 4-km analysis system
quality, as illustrated by Fig. 16, which shows aggregate 0000 UTC 4-km background RMSEs and additive biases compared to radiosonde and aircraft observations over the 4-km domain. When 4-km hybrid backgrounds were spun up over 6 h compared to continuously cycled, temperature and moisture biases and RMSEs were usually statistically significantly poorer at the 90% level (cf. the blue and red lines in Figs. 16c,e,f) and wind RMSEs were usually larger (Figs. 16a,b,d). These statistics suggest that 4-km backgrounds spun up over 6 h from downscaled 20-km ICs had inferior quality than 4-km continuously cycled backgrounds. Additionally, although smaller static BEC length scales and larger static background error variances allowed Hyb_DR_4km_static analyses to fit observations more closely than Hyb_DR_20km_static

**FIG. 11.** FSSs computed with a 75-km radius of influence for the (a) 97th, (b) 98th, (c) 99th, (d) 99.25th, (e) 99.5th, and (f) 99.75th percentile thresholds aggregated over all 55 forecasts and the first 12 h. Statistical significance compared to Hyb_SR_20km is denoted at the top of each panel by “+” and “−” symbols, where the colors denote different experiments corresponding to those in the legend (e.g., the red symbols denote differences between Hyb_DR_20km_static and Hyb_SR_20km). For a particular forecast hour, if a “+” symbol is present, then the experiment had a statistically significantly higher FSS than Hyb_SR_20km at the 90th percentile. Conversely, if a “−” symbol is present, the experiment had a statistically significantly lower FSS than Hyb_SR_20km at the 90th percentile. If no symbol is present, then the difference compared to Hyb_SR_20km was not statistically significant at the 90% level. For example, at the 97th percentile threshold, Hyb_DR_20km_static had statistically significantly higher FSSs at the 90% level for forecast hours 1–7 and 11–12 compared to Hyb_SR_20km (red “+” symbols), while 3DVAR_20km had statistically significantly lower FSSs at the 90% level compared to Hyb_SR_20km (gray “−” symbols) at all forecast hours.
(not shown), background Hyb_DR_20km_static RMSEs were usually, and sometimes statistically significantly, smaller compared to background Hyb_DR_4km_static RMSEs (cf. the green and red lines in Fig. 16). These results suggest both that 4-km static BECs may have engendered analysis fits to observations closer than desired and smaller static BEC length scales, which limited the extent to which observational content was spread to surrounding areas, were detrimental. The 4-km background RMSEs for the two 3DVAR experiments were usually highest (cf. the red, orange, and yellow lines in Fig. 16), with differences compared to Hyb_DR_20km_static statistically significant at the 90th percentile at a majority of pressure levels for all observation types. Commensurate with its poor precipitation forecasts, 3DVAR_4km_BEC4 consistently had the worst RMSEs.

These different background characteristics were consistent with variations in precipitation forecast quality. Generally, experiments with poorer (better) observation fits to backgrounds initialized comparatively bad (good) precipitation forecasts, and the background RMSEs and biases compared to observations again suggest the continuously cycling DR hybrid experiment with 20-km static BECs (i.e., Hyb_DR_20km_static) was the best 4-km analysis system.

7. Discussion and summary

Several hybrid DA systems produced analyses between 0000 UTC 4 May and 0000 UTC 30 June 2013 over a CONUS-spanning domain. Hybrid analyses were generated with both 20- and 4-km horizontal grid
Some hybrid experiments employed continuously cycling configurations while others used a “cool-start” approach to spin up 4-km backgrounds over 6 h. Continuously cycling EnSRF and 3DVAR analyses were also produced. Beginning 7 May, all 0000 UTC analyses initialized 36-h ARW model forecasts, where a large convection-allowing 4-km nest was embedded within a 20-km parent domain. Rarely, if ever before, have convection-allowing ensemble-based DA systems been successfully implemented over geographic regions as large as the 4-km domain and in frameworks suitable for operations.

The 4-km forecasts were evaluated with a focus on accumulated precipitation, and verification statistics support the following conclusions:

1) Considering all time periods, 4-km rainfall forecasts initialized from downscaled 20-km hybrid analyses were more skillful than those initialized from operational GFS and downscaled 20-km EnSRF and 3DVAR analyses, corroborating SL14.

2) There was substantial sensitivity to the static BECs used within 4-km hybrid and 3DVAR systems, with larger static BEC horizontal length scales associated with better forecasts. This finding suggests that allowing observations on the 4-km grid to more broadly impact their surroundings was beneficial and may have helped constrain model error growth. Within hybrid configurations, larger static BEC horizontal length scales may have been beneficial owing to their consistency with ensemble BEC horizontal length scales (Fig. 4), and there may have been negative impacts from “mixing” the 4-km static BECs with the 20-km ensemble BECs. These results suggest the NMC method could be
improved for HR applications, and future work should investigate methods of constructing HR static BECs.

3) Continuously cycling the 4-km backgrounds yielded better forecasts than spinning up 4-km backgrounds over 6 h.

4) In terms of both bias at heavier rainfall rates and skill, initializing 4-km forecasts with true 4-km hybrid analyses, rather than downscaled 20-km hybrid analyses, unequivocally yielded superior precipitation forecasts over the first 12 h, regardless of the static BECs or whether 4-km hybrid backgrounds were obtained by continuous cycling or cool-start approaches. However, compared to forecasts initialized from downscaled 20-km hybrid analyses (i.e., Hyb_SR_20km), 18–36-h forecasts were worse when initialized from 4-km DR hybrid analyses when the 4-km background was either spun up from 20-km ICs over 6 h or 4-km static BECs were used. But, the DR hybrid experiment that employed continuous cycling and used 20-km static BECs (i.e., Hyb_DR_20km_static) initialized comparable or better 18–36-h precipitation forecasts than Hyb_SR_20km. Benefits of HR 3DVAR analyses endured throughout the forecast when 20-km static BECs were used.

5) The analysis method was more important than IC resolution, as 4-km precipitation forecasts initialized from downscaled 20-km hybrid analyses were more skillful than 4-km precipitation forecasts initialized from true 4-km 3DVAR analyses.

6) In the cool-start framework, for 18–36-h forecasts, there was little sensitivity to whether 4- or 20-km ensembles were ingested into 4-km hybrid analyses, although some degradation occurred over the first 12 h when 4-km ensembles were combined with deterministic 4-km backgrounds. However, continuously cycling experiments are needed to confirm these results and provide further insight, because differences due to perturbation resolution may accumulate throughout assimilation cycles. Unfortunately, these experiments are very expensive to perform over domains large enough to resolve meso-α- to synoptic-scale motions.

As the impact of producing high-resolution analyses was greatest over the first ~12 h, whether it is preferable to perform convection-allowing analyses may depend somewhat on the application. However, even for 18–36-h forecasts, the best DR hybrid configuration (Hyb_DR_20km_static) produced comparable or better forecasts than Hyb_SR_20km, suggesting little risk of initializing forecasts with HR analyses, so long as the DA system is constructed carefully.

These collective findings indicate that using 20-km ensembles within 4-km hybrid analyses was a practical, effective, and affordable way to generate HR analyses incorporating beneficial flow-dependent

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Percentile Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>97</td>
</tr>
<tr>
<td>Hyb_SR_4km_Cool_start</td>
<td>r = 25-km</td>
</tr>
<tr>
<td></td>
<td>r = 75-km</td>
</tr>
<tr>
<td></td>
<td>r = 125-km</td>
</tr>
<tr>
<td>Hyb_DR_Cool_start_4km_static</td>
<td>r = 25-km</td>
</tr>
<tr>
<td></td>
<td>r = 75-km</td>
</tr>
<tr>
<td></td>
<td>r = 125-km</td>
</tr>
<tr>
<td>Hyb_DR_Cool_start_20km_static</td>
<td>r = 25-km</td>
</tr>
<tr>
<td></td>
<td>r = 75-km</td>
</tr>
<tr>
<td></td>
<td>r = 125-km</td>
</tr>
<tr>
<td>Hyb_DR_4km_static</td>
<td>r = 25-km</td>
</tr>
<tr>
<td></td>
<td>r = 75-km</td>
</tr>
<tr>
<td></td>
<td>r = 125-km</td>
</tr>
<tr>
<td>Hyb_DR_20km_static</td>
<td>r = 25-km</td>
</tr>
<tr>
<td></td>
<td>r = 75-km</td>
</tr>
<tr>
<td></td>
<td>r = 125-km</td>
</tr>
<tr>
<td>3DVAR_4km_BEC20</td>
<td>r = 25-km</td>
</tr>
<tr>
<td></td>
<td>r = 75-km</td>
</tr>
<tr>
<td></td>
<td>r = 125-km</td>
</tr>
<tr>
<td>3DVAR_4km_BEC4</td>
<td>r = 25-km</td>
</tr>
<tr>
<td></td>
<td>r = 75-km</td>
</tr>
<tr>
<td></td>
<td>r = 125-km</td>
</tr>
</tbody>
</table>

**FIG. 14.** As in Fig. 10, but significance levels for aggregate 18–36-h FSS differences compared to Hyb_SR_20km.
BECs. Nonetheless, there are many areas for future work and analysis related to the topics in this paper. For example, it is unclear whether continuously cycling at 4-km horizontal grid spacing with a 6-h period would be stable over geographic regions with sparser observation networks. Thus, future work should consider more frequent cycling (e.g., 1-h cycles) to incorporate ensemble BECs into convection-allowing analyses. Additionally, future studies should examine the assimilation of radar and geostationary satellite observations into high-resolution backgrounds to further improve short-term forecasts (e.g., Kain et al. 2010). Furthermore, the cool-start analysis systems, whereby the high-resolution backgrounds were spun up over 6 h, could potentially be improved. It is possible that the spinup process was incomplete at 6 h (e.g., Raynaud and Bouttier 2016), which may have negatively impacted the cool-start analyses. Thus, further work might examine spinup characteristics more closely and potentially lengthen the spinup period. Moreover, spinup backgrounds may be improved if hourly DA occurred during the spinup period. Finally, additional work is needed to determine whether similar results are obtained over different seasons (e.g., winter) and topographically diverse regions and if large-scale forcing characteristics influence the benefit of high-resolution ICs.

The DR hybrid systems described herein do not produce convection-allowing ensemble analyses. While convection-parameterizing EnKF analysis ensembles can be downscaled to convection-allowing resolution and initialize convection-allowing ensemble forecasts (e.g., Romine et al. 2014; Schumacher and Clark 2014; Schwartz et al. 2014), it is likely that ensemble forecasts
would also benefit from increased analysis resolution, and, indeed, convection-allowing operational ensemble DA systems are being developed (e.g., Harnisch and Keil 2015).

Direct extensions of this work include incorporating four-dimensional BECs in the hybrid DA algorithm (e.g., Buehner et al. 2010; Wang and Lei 2014; Kleist and Ide 2015b; Lorenc et al. 2015) within both SR and DR hybrid configurations in hopes of yielding even further forecast improvements, and if computing resources permit, convection-allowing analyses with continuously cycling convection-allowing ensembles will be performed to further understand convection-allowing hybrid analysis sensitivity to ensemble perturbation resolution.

Acknowledgments. Computing for these experiments was provided by the Yellowstone supercomputer (ark:/85065/d7wd3xhc) maintained by NCAR’s Computational and Information Systems Laboratory. Three anonymous reviewers provided constructive comments that improved the clarity of this paper. Thanks to Chris Snyder and Ryan Sobash (NCAR/MMM) for reviewing an earlier version of this manuscript.

REFERENCES


1802 MONTHLY WEATHER REVIEW VOLUME 144

1350-4827/05001763.

117WAF-D-10-05088.1.

1350-4827/05001763.

117WAF-D-10-05088.1.

1350-4827/05001763.

117WAF-D-10-05088.1.

117WAF-D-10-05088.1.

1350-4827/05001763.

117WAF-D-10-05088.1.

1350-4827/05001763.

117WAF-D-10-05088.1.

1350-4827/05001763.

117WAF-D-10-05088.1.

1350-4827/05001763.


