Toward Unifying Short-Term and Next-Day Convection-Allowing Ensemble Forecast Systems with a Continuously Cycling 3-km Ensemble Kalman Filter over the Entire Conterminous United States

CRAIG S. SCHWARTZ, a GLEN S. ROMINE, a and DAVID C. DOWELL b

a National Center for Atmospheric Research, Boulder, Colorado
b NOAA/Earth System Research Laboratory, Boulder, Colorado

ABSTRACT: Using the Weather Research and Forecasting Model, 80-member ensemble Kalman filter (EnKF) analyses with 3-km horizontal grid spacing were produced over the entire conterminous United States (CONUS) for 4 weeks using 1-h continuous cycling. For comparison, similarly configured EnKF analyses with 15-km horizontal grid spacing were also produced. At 0000 UTC, 15- and 3-km EnKF analyses initialized 36-h, 3-km, 10-member ensemble forecasts that were verified with a focus on precipitation. Additionally, forecasts were initialized from operational Global Ensemble Forecast System (GEFS) initial conditions (ICs) and experimental “blended” ICs produced by combining large scales from GEFS ICs with small scales from EnKF analyses using a low-pass filter. The EnKFs had stable climates with generally small biases, and precipitation forecasts initialized from 3-km EnKF analyses were more skillful and reliable than those initialized from downscaled GEFS and 15-km EnKF ICs through 12-18 and 6-12 h, respectively. Conversely, after 18 h, GEFS-initialized precipitation forecasts were better than EnKF-initialized precipitation forecasts. Blended 3-km ICs reflected the respective strengths of both GEFS and high-resolution EnKF ICs and yielded the best performance considering all times: blended 3-km ICs led to short-term forecasts with similar or better skill and reliability than those initialized from unblended 3-km EnKF analyses and −18–36-h forecasts possessing comparable quality as GEFS-initialized forecasts. This work likely represents the first time a convection-allowing EnKF has been continuously cycled over a region as large as the entire CONUS, and results suggest blending high-resolution EnKF analyses with low-resolution global fields can potentially unify short-term and next-day convection-allowing ensemble forecast systems under a common framework.

KEYWORDS: Data assimilation; Ensembles; Model evaluation/performance; Numerical weather prediction/forecasting

1. Introduction

Convection-allowing ensembles (CAEs) produce better precipitation and severe weather forecasts than coarser-resolution, convection-parameterizing ensembles (e.g., Clark et al. 2009; Duc et al. 2013; Iyer et al. 2016; Schellander-Gorgas et al. 2017), are operational at many weather forecasting offices (e.g., Gebhardt et al. 2011; Peralta et al. 2012; Hagelin et al. 2017; Raynaud and Bontabbio 2017; Firank et al. 2018; Klassa et al. 2018), and have proven useful and valuable for various meteorological applications around the world (e.g., Xue et al. 2007; Clark et al. 2012; Evans et al. 2014; Maurer et al. 2017; Zhang 2018; Cafaro et al. 2019; Porson et al. 2019; Schwartz et al. 2019). Thus, as computing power has increased, CAE domains have gradually enlarged, with operational global CAEs on the horizon.

While CAEs can be initialized by downscaling coarser-resolution, convection-parameterizing analyses, convection-allowing numerical weather prediction (NWP) models are typically best when initialized from correspondingly convection-allowing analyses, particularly for short-term forecasts (e.g., Ancell 2012; Harnisch and Keil 2015; Johnson et al. 2015; Johnson and Wang 2016; Raynaud and Bontabbio 2016; Schwartz 2016; Gustafsson et al. 2018). Therefore, to produce the best possible CAE forecasts over ever-expanding domains, convection-allowing data assimilation (DA) systems over large areas are needed to provide optimal initial conditions (ICs).

However, there are obstacles to implementing convection-allowing DA systems over domains large enough to resolve mesoalpha- to synoptic-scale features, especially when using state-of-the-science ensemble-based DA algorithms like the ensemble Kalman filter (EnKF; Evensen 1994; Houtekamer and Zhang 2016), which produces flow-dependent analysis components and has become popular for initializing CAEs (e.g., Jones and Stensrud 2012; Melhauser and Zhang 2012; Schumacher and Clark 2014; Schwartz et al. 2014, 2015a,b, 2019). One challenge is simply computational expense, which grows directly with domain size, 1 and accordingly, most convection-allowing EnKFs and their associated CAE forecasts have relatively small domains centered on a single European country (e.g., Schraff et al. 2016; COSMO 2020) or a small portion of the conterminous United States (CONUS). For example, NOAA’s experimental “Warn-on-Forecast” (WoF; Stensrud et al. 2009, 2013) system, initialized from 36-member 3-km EnKF analyses, covers less than

1 Mixed-resolution DA systems (e.g., Gao and Xue 2008; Rainwater and Hunt 2013; Li et al. 2015) possessing both convection-allowing and convection-parameterizing resolution components can lessen costs and make large-domain convection-allowing analyses more feasible (e.g., Schwartz 2016; Rogers et al. 2017).
1000 km × 1000 km (Wheatley et al. 2015; Jones et al. 2016, 2018, 2020; Skinner et al. 2018).

Fortunately, computing challenges can be overcome with increased resources, and recently, several studies initialized CAE forecasts from 40-member EnKF analyses with 3-km or finer horizontal grid spacing over the entire CONUS (Duda et al. 2019; Gasperoni et al. 2020; Johnson et al. 2020). Similarly, NOAA’s real-time, experimental High-Resolution Rapid Refresh Ensemble (HRRRE) is initialized from CONUS-spanning, 3-km, 36-member EnKF analyses (Dowell et al. 2016; Ladwig et al. 2018). However, 36–40-member EnKFs are likely smaller than desirable, considering that operational global EnKFs run by the United States and Canada, respectively, have 80 and 256 members, and generally, EnKFs benefit from larger ensembles (e.g., Zhang et al. 2013; Houtekamer et al. 2014).

But, even with unlimited resources, there are fundamental scientific concerns that must be addressed to develop stable, high-quality, convection-allowing EnKFs over large regional domains, especially in continuously cycling limited-area EnKFs where external models are relegated to providing boundary conditions. In particular, model physics deficiencies can lead to accumulation of biases throughout EnKF DA cycles, potentially degrading analysis system performance and subsequent forecasts (e.g., Torn and Davis 2012; Romine et al. 2013; Cavallo et al. 2016; Wong et al. 2020). Although all continuously cycling limited-area EnKFs are prone to bias accumulation, this issue may be exacerbated as both model resolution and domain size increase: biases may accumulate more in high-resolution EnKFs than low-resolution EnKFs because of rapid small-scale error growth (e.g., Lorenz 1969; Zhang et al. 2003; Hohenegger and Schär 2007; Judt 2018), and EnKFs over large domains may suffer from bias accumulations more than EnKFs over small domains because of reduced influence from lateral boundaries provided by potentially less biased global models (e.g., Warner et al. 1997; Romine et al. 2014; Schumacher and Clark 2014).

Given these scientific and computing challenges, operational convection-allowing continuously cycling EnKFs and attendant CAEs over Europe have small domains (e.g., Schraff et al. 2016; COSMOS 2020), while large-domain convection-allowing EnKFs over the CONUS (e.g., Duda et al. 2019; Gasperoni et al. 2020; Johnson et al. 2020; HRRRE) employ “partial cycling” strategies that periodically discard convection-allowing analysis cycles and replace them with coarser-resolution, large-scale external analyses in hopes of tempering bias accumulations (e.g., Hsiao et al. 2012; Benjamin et al. 2016; Wu et al. 2017). This partial cycling approach over the CONUS seems justified, as Schwartz et al. (2020) showed that a limited-area continuously cycling EnKF with convection-parameterizing resolution did not initialize better CAE precipitation forecasts over the CONUS than downscaled global analyses.

Nonetheless, as discussed at length by Schwartz et al. (2019), continuously cycling EnKFs have many attractive properties for CAE initialization, including the ability to diagnose model biases while simultaneously producing flow-dependent ICs that are dynamically consistent with and span all possible resolvable scales of the convection-allowing forecast model. Thus, despite formidable challenges, it is desirable to further explore and develop continuously cycling EnKFs over large geographic areas at convection-allowing resolutions for CAE initialization purposes.

Accordingly, we produced continuously cycling, 80-member, 3-km EnKF analyses with a 1-h cycling period for 4 weeks over a computational domain spanning the entire CONUS. EnKF analysis ensembles then initialized 36-h, 3-km, 10-member CAE forecasts. For comparison, 3-km CAE forecasts were also initialized by downscaling both 15-km EnKF analyses and global ICs produced for NCEP’s operational Global Ensemble Forecast System (GEFS; Zhou et al. 2017). The impact of assimilating radar observations into the 3-km EnKF was also assessed. Relative to the EnKF described in Schwartz et al. (2020), our EnKFs used more advanced observation processing, an upgraded NWP model, and a shorter cycling period, and inclusion of 3-km EnKF DA was also new. To our knowledge, this work presents the first time convection-allowing continuously cycling EnKF analyses have been produced over the entire CONUS.

Results indicated benefits of EnKF-initialized forecasts with respect to GEFS-initialized forecasts diminished with forecast length, presumably because large-scale fields were better represented in GEFS ICs and became more important at longer forecast ranges. These findings motivated experimentation with a “blending” approach combining large-scale fields from an external (e.g., global) NWP model with small-scale fields from a limited-area model, which can be achieved by augmenting a variational cost function with a global model constraint (e.g., Guidard and Fischer 2008; Dahlgren and Gustafsson 2012; Vendrasco et al. 2016; Keresturi et al. 2019) or using filtering to perform scale separation (e.g., Yang 2005; Wang et al. 2011; Caron 2013; H. Wang et al. 2014; Y. Wang et al. 2014; Hsiao et al. 2015; Zhang et al. 2015; Feng et al. 2020); we used a low-pass filter to combine large scales from GEFS ICs with small scales from EnKF analyses. These previous studies collectively suggested blended limited-area ICs improved forecasts compared to those initialized from unblended limited-area ICs, including for a CAE within a perturbed-observation variational DA framework (Keresturi et al. 2019). However, our application of blending within the context of a large-domain convection-allowing continuously cycling EnKF was unique, and, as described below, blending global fields with high-resolution EnKF analyses can potentially unite short-term and next-day (18–36-h) CAE forecast systems under a common framework.

2. Model configurations, EnKF settings, and experimental design

a. Forecast model

All forecasts were produced by version 3.9.1.1 of the Advanced Research Weather Research and Forecasting (WRF) Model (Skamarock et al. 2008; Powers et al. 2017) over a nested computational domain (Fig. 1a). The horizontal grid spacing was 15 km in the outer domain and 3 km in the nest, and time steps were 60 and 12 s in the 15- and 3-km
domains, respectively. Both domains had 51 vertical levels distributed as in the Rapid Refresh model (Benjamin et al. 2016) with a 15-hPa top. Physical parameterizations were identical across the two domains (Table 1), except no cumulus parameterization was employed on the convection-allowing 3-km grid, and all ensemble members used common physics and dynamics options.

b. EnKF DA systems

1) EnKF Experiments and Configurations

Two primary DA experiments with 80-member ensembles were performed using an ensemble adjustment Kalman filter (Anderson 2001, 2003; Anderson and Collins 2007), a type of EnKF, as implemented in the Data Assimilation Research Testbed

---

**TABLE 1. Physical parameterizations for all WRF Model forecasts. Cumulus parameterization was only used on the 15-km domain.**

<table>
<thead>
<tr>
<th>Physical parameterization</th>
<th>WRF 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 (15-km domain only)</td>
<td>Tiedtke (1989); Zhang et al. (2011)</td>
</tr>
</tbody>
</table>
(DART; Anderson et al. 2009) software. The first EnKF experiment only produced analyses on the 15-km domain (Fig. 1a), and the 3-km domain was removed during WRF Model advances between EnKF analyses. Conversely, the second EnKF experiment produced separate, independent analyses on both the 15- and 3-km domains, with nested WRF Model forecasts between EnKF analyses. During these nested forecasts, which were ~45 times more expensive than the single-domain 15-km model advances, one-way feedback was employed such that the 15-km EnKF DA system was unaffected by the 3-km EnKF DA system (i.e., 15-km fields in the nested- and single-domain EnKF DA systems were identical), permitting a clean comparison of analysis and forecast sensitivity to EnKF resolution. The 15- and 3-km EnKFs updated identical state variables (Table 2), with hydrometers included in anticipation of experimentation with radar DA (section 4c).

Initial 80-member ensembles were produced by interpolating the 0.25° NCEP Global Forecast System (GFS) analysis at 0000 UTC 23 April 2017 onto the 15-km domain and adding random, correlated, Gaussian noise with zero mean (e.g., Barker 2005; Torn et al. 2006) drawn from background error covariances provided by the WRF Model’s DA system (Barker et al. 2012). The randomly produced 15-km ensemble was then downscaled onto the 3-km grid to initialize the 3-km EnKF, ensuring initial 15- and 3-km ensembles were identical aside from interpolation errors. These randomly generated ensembles served as prior (before assimilation) ensembles for the first EnKF analyses, and the posterior (after assimilation) ensembles at 0000 UTC 23 April 2017 initialized 1-h, 80-member ensemble forecasts that became prior ensembles for the next EnKF analyses at 0100 UTC 23 April 2017. Analysis–forecast cycles with a 1-h period continued until 0000 UTC 20 May 2017 (649 total DA cycles). This experimental period (23 April–20 May 2017) was similar to that in Schwartz (2019), which featured several heavy precipitation episodes primarily driven by strong synoptic forcing, a broad overall precipitation maximum centered in Missouri (Fig. 1b), and a variety of flow patterns (Figs. 1c–e).

During EnKF cycles, soil states freely evolved for each member, sea surface temperature was updated daily from NCEP’s 0.12° analyses (e.g., Gemmill et al. 2007), and identical randomly perturbed lateral boundary conditions (LBCs) were applied to the 15-km domain in each DA system, with perturbations for individual members generated using the same method to produce initial ensembles at 0000 UTC 23 April 2017. The first two days of cycling were regarded as spinup.

Spurious correlations due to sampling error were mitigated with a sampling error correction scheme (Anderson 2012) and covariance localization [Eq. (4.10) of Gaspari and Cohn (1999)]. Vertical localization limited analysis increments to ±1.0 scale height (in log pressure coordinates) away from an observation in both the 15- and 3-km EnKFs. However, horizontal localizations differed depending on EnKF resolution: 15-km EnKF analysis increments were forced to zero 1280 km from an observation, but to lessen expense and complete 3-km EnKF analyses quickly enough for operational applications, 3-km EnKF analysis increments were forced to zero 640 km from an observation, except rawinsonde observations could produce increments up to 1280 km away (Table 2). The vertical and 15-km EnKF horizontal localization distances were guided by previous experiences with DART (e.g., Romine et al. 2013, 2014; Schwartz et al. 2015a,b, 2019), and while our 3-km EnKF horizontal localization distances were similar to Johnson et al. (2015), they were larger than those in many other convection-allowing EnKFs (e.g., Harnisch and Keil 2015; Yussouf et al. 2015, 2016; Degelia et al. 2018; Gasperoni et al. 2020, Jones et al. 2020). However, these studies with smaller localization distances either used partial cycling strategies or only continuously cycled for a short period (days), and we believed that larger localization distances were necessary to provide stronger observational constraints in a large-domain continuously cycling 3-km EnKF.

EnKF spread was maintained by applying covariance inflation to posterior state-space perturbations about the ensemble mean following Whitaker and Hamill (2012)’s “relaxation-to-prior spread” algorithm with an inflation parameter $\alpha = 1.06$ in both the 15- and 3-km EnKFs. As noted by Schwartz and

---

**Table 2. Summary of EnKF configurations.**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>15-km EnKF</th>
<th>3-km EnKF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble size</td>
<td>80 members</td>
<td>80 members</td>
</tr>
<tr>
<td>Updated WRF Model variables</td>
<td>Zonal and meridional wind components; perturbation geopotential height, potential temperature, and dry surface pressure; and water vapor, graupel, snow, and rain mixing ratios</td>
<td>Zonal and meridional wind components; perturbation geopotential height, potential temperature, and dry surface pressure; and water vapor, graupel, snow, and rain mixing ratios</td>
</tr>
<tr>
<td>Horizontal localization full width</td>
<td>1280 km</td>
<td>640 km, except 1280 km for rawinsonde observations</td>
</tr>
<tr>
<td>Vertical localization full width</td>
<td>1.0 scale height</td>
<td>1.0 scale height</td>
</tr>
<tr>
<td>Inflation factor ($\alpha$)</td>
<td>1.06</td>
<td>1.06</td>
</tr>
<tr>
<td>Horizontal thinning for aircraft and satellite-tracked wind observations</td>
<td>30 km</td>
<td>15 km</td>
</tr>
<tr>
<td>Vertical thinning for aircraft and satellite-tracked wind observations</td>
<td>25 hPa</td>
<td>25 hPa</td>
</tr>
</tbody>
</table>
Liu (2014), $\alpha > 1$ meant inflated posterior spread was greater than prior spread, which, while counterintuitive, was necessary to maintain reasonable spread given absence of other spread-inducing methods like multiphysics ensembles, additive inflation, or stochastic physics. Several iterative weeklong trials with 15-km EnKFs were performed to settle on $\alpha = 1.06$, which provided acceptable prior observation-space statistics for the assumed observation errors (section 3).

2) OBSERVATIONS

Although DART has observation processing capabilities, we instead used NCEP’s operational Gridpoint Statistical Interpolation (GSI) DA system (Kleist et al. 2009; Shao et al. 2016) for observation processing, which, relative to DART, has more sophisticated quality control, observation thinning, and observation error assignment capabilities. In addition, GSI’s observation operators were used instead of DART’s built-in observation operators to produce model-simulated conventional observations. Initially specified observation errors were based on the HRRRE and identical in the 15- and 3-km EnKFs (Fig. 2; Table 3); GSI adjusted these errors to produce “final” observation error standard deviations $\sigma_o$ actually used in the assimilation, as described by several texts (e.g., Schwartz and Liu 2014; Developmental Testbed Center 2016; Johnson and Wang 2017). These adjustments often inflated initially specified observation errors (Fig. 2).

Time windows for the observation platforms varied and were based on Rapid Refresh model (Benjamin et al. 2016) and
HRRRE settings, with generally smaller windows for frequently reporting, stationary platforms, like METAR observations (Table 3), and all observations were assumed valid at the analysis time. Moisture observations were initially processed as specific humidity, but because GSI requires moisture observation errors in terms of relative humidity, moisture observations were ultimately converted to and assimilated as relative humidity using the prior ensemble mean saturation specific humidity. Satellite-tracked wind and aircraft observations were thinned such that remaining observations were spaced 25 hPa apart vertically and 30 and 15 km apart horizontally in the 15- and 3-km EnKFs, respectively (Table 2); these different horizontal thinnings were chosen so the 15- and 3-km EnKFs had equal numbers of satellite-tracked wind and aircraft observations within their respective horizontal localization radii. Radiance observations were not assimilated since they generally yield small impacts over the CONUS (Lin et al. 2017) given the multitude of available conventional observations. Additionally, the EnKFs did not assimilate radar observations, although an auxiliary experiment was performed where radar observations were assimilated with a 3-km EnKF (section 4c).

Observations were subject to numerous quality control procedures, such as excluding observations from specific aircraft with known biases and applying an “outlier check” to reject observations whose ensemble mean innovations were greater than some threshold 

\[ a \]  

where \( a \) varied from 2.5 to 10 depending on observation type and platform (Table 3). These \( a \) were generally fairly lenient and allowed most observations to pass the outlier check, which, along with our relatively large localization distances, reflected a philosophy that we wanted observations to heavily constrain the 1-h WRF Model forecasts between EnKF analyses. Overall, the EnKFs assimilated 30 000–100 000 conventional observations each cycle, with a relative dearth of overnight observations due to fewer commercial flights and maxima at 0000 and 1200 UTC reflecting the majority of rawinsonde launches (Fig. 3). Ultimately, GSI-provided observations, final observation errors, and prior model-simulated observations for each ensemble member were ingested directly into DART for use in EnKF DA.

### Table 3: Conventional observations that were assimilated and their outlier check thresholds, time windows, and initially specified observation error standard deviations.

<table>
<thead>
<tr>
<th>Observing platform</th>
<th>Observation type</th>
<th>Initial observation error</th>
<th>Outlier check threshold (a)</th>
<th>Time window (h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rawinsonde</td>
<td>Surface pressure</td>
<td>Fig. 2d</td>
<td>5</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>Temperature</td>
<td>Fig. 2b</td>
<td>7</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>Relative humidity</td>
<td>Fig. 2c</td>
<td>7</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>Wind</td>
<td>Fig. 2a</td>
<td>10</td>
<td>1.5</td>
</tr>
<tr>
<td>Aircraft</td>
<td>Temperature</td>
<td>Fig. 2b</td>
<td>7</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>Relative humidity</td>
<td>Fig. 2c</td>
<td>7</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>Wind</td>
<td>Fig. 2a</td>
<td>10</td>
<td>0.75</td>
</tr>
<tr>
<td>Wind profiler</td>
<td>Wind</td>
<td>Fig. 2a</td>
<td>5</td>
<td>0.4</td>
</tr>
<tr>
<td>Global positioning system radio occultation (GPSRO)</td>
<td>Refractivity</td>
<td>1% of observation value</td>
<td>10</td>
<td>3.0</td>
</tr>
<tr>
<td>Infrared and water vapor channel satellite-tracked wind</td>
<td>Wind</td>
<td>Fig. 2a</td>
<td>2.5</td>
<td>1.5</td>
</tr>
<tr>
<td>Ship and buoy</td>
<td>Surface pressure</td>
<td>0.44 hPa</td>
<td>5</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>Temperature</td>
<td>0.8 K</td>
<td>7</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>Relative humidity</td>
<td>3.9%</td>
<td>7</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>Wind</td>
<td>1.45 m s(^{-1})</td>
<td>5</td>
<td>1.5</td>
</tr>
<tr>
<td>SYNOP and METAR</td>
<td>Surface pressure</td>
<td>0.54 hPa</td>
<td>5</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>Temperature</td>
<td>2.3 K</td>
<td>5</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>Relative humidity</td>
<td>3.4%</td>
<td>7</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>Wind</td>
<td>1.2 m s(^{-1})</td>
<td>5</td>
<td>0.25</td>
</tr>
<tr>
<td>Oklahoma and West Texas mesonet</td>
<td>Surface pressure</td>
<td>0.35 hPa</td>
<td>5</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>Temperature</td>
<td>1.5 K</td>
<td>5</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>Relative humidity</td>
<td>4%</td>
<td>7</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>Wind</td>
<td>1.1 m s(^{-1})</td>
<td>5</td>
<td>0.1</td>
</tr>
</tbody>
</table>

---

\([2]\) The “innovation” is the difference between an observation and the prior model-simulated observation.

### 3) Forecast Initialization

EnKF analysis ensembles initialized 36-h 10-member ensemble forecasts over the nested computational domain (Fig. 1a) at 0000 UTC between 25 April and 20 May 2017 (inclusive; 26 forecasts). Although 80 EnKF analysis members were available, due to computing constraints, 36-h forecasts were only initialized from members 1–10; 10-member CAEs are sufficient to provide skillful and valuable probabilistic forecasts (e.g., Clark et al. 2011, 2018; Schwartz et al. 2014) and similar in size as the HRRRE and NCEP's operational High-Resolution Ensemble Forecast system (Jirak et al. 2018). Choosing members 1–10 was effectively the same as randomly selecting 10 members since all ensemble members had identical configurations (e.g., Schwartz et al. 2014). In principle, free forecasts could have been initialized every hour, but given finite resources, forecasts were solely initialized at 0000 UTC.
which allowed us to focus on both short-term and next-day forecast periods featuring active convection.

When initializing 36-h forecasts from 15-km EnKF analyses, the 3-km nest was initialized by downscaling 15-km EnKF analyses onto the 3-km grid. Conversely, downscaling was unnecessary to initialize 36-h forecasts from the 3-km EnKF; 3-km ICs were simply 3-km EnKF analysis members. For both sets of EnKF-initialized 36-h forecasts, perturbation members 1–10 from the GEFS (Zhou et al. 2017) with 0.5° horizontal grid spacing provided LBCs at 3-h intervals for the 15-km domain, which in turn provided LBCs for the 3-km nest. While random LBCs could have been used for the 36-h forecasts as in the EnKF DA system, we believed it was more appropriate to use flow-dependent LBCs for these longer unconstrained forecasts.

c. Benchmark ensemble

To serve as a benchmark for the EnKF-initialized CAE forecasts, 36-h forecasts on the nested grid (Fig. 1a) with the configurations in section 2a were initialized by interpolating 0.5° ICs from perturbation members 1–10 of the GEFS onto the computational domain at 0000 UTC daily between 25 April and 20 May 2017 (inclusive), with LBCs provided by GEFS forecasts identically as in the EnKF-initialized CAEs. As described by Zhou et al. (2017), GEFS ICs were produced by adding 6-h forecast perturbations from a global EnKF DA system (Whitaker and Hamill 2002) to “hybrid” variational-ensemble analyses produced for NCEP’s deterministic GFS (e.g., Wang and Lei 2014; Kleist and Ide 2015a,b). Relative to the limited-area EnKF analyses, GEFS ICs were much coarser but reflected assimilation of many more observations, including satellite radiances. Overall, comparison of GEFS- and EnKF-initialized CAE forecasts provides insight about whether the vastly more expensive EnKF initialization procedure was warranted.

d. Blending

Based on performance of the EnKF- and GEFS-initialized CAE forecasts (section 4b), additional ensemble ICs were created by “blending” small scales from EnKF analyses with large scales from GEFS ICs. Blending was solely performed at 0000 UTC between 25 April and 20 May 2017 (inclusive) immediately after EnKF DA and before initializing 36-h CAE forecasts; blending was not employed within the context of continuously cycling EnKF DA, as the blended 0000 UTC fields were not used to initialize 1-h WRF Model forecasts that served as priors for the next DA cycle.
Specifically, ICs from corresponding GEFS and EnKF ensemble members were blended on both the 15- and 3-km domains to create new initial ensembles using

$$x_{\text{blend}}^i = (\text{EnKF}^i - \text{EnKF}_{\text{FILT}}^i) + \text{GEFS}_{\text{FILT}}^i,$$

where $x_{\text{blend}}^i$ represents the blended ICs for the $i$th ensemble member, EnKF is the EnKF analysis for the $i$th member, and EnKF$_{\text{FILT}}^i$ and GEFS$_{\text{FILT}}^i$ are the low-pass filtered EnKF and GEFS ICs for the $i$th member, respectively, for $i = 1, \ldots, 10$. To perform the scale separation, a low-pass, sixth-order implicit tangent filter (e.g., Raymond 1988; Raymond and Garder 1991) as implemented by several studies (e.g., Yang 2005; H. Wang et al. 2014; Hsiao et al. 2015; Feng et al. 2020) and given by

$$H(L) = [1 + \tan^{-6}(\pi\Delta x/L_x)\tan^2(\pi\Delta x/L_x)]^{-1}$$

was employed (Fig. 4), where $\Delta x$ is the horizontal grid spacing (either 15’ or 3 km), $L$ the wavelength, $H(L)$ the scale-dependent response function, and $L_x$ a specified filter cutoff (km) physically representing the spatial scale (wavelength) where the blended ICs (e.g., $x_{\text{blend}}^i$) had equal contributions from GEFS and EnKF initial states [i.e., when $L = L_x$, $H(L) = 0.5$]. Blending was applied at all 51 vertical levels to zonal and meridional wind components; perturbation geopotential height, potential temperature, and dry surface pressure; and water vapor mixing ratio, and the cutoff length was height and variable invariant.

We produced blended ICs using filter cutoff lengths $L_x$ of 640, 960, and 1280 km, guided by EnKF horizontal localization lengths and previous work suggesting values between 640 and 1280 km were appropriate (e.g., H. Wang et al. 2014; Hsiao et al. 2015; Feng et al. 2020). CAE forecasts initialized from these three sets of blended ICs objectively had similar skill, although $L_x = 960$ km yielded slightly better results. Therefore, results are shown only for the 960-km cutoff.

3. EnKF performance

To assess EnKF performance, we examined the observation-space bias and relationship between the prior ensemble mean root-mean-square error (RMSE) and “total spread,” the square root of the sum of the observation error variance $\sigma_o^2$ and ensemble variance of the simulated observations (Houtekamer et al. 2005). Ideally, the ratio of total spread to RMSE [termed the consistency ratio (CR; Dowell and Wicker 2009)] should be near 1.0. To fairly compare the 15- and 3-km EnKFs, we restricted this analysis solely to those observations assimilated by both EnKFs, although overall findings were unchanged when computing identical statistics with inhomogeneous samples. We focused on aircraft and rawinsonde observations because of their large impacts on springtime forecasts over the CONUS (James and Benjamin 2017).

Ensemble mean additive biases (model minus observations) and RMSEs aggregated over all prior ensembles (1-h forecasts) between 0000 UTC 25 April and 0000 UTC 20 May 2017 (inclusive) were similar in the 15- and 3-km EnKFs with respect to zonal wind and temperature observations at most levels (Figs. 5a,b,d,e), while biases and RMSEs for moisture were typically smaller in the 3-km EnKF (Figs. 5c,f). Magnitudes of temperature biases were typically $<0.1$ K, except near the surface and in the upper troposphere for rawinsonde observations (Fig. 5a); the latter is consistent with other continuously cycling EnKFs over the CONUS (e.g., Romine et al. 2013; Schumacher and Clark 2014; Schwartz and Liu 2014; Cavallo et al. 2016; Schwartz 2016) and likely due to closer fits to the more numerous aircraft observations that may have systematically warm biases compared to rawinsonde observations (Ballish and Krishna Kumar 2008). That upper-tropospheric temperature biases relative to aircraft observations (Fig. 5d) were smaller than and opposite the sign of temperature biases relative to rawinsonde observations (Fig. 5a) further supports this reasoning.

Prior total spreads were similar in both EnKFs (Fig. 5) and CRs were usually between 0.8 and 1.2, although CRs suggest
moisture observation errors could potentially be decreased. While more spread may have been expected in the 3-km EnKF because small-scale errors grow rapidly upscale (e.g., Lorenz 1969; Zhang et al. 2003; Hohenegger and Schär 2007), cumulus parameterization in the 15-km DA system may have served as an error source that compensated for missing storm-scale structures, and assimilating copious observations each cycle (Fig. 3) with fairly large localization distances highly constrained the 15- and 3-km EnKFs, limiting spread growth during 1-h WRF Model integrations between analyses. In balance, these factors potentially contributed to the similar 15- and 3-km prior spreads.

Overall, systematic biases were usually small and EnKF performance appeared acceptable. Moreover, after the first two days, prior total spread and ensemble mean biases were steady throughout the cycles (Fig. 6), and observation rejection rates varied little with time (not shown). These results indicate the continuously cycling EnKFs maintained stable climates, which is particularly noteworthy for the 3-km EnKF, as it has not previously been demonstrated that a convection-allowing EnKF can be continuously cycled over a large domain without deleterious consequences like a drifting model climate or filter divergence [see appendix A of Houtekamer and Zhang (2016) for a succinct summary of filter divergence].

4. Precipitation forecast verification

Hourly accumulated precipitation forecasts were verified against Stage IV (ST4) analyses (Lin and Mitchell 2005) produced at NCEP considered as “truth.” Objective evaluations were performed over the CONUS east of 105°W (hereafter the “verification region”; Fig. 1a), where ST4 analyses were most robust (e.g., Nelson et al. 2016). For metrics requiring a common grid for forecasts and observations, we used a budget algorithm (e.g., Accadia et al. 2003) to interpolate forecast precipitation to the ST4 grid (4.763-km horizontal grid spacing). Otherwise, metrics were computed from native grid output.

The following statistics were aggregated over all twenty-six 3-km forecasts initialized at 0000 UTC.

a. Precipitation climatologies

To assess precipitation climatologies, aggregate domain-total precipitation per grid point and fractional coverages of 1-h accumulated precipitation meeting or exceeding various

---

**Fig. 5.** Ensemble mean additive bias (model minus observations; short-dashed lines), ensemble mean RMSE (solid lines), total spread (long-dashed lines), and consistency ratio (CR; solid lines with circles) for (a) rawinsonde temperature (K), (b) rawinsonde zonal wind (m s\(^{-1}\)), (c) rawinsonde relative humidity (%), (d) aircraft temperature (K), (e) aircraft zonal wind (m s\(^{-1}\)), and (f) aircraft relative humidity (%) observations aggregated over all prior ensembles (1-h forecasts) between 0000 UTC 25 Apr and 0000 UTC 20 May 2017 (inclusive). These statistics were computed for those observations assimilated by both the 15- and 3-km EnKFs. Sample size at each pressure level is shown at the right of each panel. Vertical lines at x = 0 and x = 1 are references for biases and CRs, respectively.
accumulation thresholds (e.g., 2.5 mm h\(^{-1}\)) were calculated on native grids over the verification region. Additionally, spatial patterns of total precipitation over all 26 forecasts were examined, which were similar in the various ensembles and generally agreed with observations (e.g., Fig. 1b), including the southwest–northeast-oriented maximum across Missouri and adjacent areas. Although magnitudes of these maxima differed across the ensembles, these differences were manifested by the following domain-average statistics, so spatial variations of precipitation climatologies are not discussed further.

1) IMPACT OF ANALYSIS RESOLUTION

Differences between ensembles were largest over the first 12 h, when GEFS-initialized forecasts were spinning up precipitation from coarse 0.5° ICs. While this spinup meant GEFS-initialized forecasts underpredicted total precipitation and areal coverages (Fig. 7) and areal coverages (Fig. 8) over the first 5 h, ultimately, the spinup process yielded too much 6–12-h total precipitation and excessive coverages \(\geq 2.5\) mm h\(^{-1}\). Forecasts initialized from 15-km EnKF analyses also overpredicted total precipitation over the first 12 h, accompanied by excessive coverages for thresholds \(\geq 5.0\) mm h\(^{-1}\).

Overall, forecasts initialized from unblended 3-km EnKF analyses had precipitation climatologies best matching observations over 12 h, but there were shortcomings. For example, although at 1 h, unblended 3-km EnKF analyses produced forecasts with areal coverages closest to observations (Fig. 8), coverages rapidly decreased between 2 and 3 h and were further from those observed between 2 and 12 h for the 1.0 and 2.5 mm h\(^{-1}\) thresholds (Figs. 8a,b) compared to forecasts with 15-km or blended 3-km ICs, suggesting poor maintenance of stratiform precipitation regions after initialization. However, forecasts with unblended 3-km ICs had 6–12-h areal coverages at the 5.0 mm h\(^{-1}\) threshold well-matching observations (Fig. 8c) and 2–6-h coverages at the 10.0–50.0 mm h\(^{-1}\) thresholds closer to observations than forecasts with GEFS and 15-km EnKF ICs (Figs. 8d–f). Furthermore, 2–12-h domain-total precipitation was clearly best in forecasts with unblended 3-km ICs (Fig. 7).

Despite differences between the ensembles through 12 h, domain-total precipitation and areal coverages were broadly similar between 18 and 36 h, with too much total precipitation (Fig. 7) and general underprediction and overprediction of areal coverages at the 1.0 and 10.0–50.0 mm h\(^{-1}\) thresholds, respectively (Figs. 8a,d–f). Collectively, for precipitation

FIG. 6. Prior (1-h forecast) total spread (long-dashed lines) and ensemble mean additive bias (model minus observations; short-dashed lines) for (a) rawinsonde temperature (K), (b) rawinsonde zonal wind (m s\(^{-1}\)), (c) aircraft temperature (K), and (d) aircraft zonal wind (m s\(^{-1}\)) observations between 150 and 1000 hPa as a function of time. In (c) and (d) values are plotted every hour between 0000 UTC 23 Apr and 0000 UTC 20 May 2017 (inclusive) and smoothed with a 6-h running average, while in (a) and (b) values are plotted every 12 h between 0000 UTC 23 Apr and 0000 UTC 20 May 2017 (inclusive) without smoothing. These statistics were computed for those observations assimilated by both the 15- and 3-km EnKFs. The \(x\)-axis labels represent 0000 UTC for a specific month and day in 2017 (e.g., the marker for “0511” denotes 0000 UTC 11 May 2017). Dashed lines at \(y = 0\) are for reference.
Fig. 7. Average 1-h accumulated precipitation (mm) per grid point over all twenty-six 3-km forecasts and the verification region (CONUS east of 105°W) computed on native grids as a function of forecast hour. Red, blue, gold, and black shadings represent envelopes of the 10 members comprising the ensembles with 3-km EnKF ICs, 15-km EnKF ICs, GEFS ICs, and blended 3-km ICs, respectively, and darker shadings indicate intersections of two or more ensemble envelopes. Values on the y-axis represent ending forecast hours of 1-h accumulation periods (e.g., an x-axis value of 24 is for 1-h accumulated precipitation between 23 and 24 h). ST4 data during the 0–12- and 24–36-h forecast periods were identical except for 1 day (the former included data between 0000 and 1200 UTC 25 Apr–20 May while the latter instead included data between 0000 and 1200 UTC 26 Apr–21 May), and because domain-total ST4 precipitation between 0000–1200 UTC 21 May was much larger than that between 0000 and 1200 UTC 25 Apr, average 24–36-h domain-total ST4 precipitation was greater than average 0–12-h domain-total ST4 precipitation.

climatologies, these findings suggest benefits of convection-allowing analyses relative to convection-parameterizing analyses are primarily confined to short-term forecasts and heavier rainfall rates.

2) IMPACT OF BLENDING

With respect to forecasts initialized from unblended 3-km EnKF analyses, forecasts with blended 3-km ICs (using a 960-km cutoff) had similar 18–36-h areal coverages and total precipitation but higher domain-total precipitation and areal coverages over the first 6–12 h that typically compared worse to observations through 3 h (Figs. 7 and 8). Examination of individual forecasts indicated blended 3-km ICs mostly enhanced 1–3-h forecast precipitation within and near precipitation entities also predicted by forecasts with unblended 3-km ICs and that widespread spurious features did not cause the overprediction. This behavior is illustrated by the forecast initialized at 0000 UTC 1 May 2017, which had the largest difference of domain-total precipitation (e.g., Fig. 7) between member 1 in the CAEs with blended and unblended 3-km ICs across all twenty-six 36-h forecasts (Fig. 9). While both 1–3-h precipitation forecasts had similar spatial patterns, blended ICs led to more numerous cells in places with scattered rainfall, and these additional entities were usually erroneous compared to observations (black and gold circles in Fig. 9). Additionally, within features, the forecast with blended ICs had heavier rainfall maxima than ST4 observations and the forecast with unblended ICs (red circles in Figs. 9b,c,e,f,h,i).

Thus, overall, it appears blending did not improve short-term precipitation climatologies, likely due to imbalances created by blending (e.g., Yang 2005; H. Wang et al. 2014). Additional steps like digital filter initialization (DFI) applied to blended ICs (e.g., Yang 2005) may potentially lessen these imbalances, but DFI could result in spinups that are smoother than desirable for short-term high-resolution NWP model applications.

b. Ensemble precipitation verification

As in many studies, we used percentile thresholds to define events (e.g., the 95th percentile, which selects the top 5% of values), which removes bias and permits a thorough assessment of spatial performance given a model’s climate (e.g., Roberts and Lean 2008; Mittermaier and Roberts 2010; Mittermaier et al. 2013; Dey et al. 2014; Gowan et al. 2018; Woodhams et al. 2018; Schwartz 2019). Our application of percentile thresholds exactly followed section 5a(1) of Schwartz (2019), where physical thresholds corresponding to percentile thresholds were obtained separately for observations and each ensemble member on the ST4 grid for each precipitation accumulation interval. These physical thresholds were ultimately used to determine forecast and observed event occurrence. To help interpret subsequent objective statistics, mean physical thresholds corresponding to specific percentile thresholds are provided in Fig. 10. As with areal coverages (Fig. 8), the largest differences among the ensembles’ percentiles were over the first 6–12 h.

After interpolating precipitation forecasts to the ST4 grid, a “neighborhood approach” (e.g., Theis et al. 2005; Ebert 2008, 2009) was used to produce “neighborhood ensemble probabilities” (NEPs; Schwartz et al. 2010; Schwartz and Sobash 2017) that were ultimately verified. In short, NEPs were computed at the i-th grid point by averaging point-based ensemble probabilities over all grid points within the neighborhood of the i-th point, which incorporates spatial uncertainty and reflects the inherent inaccuracy of high-resolution NWP models at individual grid points. We produced NEPs for neighborhood length scales r between 5 and 200 km, which represented radii of circular neighborhoods. Please see section 2a of Schwartz and Sobash (2017) for more information about constructing and verifying NEPs and Eqs. (1)–(3) in Schwartz (2019), which explicitly describe NEP computation when using percentile thresholds.

Statistical significance testing followed section 5a(3) of Schwartz (2019). Specifically, a pairwise difference bootstrap technique with 10000 resamples was used to determine whether aggregate differences between two ensembles’ statistics were
statistically significant at the 95% level (e.g., Hamill 1999; Wolff et al. 2014).

1) ATTRIBUTES STATISTICS AND RANK HISTOGRAMS

To assess calibration, attributes diagrams (Wilks 2011) were produced with forecast probability bins of 0%–5%, 5%–15%, 15%–25%, ..., 85%–95%, and 95%–100%; curves on the diagonal indicate perfect reliability. Varying r changes sharpness and the resulting NEP distribution (Schwartz and Sobash 2017), which impacts reliability. Over the 1–12- and 18–36-h forecast periods, the smallest r yielding near-perfect reliability for any experiment was r = 90 km and r = 125 km, respectively, so we focus on reliability computed with those r.

Over the first 12 h for r = 90 km, the ensemble initialized from unblended 3-km EnKF analyses was statistically significantly more reliable than the ensembles initialized from GEFS and 15-km EnKF ICs, with the GEFS-initialized ensemble having the worst reliability (Fig. 11). Conversely, between 18 and 36 h for r = 125 km, the GEFS-initialized ensemble was regularly statistically significantly more reliable than the ensembles with unblended 15- and 3-km EnKF ICs, and the ensemble with 15-km ICs usually had comparable or better reliability than the ensemble with unblended 3-km ICs (Fig. 12). Except for the 99.9% threshold, all ensembles had skill with respect to forecasts of sample climatology.

These findings suggest aspects of GEFS ICs were beneficial for next-day (18–36-h) forecasts, which motivated blending GEFS and EnKF initial states. Indeed, blended 3-km ICs led to 18–36-h forecasts with comparable or better reliability as GEFS-initialized forecasts and statistically significantly better reliability than the ensemble with unblended 3-km ICs (Fig. 12). Over the first 12 h, differences between the ensembles with blended and unblended 3-km ICs were also often statistically significant, suggesting that blending can additionally improve short-term forecast reliability (Fig. 11).

Rank histograms (e.g., Hamill 2001) based on domain-total precipitation (e.g., Schwartz et al. 2014, 2020) corroborated...
attributes statistics. Specifically, over the first 12 h, bin counts in the ensemble with unblended 3-km ICs were closer to optimal in most bins compared to those for the ensembles with GEFS and 15-km EnKF ICs (Fig. 13a), which was quantified by the smaller-is-better reliability index (RI; Delle Monache et al. 2006). Blended 3-km ICs yielded slightly lower 1–12-h RIs than unblended 3-km ICs, but the difference was small compared to that between 18 and 36 h (Fig. 13b), where rank histograms and RIs indicated more observations fell within the ensemble and dispersion was improved when GEFS initial states were either
used as standalone ICs or combined with 3-km EnKF analyses through blending.

2) SPREAD AND SPECTRA

Improved reliability and rank histograms engendered by GEFS and blended 3-km ICs was associated with increased ensemble spread. In particular, the ensembles with GEFS and blended 3-km ICs had statistically significantly more 24–30-h precipitation spread compared to the ensembles with unblended EnKF ICs (Fig. 14). Additionally, blended 3-km ICs led to significantly more spread than unblended 3-km ICs over the first 6 h that may have improved reliability statistics and rank histograms, even though this enhanced spread reflected excessive early precipitation (e.g., Figs. 7–9). The greater spread through ~18 h in the ensembles with GEFS and 15-km ICs relative to that from the ensemble with unblended 3-km ICs may reflect a substantial contribution from the small, yet intense precipitation entities that were more numerically predicted when forecasts had downscaled, rather than 3-km, ICs (Figs. 8c–f and 10d–f).

To further understand spread characteristics, perturbation power spectra were computed with the discrete cosine transform (Denis et al. 2002), which is well suited for obtaining spectra from limited-area models. Perturbation spectra were determined with respect to the ensemble mean over the entire 3-km domain except for the 15 points nearest each lateral boundary. Final spectra were averaged over all 10 perturbations and 26 forecasts.

At 1 h, 500-hPa perturbation kinetic energy (PKE) in the ensemble with blended 3-km ICs broadly followed PKEs of the GEFS-initialized ensemble at scales $\geq 500$ km and the ensemble with unblended 3-km ICs at smaller scales, reflecting the blending procedure (Fig. 15a). Compared to the GEFS-initialized ensemble, the ensemble with unblended 3-km ICs had more 1-h forecast PKE at most scales (Fig. 15a), with enhanced large-scale power possibly reflecting upscale error growth with time through the continuous 3-km DA cycles. But, PKE in the GEFS-initialized ensemble grew fastest between 3 and 6 h (Figs. 15b,c) and was largest at all scales after 6 h (Figs. 15d–f).
while unblended 3-km ICs yielded the least 12–36-h PKE at scales > 100 km. Thus, more robust large-scale perturbation growth and kinetic energy in the GEFS-initialized ensemble was associated with its superior 18–36-h forecast reliability and rank histograms relative to the ensembles with unblended EnKF ICs. However, blending GEFS ICs with 3-km EnKF analyses promoted large-scale PKE growth after 6 h, and by 24–36 h, the ensembles initialized from GEFS and blended 3-km ICs had comparable large-scale PKEs, indicating blending successfully recovered these apparently favorable large-scale spectral characteristics that benefited reliability statistics and rank histograms.

### 3) FRACTIONS SKILL SCORES

Forecast skill was further evaluated with the fractions skill score (FSS; Roberts and Lean 2008), where FSS = 1 indicates a perfect forecast and FSS = 0 means no skill. We present FSSs for \( r = 100 \) km, although conclusions were unchanged when...
FSSs were computed with different neighborhood length scales. Moreover, areas under the relative operating characteristic curve (Mason 1982; Mason and Graham 2002) provided identical conclusions as FSSs and are not discussed further.

Forecasts initialized from unblended 3-km EnKF analyses had higher FSSs than those initialized from downscaled 15-km EnKF analyses through 6–12 h, both when aggregated over all forecasts (Fig. 16) and on an hour-by-hour basis (Figs. 17a–d), with many instances of significant differences. However, after 6–12 h, the ensembles with unblended 15- and 3-km EnKF ICs usually had statistically indistinguishable FSSs. Compared to the GEFS-initialized ensemble, the unblended EnKF-initialized ensembles had statistically significantly higher aggregate FSSs through 12–18 h but comparable or lower aggregate FSSs thereafter (Fig. 16), similar to attributes statistics. These 1–12-h forecast benefits from unblended 3-km EnKF ICs compared to GEFS ICs were evident for most hourly forecasts (Figs. 17i–l), while individual 1-h accumulated precipitation forecasts over the 18–36-h period from the ensemble with GEFS ICs were frequently comparable to or better than those from the ensemble with unblended 3-km EnKF ICs (Figs. 17m–p).

Blended 3-km ICs led to FSSs mirroring those from unblended 3-km EnKF ICs over the first 12–18 h (Fig. 16), indicating blending preserved short-term forecast benefits of increased analysis resolution for spatial placement. Furthermore, after 18–24 h, the ensemble with blended 3-km ICs had higher FSSs than the ensemble with unblended 3-km EnKF ICs both on an hourly basis (Figs. 17e–h) and in aggregate that were similar to or higher than FSSs from the GEFS-initialized ensemble.

4) SYNTHESIS

FSSs, attributes statistics, and rank histograms revealed clear benefits of convection-allowing analyses compared to convection-parameterizing analyses for 1–12-h precipitation forecasts, consistent with previous work (e.g.,
Johnson et al. 2015; Johnson and Wang 2016; Schwartz 2016; Gustafsson et al. 2018). But, these improvements from convection-allowing ICs did not persist to next-day forecast ranges, where GEFS-initialized forecasts outperformed EnKF-initialized forecasts. However, blended 3-km ICs led to similar or better 18–36-h forecasts than GEFS ICs, suggesting that blending large-scale fields from a global model with convection-allowing EnKF analyses can improve next-day CAE forecast dispersion, skill, and reliability while preserving short-term forecast benefits of increased IC resolution. Thus, when considering all forecast ranges, blending yielded initial ensembles that produced the best probabilistic forecasts.

c. Impact of hourly radar DA

Because our 3-km EnKF was highly constrained, we wondered whether assimilating radar observations could realize meaningful analysis and forecast improvements. So, to assess the impact of assimilating radar reflectivity observations, another EnKF was configured exactly as the nested 15-/3-km EnKF DA system (section 2b), except reflectivity observations throughout the CONUS were assimilated into 3-km analyses along with conventional observations hourly from 1900 to 0000 UTC. Although reflectivity observations could easily be assimilated more frequently in our framework, hourly radar DA mimics the HRRRE configuration. Backgrounds for 1900 UTC radar-assimilating EnKF analyses were provided by 1-h forecasts initialized from 1800 UTC posterior ensembles from the nested 15-/3-km EnKF assimilating solely conventional observations. Thus, the impact of assimilating reflectivity observations was confined to a 6-h period each day. This approach was adopted primarily to avoid the expense of continuously cycling another 3-km EnKF over the entire 4-week period. However, assimilating radar observations for just a few hours was methodologically consistent with numerous other high-resolution DA systems, including the WoF system (e.g., Wheatley et al. 2015; Jones et al. 2016; Skinner et al. 2018), and 6 h of assimilating
radar observations was more than sufficient to assess the data impact (e.g., Johnson and Wang 2017 and references therein).

Specific radar DA configurations mostly followed Duda et al. (2019) and references therein (Table 4), and like the other EnKFs, 0000 UTC analysis ensembles initialized 36-h, 3-km, 10-member CAE forecasts. Furthermore, to examine the interplay of blending and radar DA, we also created a set of ICs by blending GEFS ICs with radar-assimilating 3-km EnKF analyses using a 960-km filter cutoff. Assimilating reflectivity observations generally improved FSSs over the first 3 h but had small impacts thereafter (Fig. 18), similar to other studies finding short-lived benefits of radar DA (e.g., Kain et al. 2010; Johnson et al. 2015; Fabry and Meunier 2020). Within the radar-assimilating experiments, blending boosted FSSs at later times, as with the nonradar DA experiments. Assimilating reflectivity observations negligibly impacted attributes statistics, although assimilation of 100 000–200 000 radar observations each cycle lessened precipitation spread over the first hour (not shown). While more frequent assimilation cycles could potentially realize additional improvements from radar DA, it is unlikely that the small-scale information from radar observations can consistently yield forecast improvements after the shortest forecast ranges, especially in an EnKF highly constrained by other observations. Nonetheless, these experiments suggest feasibility of performing radar-assimilating, WoF-like analyses over large domains in a continuously cycling EnKF DA framework.

5. Summary and conclusions

EnKF DA systems with 80 members and 15- and 3-km horizontal grid spacings were continuously cycled with a 1-h period for 4 weeks over a computational domain spanning the entire CONUS. However, our EnKFs were highly constrained by observations, and whether convection-allowing EnKFs can be continuously cycled without deleterious consequences over large data-sparse domains is unclear.

At 0000 UTC, EnKF analyses initialized 36-h, 10-member CAE forecasts with 3-km horizontal grid spacing that were evaluated with a focus on precipitation. CAE forecasts were also initialized from NCEP’s operational GEFS and “blended” ICs produced by using a low-pass filter to combine large scales from GEFS ICs with small scales from EnKF analysis members. Precipitation forecasts initialized from continuously cycling EnKF analyses outperformed GEFS-initialized forecasts through 12–18 h, and benefits from initializing 3-km forecasts from corresponding 3-km analyses, rather than downscaled 15-km analyses, were
realized through 6–12 h. But, after 18 h, GEFS-initialized forecasts were comparable to or better than EnKF-initialized forecasts, indicating limitations of limited-area continuously cycling EnKFs as initialization tools for next-day CAE precipitation forecasts, consistent with Schwartz et al. (2020). Benefits of assimilating radar reflectivity observations into the 3-km EnKF were confined solely to 1–3-h forecasts.

Although blending sometimes degraded precipitation climatologies over the first 12 h, forecasts initialized from blended 3-km ICs reflected the respective strengths of both GEFS and 3-km EnKF ICs. Specifically, through 12–18 h, forecasts initialized from blended 3-km ICs had similar or better skill, reliability, and dispersion than those initialized from unblended 3-km EnKF analyses, while after 18–24 h, forecasts with blended 3-km ICs were comparable to or better than those with GEFS ICs. Therefore, blending produced ICs yielding the best performance when considering the entire 36-h forecast, indicating how combining large-scale global fields with high-resolution, limited-area EnKF analyses can potentially unify short-term WoF-like and next-day CAE guidance systems under a common framework.

There are many avenues for additional research and improvements. For example, while using identical inflation factors and observation errors in the 15- and 3-km EnKFs provided reasonable results, these choices may have been suboptimal. In particular, because observation errors are the sum of measurement and representativeness errors and representativeness errors are resolution dependent (e.g., Ben Bouallegue et al. 2020), observation errors should arguably be

---

**Fig. 16. Fractions skill scores (FSSs) over the verification region (CONUS east of 105°W) with a 100-km neighborhood length scale for the (a) 90th, (b) 95th, (c) 97.5th, (d) 99th, (e) 99.5th, and (f) 99.9th percentile thresholds aggregated over all twenty-six 3-km forecasts of 1-h accumulated precipitation as a function of forecast hour. Values on the x axis represent ending forecast hours of 1-h accumulation periods (e.g., an x-axis value of 24 is for 1-h accumulated precipitation between 23 and 24 h). The y-axis scales are different in each panel. Symbols along the top axis indicate forecast hours when differences between two ensembles were statistically significant at the 95% level as in Fig. 11 and denote the ensemble with statistically significantly higher FSSs. Note that the maximum FSS is 1.0; the area above 1.0 was added to make room for statistical significance markers.**
tuned for each domain, which, in turn, might require adjusting inflation factors. Thus, it may be possible to further improve our 3-km EnKF.

Additionally, our blending procedure did not impact the continuously cycling EnKF DA systems, and future work might assess whether incorporating large scales from global analyses into hourly limited-area DA cycles is beneficial. Furthermore, blending could potentially be optimized by dynamically determining the filter cutoff scale (e.g., Feng et al. 2020) or using height- and variable-specific cutoffs (e.g., Zhang et al. 2015), and efforts to mitigate blending-induced initial imbalances tailored for high-resolution models are needed. Moreover, next-day forecast benefits of blending suggest further exploring the value of mixed-resolution ensemble-based DA systems for convective applications may be worthwhile. Also, blending and partial cycling DA approaches should be compared; while both methods introduce external large-scale information into limited-area ICs, whether either method is

FIG. 17. (a)–(d) Histogram [expressed as probabilities (%)] of FSS differences with \( r = 100 \) km between the ensembles with 3-km EnKF ICs and 15-km EnKF ICs (3-km ICs minus 15-km ICs) computed from all twenty-six 0–1-, 1–2-, ..., 11–12-h 3-km forecasts of 1-h accumulated precipitation for the (a) 90th, (b) 95th, (c) 99th, and (d) 99.9th percentile thresholds. (e)–(h) As in (a)–(d), but for differences from all twenty-six 18–19-, 19–20-, ..., 34–35-, and 35–36-h 3-km forecasts of 1-h accumulated precipitation between the ensembles with blended and unblended 3-km ICs (blended 3-km ICs minus unblended 3-km ICs). (i)–(l), (m)–(p) As in (a)–(d) and (e)–(h), respectively, but for differences between the ensembles with unblended 3-km EnKF and GEFS ICs (3-km ICs minus GEFS ICs). Values on the \( x \) axis denote the leftmost points of each bin, and bin widths were 0.025 (e.g., the bars with left edges at 0.05 are for bins spanning 0.05–0.075). Colors of the bars correspond to the legend and indicate the experiment with the higher FSS in that bin.
preferable is unclear. It is also important to note that our blending methodology [Eq. (1)] changed the large-scale component of both the IC perturbations and the initial ensemble mean state, differing from an approach of blending perturbations derived from two different ensembles without changing the spectral representation of the initial ensemble mean (e.g., Caron 2013). Therefore, we cannot determine whether the 18–36-h forecast improvements from blending were due to altering the large-scale IC perturbations or large-scale initial ensemble mean, and it would be interesting to refine attribution in future work.

Finally, computing availability limited our cycling period to just 4 weeks, and additional experimentation is needed over longer periods, different seasons, and varied geographic

<table>
<thead>
<tr>
<th>TABLE 4. Settings for assimilation of radar reflectivity observations.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Radar observation source</strong></td>
</tr>
<tr>
<td>Horizontal localization full width</td>
</tr>
<tr>
<td>Vertical localization full width</td>
</tr>
<tr>
<td>Observation error standard deviation</td>
</tr>
<tr>
<td>Outlier check</td>
</tr>
<tr>
<td>Observation operator</td>
</tr>
<tr>
<td>Excluded observations</td>
</tr>
<tr>
<td>Assimilation of nonprecipitation observations</td>
</tr>
<tr>
<td>Minimum allowed forward operator value</td>
</tr>
</tbody>
</table>

FIG. 18. As in Fig. 16, but for comparisons focusing on the impact of assimilating radar reflectivity observations and aggregated over 24, rather than 26, forecasts (necessitated due to missing radar observations that precluded radar data assimilation sensitivity experiments for the forecasts initialized at 0000 UTC 13 and 14 May 2017). Gray curves are often beneath the black curves, especially in (a) and (b).
regions to further understand large-domain convection-allowing continuously cycling EnKF performance and whether benefits of blending are regime- and location-dependent. Nonetheless, this work suggests a combination of blending and high-resolution EnKF DA may represent a promising pathway toward an operational ensemble-based convection-allowing analysis-forecast system suitable for both nowcasting and next-day prediction over the CONUS.

Acknowledgments. This work was partially funded by NCAR’s Short-term Explicit Prediction (STEP) program and NOAA/OAR Office of Weather and Air Quality Grants NA17OAR4590182, NA17OAR4590114, and NA17OAR4590122. All forecasts were produced on NCAR’s Cheyenne supercomputer (Computational and Information Systems Laboratory 2017). Thanks to three anonymous reviewers for their constructive comments that improved this paper. The National Center for Atmospheric Research is sponsored by the National Science Foundation.

Data availability statement. Data supporting the findings of this work are available from the corresponding author. Specifically, raw precipitation forecasts from all numerical experiments are available from the corresponding author, and ST4 observations used for verification are available from https://doi.org/10.5065/D69Z93M3. Each of the 649 EnKF DA cycles required at least 250 GB of disk space, so data for each cycle were not saved. However, with GFS GRIB files (available from https://doi.org/10.5065/D65D8PWK) and model configuration and observation files (both available from the corresponding author), EnKF DA cycles can be reproduced. Files used to initialize the 36-h ensemble forecasts and the corresponding GFS files used for LBCs and blending were saved and are available upon request.

REFERENCES


