

Initial Conditions for Convection-Allowing Ensembles over the Conterminous United States

CRAIG S. SCHWARTZ, MAY WONG, GLEN S. ROMINE, RYAN A. SOBASH, AND KATHRYN R. FOSSELL

National Center for Atmospheric Research, Boulder, Colorado

(Manuscript received 16 December 2019, in final form 13 February 2020)

ABSTRACT

Five sets of 48-h, 10-member, convection-allowing ensemble (CAE) forecasts with 3-km horizontal grid spacing were systematically evaluated over the conterminous United States with a focus on precipitation across 31 cases. The various CAEs solely differed by their initial condition perturbations (ICPs) and central initial states. CAEs initially centered about deterministic Global Forecast System (GFS) analyses were unequivocally better than those initially centered about ensemble mean analyses produced by a limited-area single-physics, single-dynamics 15-km continuously cycling ensemble Kalman filter (EnKF), strongly suggesting relative superiority of the GFS analyses. Additionally, CAEs with flow-dependent ICPs derived from either the EnKF or multimodel 3-h forecasts from the Short-Range Ensemble Forecast (SREF) system had higher fractions skill scores than CAEs with randomly generated mesoscale ICPs. Conversely, due to insufficient spread, CAEs with EnKF ICPs had worse reliability, discrimination, and dispersion than those with random and SREF ICPs. However, members in the CAE with SREF ICPs undesirably clustered by dynamic core represented in the ICPs, and CAEs with random ICPs had poor spinup characteristics. Collectively, these results indicate that continuously cycled EnKF mean analyses were suboptimal for CAE initialization purposes and suggest that further work to improve limited-area continuously cycling EnKFs over large regional domains is warranted. Additionally, the deleterious aspects of using both multimodel and random ICPs suggest efforts toward improving spread in CAEs with single-physics, single-dynamics, flow-dependent ICPs should continue.

1. Introduction

Convection-allowing ensembles (CAEs) provide useful and valuable forecast guidance (e.g., Clark et al. 2012; Evans et al. 2014; Schwartz et al. 2019) and are now operational at many meteorological offices (e.g., Gebhardt et al. 2011; Peralta et al. 2012; Hagelin et al. 2017; Raynaud and Bouttier 2017; Jirak et al. 2018; Klasa et al. 2018). Nonetheless, uncertainty remains about optimal CAE design, especially regarding initial condition perturbations (ICPs),¹ which are needed to generate forecast diversity before perturbations from lateral boundary conditions (LBCs) and model error representation schemes engender substantial spread (e.g., Hohengger et al. 2008; Vié et al. 2011;

Peralta et al. 2012; Kühnlein et al. 2014; Romine et al. 2014; Zhang 2019).

There are several broad approaches for producing CAE ICPs. One simple method is to add random noise to a deterministic field (e.g., Hohengger and Schär 2007; Johnson et al. 2014; Raynaud and Bouttier 2016; hereafter RB16). While this operation is trivial, randomly produced ICPs are not flow dependent, a potential limitation.

Another straightforward method for initial condition (IC) generation is to downscale preexisting coarse-resolution analyses or short-term forecasts from either an ensemble or collection of deterministic numerical weather prediction (NWP) models directly onto the CAE grid (Jones and Stensrud 2012; Duc et al. 2013; Romine et al. 2014; Schumacher and Clark 2014; Schwartz et al. 2015a,b, 2019; Tennant 2015; Clark 2017; Schellander-Gorgas et al. 2017; Jirak et al. 2018; Klasa et al. 2018; Cafaro et al. 2019; Porson et al. 2019). Downscaling means the ICPs directly reflect the NWP model and data assimilation (DA) system underlying the coarse-resolution fields, and finescale details are not introduced into the CAE ICs.

¹ Herein, “perturbation” is defined as the difference between an individual ensemble member and a reference field, which is usually the ensemble mean.

Corresponding author: Craig Schwartz, schwartz@ucar.edu

DOI: 10.1175/MWR-D-19-0401.1

© 2020 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy](#) (www.ametsoc.org/PUBSReuseLicenses).

Alternatively, ICPs derived from coarser-resolution models can be recentered about a deterministic analysis of one's choosing, which could either be interpolated onto or produced directly on the CAE grid (Xue et al. 2007; Kong et al. 2008, 2009; Peralta et al. 2012; Kühnlein et al. 2014; Tennant 2015; RB16; Raynaud and Bouttier 2017; Hagelin et al. 2017). Thus, while these ICPs again reflect the external modeling system, the initial ensemble center could possess finescale structures that are imparted to individual ensemble members during recentering. However, recentering can be complex; European studies examining how recentering affects ensemble forecasts revealed mixed results (e.g., Lang et al. 2015; Tennant 2015; RB16), and recentering analysis ensembles about deterministic "hybrid" variational-ensemble analyses within DA contexts yields little impact (e.g., Clayton et al. 2013; Wang et al. 2013; Pan et al. 2014; Schwartz et al. 2015c).

Still another approach for generating CAE ICPs is to produce them directly on the CAE grid with an ensemble DA system, which provides flow-dependent ICPs fully consistent with the CAE forecast model that span all possible resolvable scales (e.g., Vié et al. 2011; Bouttier et al. 2012; Harnisch and Keil 2015; Wheatley et al. 2015; Johnson and Wang 2016; RB16; Keresturi et al. 2019). While this method is more sophisticated than and theoretically preferable to others, convective-scale DA is still evolving and computationally expensive.

Within each of these overarching methods, there are many options for producing CAE ICs: random noise can be generated in a variety of manners with different correlation length scales; coarse-resolution analyses are available from numerous NWP models with varied resolutions and DA methods; perturbations can be derived from and centered about many potential datasets; and myriad high-resolution DA implementations are possible. Moreover, these various approaches can be combined to produce CAE ICPs (e.g., Zhang 2018, 2019).

Yet, despite the multitude of options for CAE ICP generation, few studies have rigorously examined CAE forecast sensitivity to ICPs. Perhaps the most systematic study devoted to CAE ICPs was RB16, who found ICPs provided from both correlated random noise and a high-resolution, perturbed-observation variational DA system led to better CAE forecasts than ICPs from downscaled global ensemble analyses through 9–12 h. However, RB16 reported negligible sensitivity to ICP method for 12–36-h forecasts, presumably because LBC information quickly swept through their fairly small France-centered computational domain, and it is unclear how RB16's results may translate to larger domains that are less prone to LBC impacts (e.g., Warner et al. 1997; Romine et al. 2014;

Schumacher and Clark 2014), where sensitivity to ICPs may be detectable beyond 9–12 h.

In addition, several studies have assessed the suitability of limited-area ensemble Kalman filters (EnKFs; Evensen 1994; Houtekamer and Zhang 2016) for CAE initialization in case-study or idealized frameworks, although some did not fully isolate ICP impacts. For example, Harnisch and Keil (2015) suggested a convective-scale EnKF could initialize better CAE forecasts than downscaled ICPs for three forecasts, but forecast differences were not fully attributable to ICPs given discrepancies regarding DA and LBCs between various CAEs. Similarly, although Schumacher and Clark (2014) suggested an EnKF-initialized CAE sometimes outperformed a CAE initialized by downscaling and recentering non-EnKF perturbations about a deterministic analysis for a multiday heavy rainfall case, many differences between the CAEs also limited attribution to ICPs. Conversely, Johnson and Wang (2016) performed an idealized, controlled experiment and noted ICPs produced directly on a convection-allowing grid via EnKF DA led to modestly better 9-h precipitation forecasts than when ICPs were provided by coarser-resolution EnKF analyses, but their "perfect model" framework may not apply to many real-data situations.

More broadly, limited-area EnKFs are attractive for CAE initialization, as EnKFs seamlessly meld ensemble DA and forecasting in a single step to produce analysis ensembles that can initialize CAEs. Furthermore, continuously cycling EnKFs have become increasingly popular for real-time CAE forecast applications. For instance, between 2015 and 2017, the National Center for Atmospheric Research (NCAR) produced experimental, real-time CAE forecasts over the conterminous United States (CONUS) initialized with a continuously cycling EnKF (Schwartz et al. 2015b, 2019), and in 2017, Germany began using a continuously cycling EnKF to initialize their operational CAE (Schraff et al. 2016; Pantillon et al. 2018).

However, using continuously cycling limited-area EnKFs to initialize CAEs has risks, as biases can accumulate through assimilation cycles and degrade forecasts (e.g., Hsiao et al. 2012; Torn and Davis 2012; Romine et al. 2013; Wong et al. 2020); EnKFs over large domains like the CONUS might be more susceptible to this problem than EnKFs over comparatively small European domains where prominent LBC influences may mitigate bias accumulation. Therefore, although NCAR's experimental CAE forecasts were credible and widely adopted by both researchers and forecasters (Schwartz et al. 2019), it remains unclear whether large-domain, limited-area, continuously cycling EnKFs are optimal for producing CAE ICPs. Furthermore, objective assessments

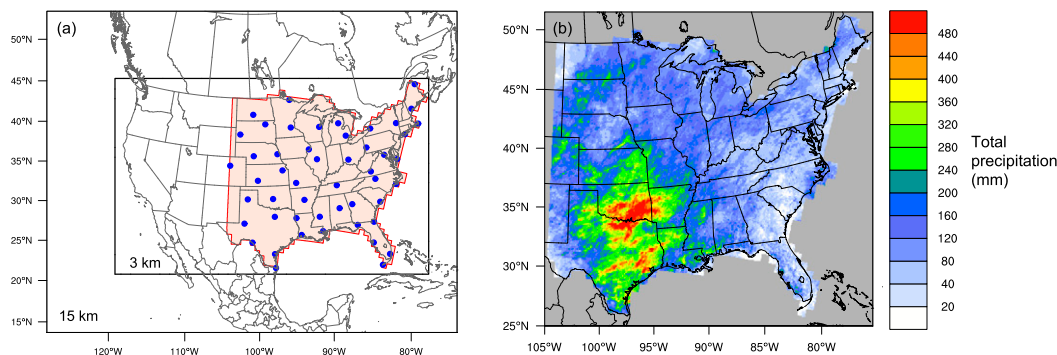


FIG. 1. (a) Computational domain. The horizontal grid spacing was 15 km in the outer domain (415×325 points) and 3 km in the nest (1581×986 points). Objective verification only occurred over the red shaded region of the 3-km domain (CONUS east of 105°W). Blue dots denote locations of rawinsonde observations used for verification. (b) Total accumulated Stage IV (ST4) precipitation (mm) between 0000 UTC 1 May and 0000 UTC 2 Jun 2015 over the verification region. This accumulation period encompassed all possible valid times of the model forecasts.

of systematic, controlled experiments designed to isolate impacts of ICPs on real-data CAE forecasts over the CONUS have yet to be reported, although subjective evaluations of two CAEs differing solely by ICPs performed during NOAA's 2019 Hazardous Weather Testbed Spring Forecasting Experiment suggested ICPs had little impact on severe weather forecasts (Clark et al. 2019).

Accordingly, to further understanding about ICP methods for CAEs, including EnKF-based approaches, this study systematically examined 31 48-h forecasts from several 10-member CAEs over the CONUS, where many of the CAEs differed solely by their ICPs. In addition, to explore the impacts of recentering ICPs, other CAEs differed solely by their central initial states. Thus, differences between various CAE forecasts were fully attributable to either ICPs or central initial state, providing insight about CAE initialization and design that has implications for development of future operational CAEs, such as those at NOAA under the Unified Forecast System (UFS) framework.

2. Model configurations and ICP strategies

a. Model configurations

All CAEs employed forecast model configurations similar to those used in NCAR's real-time CAE project (Schwartz et al. 2015b, 2019). Specifically, 48-h forecasts were produced by version 3.6.1 of the Advanced Research Weather Research and Forecasting (WRF) Model (Skamarock et al. 2008; Powers et al. 2017) over a two-way nested domain spanning the CONUS and adjacent areas (Fig. 1a). The horizontal grid spacing was 15 km in the outer domain and 3 km in the nest, where

time steps were 75 and 18.75 s, respectively. Both domains had 40 vertical levels, a 50-hPa top, and used common physical parameterizations (Table 1), except no cumulus parameterization was employed on the 3-km grid. All ensemble members used identical physics and dynamics.

During the 48-h integrations, LBCs were produced for each ensemble member by perturbing forecasts from NCEP's Global Forecast System (GFS) with random, correlated, Gaussian noise with zero mean (e.g., Barker 2005; Torn et al. 2006) drawn from the default "cv3" background error covariances (BECs) provided by the WRF Model's DA system (WRFDA; Barker et al. 2012), which were produced with the "NMC method" (Parrish and Derber 1992) based on differences between 48- and 24-h forecasts from a legacy ~ 100 -km configuration of the GFS model. Following Schwartz et al. (2015b), LBC perturbation magnitudes linearly increased throughout the forecasts to promote spread. Identical LBC sets were used for all CAEs.

b. IC generation and experimental design

Five sets of 10-member, 48-h CAE forecasts were produced over May 2015, which was the wettest month ever recorded over the CONUS (e.g., Blunden and Arndt 2016) and featured a broad precipitation maximum over the central CONUS (Fig. 1b). The CAEs were identical except for their ICs, which are now described.

1) CONTINUOUSLY CYCLING ENKF DA

Two CAEs had ICPs derived from an experimental continuously cycling ensemble adjustment Kalman filter (Anderson 2001, 2003; Anderson and Collins 2007), a

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

Physical parameterization	WRF Model option	References
Microphysics	Thompson	Thompson et al. (2008)
Longwave and shortwave radiation	Rapid Radiative Transfer Model for Global Climate Models (RRTMG) with ozone and aerosol climatologies	Mlawer et al. (1997), Iacono et al. (2008), and Teegen et al. (1997)
Planetary boundary layer	Mellor–Yamada–Janjić (MYJ)	Mellor and Yamada (1982) and Janjić (1994, 2002)
Land surface model	Noah	Chen and Dudhia (2001)
Cumulus parameterization	Tiedtke (15-km domain only)	Tiedtke (1989) and Zhang et al. (2011)

type of EnKF, implemented in the Data Assimilation Research Tested (DART) software (Anderson et al. 2009). The EnKF DA system had 80 ensemble members and produced analyses solely on the 15-km domain (Fig. 1a).

Similar to Schwartz et al. (2015a), an initial 15-km ensemble was created by adding random noise drawn from the WRFDA-provided BECs to the 0.25° 0000 UTC 26 April 2015 GFS analysis; this randomly generated ensemble served as the prior (before assimilation) ensemble for the first EnKF analysis. Then, the 0000 UTC 26 April 2015 posterior (after assimilation) ensemble initialized a 6-h, 80-member ensemble forecast that became the prior ensemble for the second EnKF analysis at 0600 UTC 26 April 2015, and this analysis–forecast cycle with a 6-h period continued until 0000 UTC 31 May 2015. Model configurations and LBC perturbation strategies during the 6-h forecast steps were identical to those described in section 2a, except forecasts were not produced on the 3-km grid. Soil states evolved freely and independently for each member during the entire cycling period, and the first 5 days of cycling were considered spinup and discarded.

The EnKF assimilated conventional observations as described by Schwartz et al. (2015b), with the addition of global positioning system radio occultation refractivity observations. Observation errors, preprocessing, and quality control were also detailed by Schwartz et al. (2015b) and included an “outlier check” to reject observations far from the prior ensemble mean, inflating errors of observations near lateral boundaries, rejecting surface observations with mismatched modeled and observed terrain heights, and superrobbing aircraft and satellite wind observations.

Specific DA settings were mostly similar to those employed by NCAR during their real-time CAE project (Schwartz et al. 2015b, 2019) and summarized in Table 2. Compared to NCAR’s real-time EnKF analyses produced in May 2015 (Schwartz et al. 2015b), the biggest differences involved ensemble size and covariance inflation. Specifically, the 80 members and posterior relaxation-to-prior-spread (RTPS) inflation (Whitaker

and Hamill 2012) differed from the real-time analyses, which had 50 members and used prior adaptive inflation (e.g., Anderson 2009). The switch to RTPS inflation was based on systematic experimentation finding little precipitation forecast sensitivity to inflation method, and because RTPS inflation is simpler, we chose it for this work. Ultimately, the EnKF configuration herein was well tuned in a spread–skill sense (Houtekamer et al. 2005) and initialized significantly better precipitation forecasts than the real-time EnKF analyses (not shown), likely due to the larger ensemble size, which benefits EnKFs (e.g., Zhang et al. 2013; Houtekamer et al. 2014).

For each 0000 UTC EnKF analysis between 1 and 31 May 2015 (inclusive), the first 10 15-km analysis members (i.e., members 1–10) initialized 48-h forecasts on the nested grid (Fig. 1a), where 3-km ICs were downscaled from the 15-km analyses that lacked storm-scale structures. Because the EnKF can be conceived as separately updating a mean and perturbations about the mean, our EnKF-based ICPs were centered about 80-member ensemble mean EnKF analyses (“EnKF_{EnKF}”; Table 3). On average, each EnKF member was equally likely to be closest to “truth”, so choosing the first 10 members to initialize 48-h forecasts was analogous to randomly picking 10 members from full 80-member analysis ensembles (e.g., Schwartz et al. 2014, 2019), and 10-member CAEs can provide skillful and valuable forecasts (e.g., Clark et al. 2011, 2018; Schwartz et al. 2014).

In addition, between 1 and 31 May 2015 (inclusive), 0000 UTC perturbations of zonal and meridional wind, potential temperature, water vapor mixing ratio, and perturbation geopotential and dry surface pressure (U , V , θ , q_v , ϕ , and μ , respectively) from analysis members 1–10 were added to corresponding 0000 UTC 0.25° GFS analyses to create another set of ICs that initialized 48-h forecasts on the nested grid (“GFS_{EnKF}”; Table 3). These ICs had identical U , V , θ , q_v , ϕ , and μ perturbations as EnKF_{EnKF} but were centered on GFS analyses, rather than EnKF mean analyses (Table 3), providing insight about sensitivity

TABLE 2. Configuration details of the DART-based EnKF DA system.

Parameter	Setting	References
Ensemble size	80 members	
Filter type	Ensemble adjustment Kalman filter (EAKF)	Anderson (2001, 2003) and Anderson and Collins (2007)
Updated WRF Model variables	Zonal and meridional wind components; perturbation geopotential height, potential temperature, and dry surface pressure; water vapor mixing ratio; cloud water, cloud ice, rain, snow, and graupel mixing ratios; rain and cloud ice number concentrations; diabatic heating	
Covariance localization function	Gaspari–Cohn	Eq. (4.10) of Gaspari and Cohn (1999)
Horizontal localization full-width	1280 km	
Vertical localization full-width	1.0 scale height, except 2.0 scale heights for surface pressure observations	
Adaptive localization threshold	2000 observations	Bishop and Hodyss (2009a,b) and Schwartz et al. (2015b)
Covariance inflation method	Posterior relaxation-to-prior-spread (RTPS)	Whitaker and Hamill (2012)
Inflation factor	1.12	Whitaker and Hamill (2012), Schwartz and Liu (2014), and Schwartz (2016)
Sampling error correction	True	Anderson (2012)

to IC center and suitability of large-domain regional continuously cycling EnKFs to initialize CAE forecasts. Compared to the EnKF analyses, GFS analyses had coarser resolution but assimilated many more observations, including satellite radiances, and reflect a well-tuned, operational deterministic forecast system. Moreover, standard WRF Model preprocessing discards GFS hydrometeor analyses, such that CAEs with GFS initial centers started with no (zero value) hydrometeors, contrasting those CAEs with EnKF mean initial centers (Table 3).

2) RANDOM ICPs

Performing limited-area EnKF DA can be expensive, so other cheaper, pragmatic methods of producing CAE ICPs were also explored. Thus, two additional sets of 10-member 48-h forecasts on the nested grid (Fig. 1a) were initialized by taking random draws of $U, V, \theta, q_v, \phi,$ and μ from the WRFDA-provided

BECs and adding them to both 0000 UTC 15-km EnKF mean (“EnKF_{RAND}”) and 0.25° GFS analyses (“GFS_{RAND}”) between 1 and 31 May 2015 (inclusive; Table 3). The random patterns differed for each initialization, but, for a particular initialization and ensemble member, identical random perturbations were added to both EnKF mean and GFS analyses (i.e., member 1 of EnKF_{RAND} and GFS_{RAND} had identical perturbations).

Length scales and variances of random perturbations can be tuned when drawing from BECs, providing many possibilities for specifying initial correlated random noise. However, we only used one set of tuning parameters where the length scales were empirically reduced by ~85% from those within the WRFDA-provided BECs, which was necessary because of our much finer grid spacing compared to the ~100-km statistics contained in the BECs. Variances were also reduced relative to those in the WRFDA-provided

TABLE 3. Description of the various CAEs in terms of their initial centers, IC perturbation methods, and initial hydrometeors.

Ensemble name	IC center	IC perturbations	Initial hydrometeors?
EnKF _{EnKF}	EnKF mean analyses	EnKF DA	Yes, with spread
EnKF _{RAND}	EnKF mean analyses	Correlated random noise	Yes, but no spread; hydrometeor values equal to the ensemble mean mixing ratios
GFS _{EnKF}	GFS analyses	EnKF DA	No
GFS _{RAND}	GFS analyses	Correlated random noise	No
GFS _{SREF}	GFS analyses	3-h SREF forecasts	No

BECs in attempt to roughly approximate spread of the other initial ensembles. Ultimately, our randomly produced ICPs had mesoscale structures and contrasted RB16, who used a similar method to produce random CAE ICPs but with convective-scale structures. Although subsequent CAE forecasts may be sensitive to length scale and variance of initial random noise, examining this sensitivity was beyond the scope of this work, and the primary purpose of constructing random ICPs was to assess whether they yielded comparable forecast quality as flow-dependent EnKF ICPs, but with substantially lower costs, by examining relative performances of CAEs with the same initial center but different ICPs (e.g., EnKF_{RAND} versus EnKF_{EnKF} and GFS_{RAND} versus GFS_{EnKF}; Table 3).

3) SREF ICPs

An additional IC set was produced by adding perturbations of U , V , θ , q_v , ϕ , and μ derived from 2100 UTC–initialized 3-h forecasts of NCEP’s Short-Range Ensemble Forecast (SREF; Du et al. 2014) system to 0000 UTC 0.25° GFS analyses (“GFS_{SREF}”; Table 3); these ICs then initialized 48-h forecasts on the nested grid between 1 and 31 May 2015 (inclusive). This inexpensive method was very similar to that used by the Center for Analysis and Prediction of Storms to produce CAE ICPs for many years (e.g., Xue et al. 2007; Kong et al. 2008, 2009; Gallo et al. 2017). Like EnKF perturbations, SREF perturbations were flow-dependent, and although the SREF system had 16-km horizontal grid spacing, data available to us had been coarsened to 32 km.

During the experimental period (May 2015) the SREF contained 21 members with diversity provided by varied dynamic cores, physics, and ICs (Du et al. 2014).² However, we only needed perturbations from 10 members, which were chosen as the 8 SREF members used to initialize the National Severe Storms Laboratory’s experimental CAE (Clark 2017) plus 2 additional members based on the Advanced Research WRF dynamic core (the “p3” and “n3” SREF members). Contrasting the single-physics, single-dynamics EnKF ICPs, these 10 SREF-based ICPs collectively reflected three dynamic cores, each associated with its own unique IC generation method, and, moreover, some physics schemes varied across SREF members with a common core (Du et al. 2014). Thus, below we refer to SREF ICPs as “multimodel ICPs,” with the understanding that differences between GFS_{SREF} members cannot be fully attributed to dynamic core,

physics, or initialization method encapsulated and entangled within their ICPs.

3. ICP characteristics and spread growth

a. Initial spread characteristics

Mean 700-hPa zonal wind spread over all 31 0000 UTC initial 15-km ensembles highlighted differences between the various ICPs. Specifically, EnKF and SREF ICPs were flow dependent (Figs. 2a,b), with relatively large spread associated with stronger mean height gradients that portend uncertainty, such as over eastern Canada and the central CONUS, and comparatively small spread associated with weaker height gradients over the southeast CONUS and West Coast. Conversely, random ICPs were not flow dependent and yielded nearly uniform mean spread reflecting the tuned BECs (Fig. 2c).

Consistent with Figs. 2a and 2b, EnKF- and SREF-based initial ensembles had comparable spreads for wind (Figs. 3a,b), while SREF-based initial ensembles had larger spread than EnKF-based initial ensembles for temperature and moisture below 250 hPa (Figs. 3c,d); these larger SREF spreads for thermodynamic variables were possibly manifestations of diverse precipitation patterns produced by the multiple models in the unconstrained 3-h SREF forecasts leveraged to obtain ICPs. Except for jet stream level, randomly produced initial ensembles had broadly comparable spreads as those with SREF and EnKF perturbations for wind but with smoother vertical structures (Figs. 3a,b). However, random ICPs had more midtropospheric temperature and low-level moisture spread than the other ICPs (Figs. 3c,d).

b. Spread evolution

1) SPREAD AND ERROR AT RAWINSONDE LOCATIONS

Ensemble mean RMSEs with respect to rawinsonde observations and standard deviations at rawinsonde locations were computed to assess spread and error growth. The ensembles with EnKF ICPs (EnKF_{EnKF} and GFS_{EnKF}; Table 3) had similar spreads throughout the forecast (Fig. 4), but the CAE initially centered about EnKF mean analyses (EnKF_{EnKF}; gray curves) had smaller RMSEs than the CAE initially centered about GFS analyses (GFS_{EnKF}; blue curves) at the initial time, indicating the EnKF (section 2b) fit rawinsonde observations more closely than GFS analyses. However, EnKF_{EnKF} ensemble mean RMSEs grew quickly and were statistically significantly worse

² The SREF configuration has since changed (e.g., Du et al. 2015).

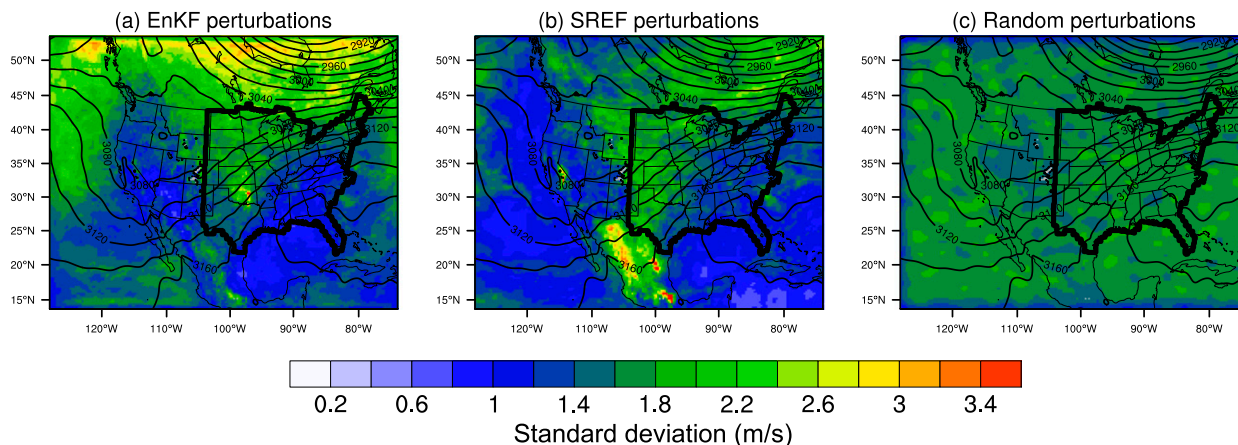


FIG. 2. Average standard deviation of 700-hPa zonal wind (m s^{-1}) over all 10-member, 15-km 0000 UTC initial ensembles between 1 and 31 May 2015 (inclusive) constructed with (a) EnKF, (b) SREF, and (c) random perturbations. Mean 700-hPa height (m; contoured every 20 m) over all 0000 UTC Global Forecast System (GFS) analyses between 1 and 31 May 2015 (inclusive) is overlaid on each panel, and the verification region (CONUS east of 105°W) is outlined.

than those from GFS_{EnKF} after initialization, indicating forecast sensitivity to initial center and suggesting GFS analyses were overall better than EnKF mean analyses.

The three CAEs with GFS initial centers (GFS_{EnKF} , GFS_{RAND} , GFS_{SREF} ; Table 3) typically had similar RMSEs but different spreads (Fig. 4), with initial spreads generally aligned with Figs. 3a–c. In particular, except for 500- and 300-hPa wind, EnKF ICPs had the smallest initial spread at rawinsonde locations due to the restorative effect of assimilating those very observations, and GFS_{EnKF} spread was smaller than GFS_{RAND} and GFS_{SREF} spread from 12 to 48 h.

While GFS_{EnKF} and GFS_{SREF} spread sometimes grew more than GFS_{RAND} spread over the entire 48-h forecast, over the first 12 h, GFS_{SREF} spread usually grew faster than GFS_{EnKF} spread and GFS_{RAND} spread growth rates were typically highest (Fig. 4). Although GFS_{RAND} initial spread was often relatively large, even when GFS_{RAND} initial spread was comparable to or smaller than that of the other ensembles, rapid spread growth still occurred over the first 12 h (Figs. 4a,b,f,g,j), suggesting GFS_{RAND} forecast spread was not simply modulated by its initial spread.

2) PERTURBATION POWER SPECTRA

To further understand spread growth characteristics over the first 12 h, perturbation power spectra were computed using the discrete Fourier transform after applying a Hanning window (e.g., Harris 1978) to enforce periodicity. Random 2-m temperature and 10-m wind ICPs had less power than SREF and EnKF ICPs for scales <500 and 250 km, respectively (Figs. 5a,f), reflecting the specified length scales used to construct random noise. However, random ICPs led to rapid error

growth over the first hour (Figs. 5b,g), with larger growth rates than EnKF and SREF ICPs at small scales, suggesting rapid GFS_{RAND} spread increases over the first 12 h (e.g., Figs. 4a,b,f–j) were driven by small-scale perturbation growth ultimately spurred by downscale propagation of random mesoscale errors (e.g., Durran and Gingrich 2014). After the first hour, GFS_{RAND} error growth rates were much slower (Figs. 5c–e,h–j), but by 12 h, at all scales GFS_{RAND} had the most perturbation energy and GFS_{EnKF} the least (Figs. 5e,j), consistent with greater low-level GFS_{RAND} 12-h forecast spread compared to GFS_{EnKF} (Figs. 4a,f).

Overall, these spectra illustrate rapid GFS_{RAND} error growth was insensitive to ICP variance magnitude; GFS_{RAND} surface temperature spread was relatively small (Fig. 3c) while its surface wind spread was relatively large (Figs. 3a,b), yet rapid GFS_{RAND} error growth occurred over the first hour for both variables. Similar evolutions were evident for other vertical levels, and after 12 h, spectra from all three ensembles gradually converged as common LBCs exerted their influence (not shown).

3) PRECIPITATION SPREAD

Precipitation development and spread over the first 18 h was sensitive to initial spread characteristics. Most notably, precipitation variances (about each ensemble’s mean) were largest in the two CAEs with random ICPs, with rapid spread increases over the first 6 h (Fig. 6a) consistent with fast low-level error growth (Figs. 4a,b,f,g, 5). Comparatively, precipitation variances were less sensitive to IC center, although the CAEs initially centered about GFS analyses had less spread than those initially centered about EnKF mean

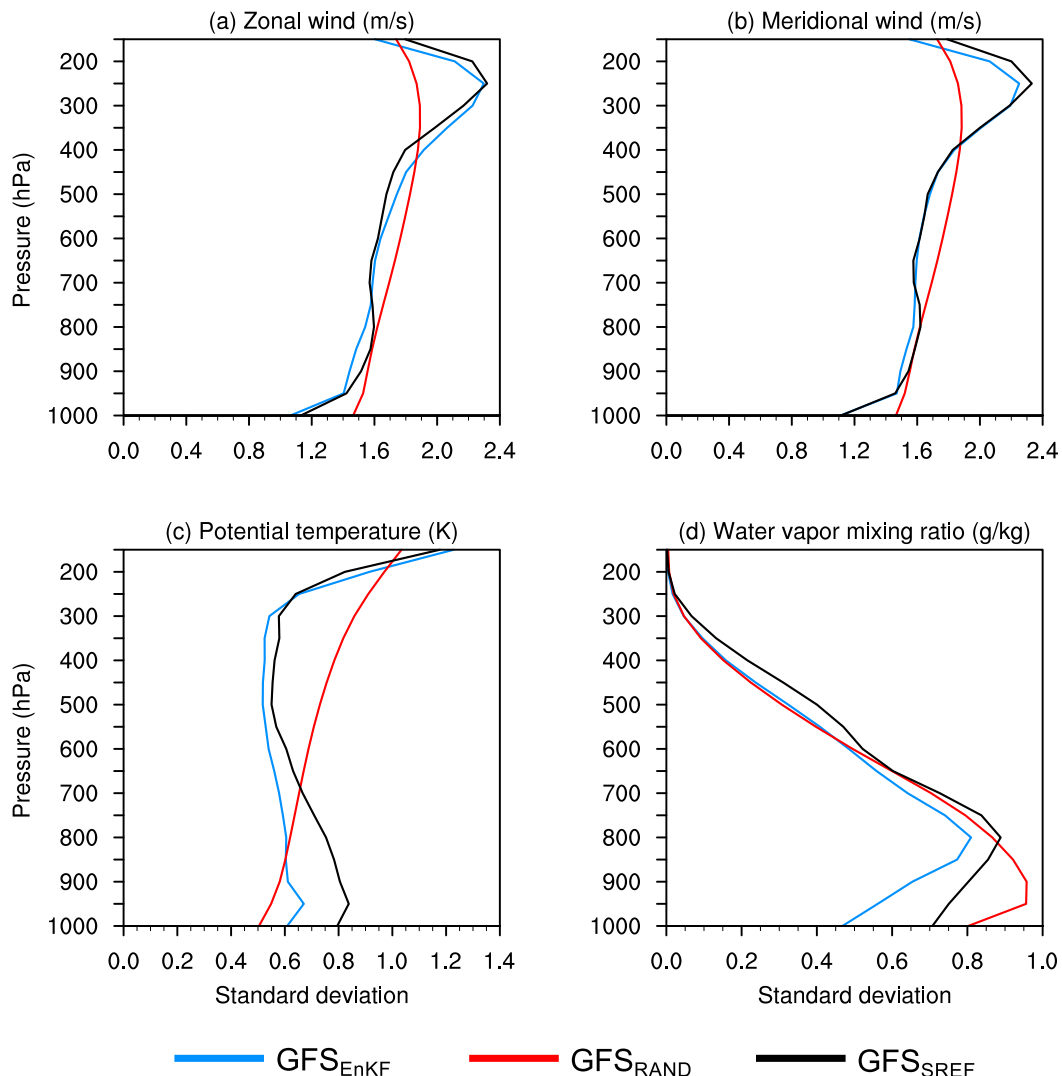


FIG. 3. Standard deviation of (a) zonal wind (m s^{-1}), (b) meridional wind (m s^{-1}), (c) potential temperature (K), and (d) water vapor mixing ratio (g kg^{-1}) as a function of pressure (hPa) over the verification region (CONUS east of 105°W) averaged over all 10-member, 15-km 0000 UTC initial ensembles between 1 and 31 May 2015 (inclusive).

analyses through 18 h, possibly because initial nonzero hydrometeor states in the CAEs initially centered about EnKF mean analyses (Table 3) contributed to spread growth over the first few hours. GFS_{SREF} had larger spread than GFS_{EnKF} through 24 h, possibly due to the multiple models reflected in SREF-based ICPs and generally consistent with greater GFS_{SREF} initial spread (Figs. 3c,d, 4) and spread growth over the first 12 h (Fig. 4). Variances computed after a bias correction (see section 4b) generally behaved similarly as uncorrected variances over the first 12–18 h but with smaller differences among the ensembles (Fig. 6b).

After 18 h, precipitation variances were more similar across all five CAEs than at earlier times. However, of

the three CAEs initially centered about GFS analyses, GFS_{SREF} had the most spread between 24 and 33 h for raw variances (Fig. 6a), while bias-corrected variances indicated more spread from random and SREF ICPs between 24 and 42 h compared to EnKF ICPs (Fig. 6b).

c. Forecast example

The forecast initialized at 0000 UTC 11 May 2015 nicely illustrates how different ICPs impacted spread growth. At this time, precipitation was ongoing in the vicinity of tropical depression Ana over southeastern North Carolina and along surface boundaries stemming from a low pressure center over South Dakota, which

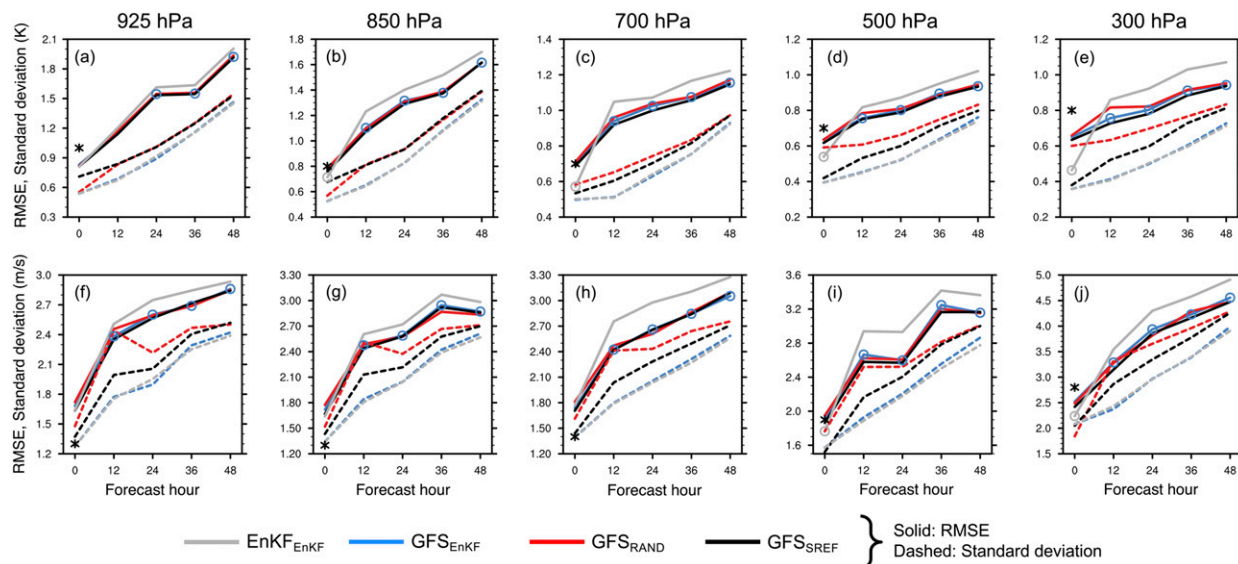


FIG. 4. (a) 925-, (b) 850-, (c) 700-, (d) 500-, and (e) 300-hPa temperature (K) ensemble mean RMSE compared to rawinsonde observations (solid lines) and standard deviation at rawinsonde locations (dashed lines) aggregated over all 31 3-km forecasts over the verification region (CONUS east of 105°W) as a function of forecast hour. See Fig. 1a for locations of rawinsonde observations. Statistical significance between EnKF_{EnKF} and GFS_{EnKF} RMSEs was determined by a bootstrap resampling technique (see section 4b) and denoted by open circles placed on the curve with the significantly lower RMSE; for example, a blue circle on a blue line indicates GFS_{EnKF} had a statistically significantly lower RMSE than EnKF_{EnKF} at the 95% level. (f)–(j) As in (a)–(e), but for wind speed (m s⁻¹). Black asterisks represent observation error standard deviations.

was associated with an upper-level trough over the Rockies and adjacent plains (Figs. 7s–u). Initial 2-m temperature EnKF and SREF perturbations indicated flow dependence with enhanced spread around these features (Figs. 7a,g,s,t), whereas random perturbations did not reflect these phenomena (Figs. 7m,u).

The 2-m temperature perturbation magnitudes were initially small (Figs. 7a,g,m,s–u), but by 1 h, spread substantially increased. At 1 h, GFS_{EnKF} and GFS_{SREF} spread primarily reflected the surface low pressure system and attendant fronts (Figs. 7b,h), while GFS_{RAND} had not fully developed flow-dependent characteristics (Fig. 7n). However, after 3 h, all spread patterns reflected synoptic-scale features (Figs. 7c,i,o), and by 6–12 h, the three ensembles had comparable structures near the fronts (Figs. 7d–f,j–l,p–r), with GFS_{SREF} spread highest along the boundaries. Conversely, in the weak forcing regime over the southeastern CONUS and Ohio Valley, GFS_{RAND} possessed much more spread than GFS_{SREF} and GFS_{EnKF} that peaked from 6 to 9 h (Figs. 7d,e,j,k,p,q), consistent with Johnson et al. (2014), who suggested random noise was most likely to promote spread growth in weak forcing scenarios. It appears that these random spread patterns were initially organized on small scales (Figs. 7n,o), consistent with perturbation spectra indicating rapid small-scale error growth over the first several hours (Figs. 5a–c,f–h).

Regarding precipitation, GFS_{EnKF} had more spread than GFS_{RAND} and GFS_{SREF} at 1 h (Figs. 8a,f,k), and although GFS_{RAND} 2-m temperature structures were not fully flow-dependent at this time (Fig. 7n), GFS_{RAND} precipitation spread represented flow-dependent features (Fig. 8k). This finding was similar to RB16, who noted flow-dependent precipitation structures quickly developed in a CAE with storm-scale random ICPs. By 3 h, precipitation spread had grown substantially in all CAEs (Figs. 8b,g,l), and by 6–12 h, there were generally more enhanced and wider areas of nonzero spread in GFS_{RAND} than GFS_{EnKF} in the vicinity of frontally forced precipitation (Figs. 8c–e,h–j,m–o). Consistent with Fig. 6a, GFS_{RAND} spread peaked at 6 h, and while GFS_{SREF} and GFS_{EnKF} had similar patterns, there was slightly more GFS_{SREF} spread from 3 to 12 h.

Between 3 and 9 h, GFS_{RAND} precipitation spread was particularly large over the weakly forced southeast CONUS and Ohio Valley, whereas GFS_{EnKF} and GFS_{SREF} had much less spread in similar locales (Figs. 8b–d,g–i,l–n). Areas of enhanced GFS_{RAND} precipitation spread often appeared to be preceded by relatively large GFS_{RAND} 2-m temperature perturbations (Figs. 7m–p) and were accompanied by low probabilities of precipitation over wide areas where observed precipitation did not occur (not

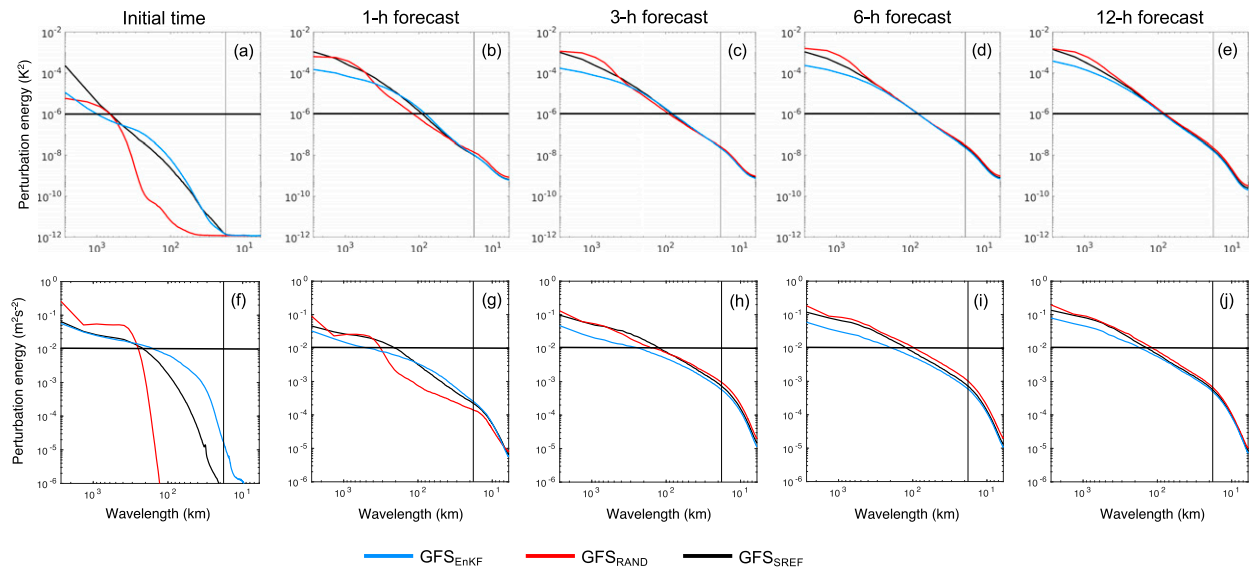


FIG. 5. (a)–(e) Average perturbation energy for 2-m temperature (K^2) as a function of wavelength (km) computed over all 31 3-km forecasts and the 3-km domain east of $105^\circ W$, excluding 16 and 42 grid points from the eastern and northern/southern boundaries, respectively, for (a) analyses and (b) 1-, (c) 3-, (d) 6-, and (e) 12-h forecasts. (f)–(j) As in (a)–(e), but for 10-m kinetic energy ($m^2 s^{-2}$). Vertical lines denote 6 times the horizontal grid spacing (3 km), the approximate effective resolution of the forecasts (Skamarock 2004), and horizontal lines are for reference to help visualize changes across forecast hours. Note that the initial conditions had 15-km horizontal grid spacing, which is manifested by lack of small-scale power in (a) and (f).

shown). Thus, at least for this case, random ICPs led to false alarms in some members in areas with weak forcing.

d. Summary

The preceding analyses suggest random ICPs promoted rapid short-term error growth, primarily driven by small-scale perturbations (Fig. 5), while EnKF and SREF ICPs had comparatively slower error growth rates. Accordingly, compared to the other ICP strategies, random ICPs generally yielded more spread over the first 12–18 h (Figs. 4, 6). As the next section shows, this additional spread from random ICPs was sometimes helpful, yet did not always possess favorable characteristics.

4. Precipitation verification

Hourly accumulated precipitation forecasts were objectively compared to NCEP’s Stage IV (ST4) analyses (Lin and Mitchell 2005) over the CONUS east of $105^\circ W$ (Fig. 1a), where ST4 data were most robust (e.g., Nelson et al. 2016) and considered as “truth.” While some statistics were computed on native grids, many verification metrics require a common grid for forecasts and observations. So, for these metrics, all precipitation forecasts were interpolated to the ST4 grid (4.763-km horizontal grid spacing) using a precipitation-conserving budget

interpolation algorithm (e.g., Accadia et al. 2003). We primarily focused on precipitation because it is an important sensible weather field and depends on many physical processes, thus providing an overall summary of model performance.

Statistics presented in this section are aggregated over all 31 forecasts.

a. Bias characteristics

1) TOTAL PRECIPITATION

Total precipitation over the verification region (Fig. 1a) normalized by number of grid points in the verification region was determined for each member on native grids (Fig. 9). To concisely summarize results, only the mean and range (maximum minus minimum; lines with circle markers) are shown for all five CAEs (Fig. 9a), while individual GFS_{SREF} members are shown in Fig. 9b.

The largest differences between the CAEs occurred over the first 12 h, where the two CAEs with random ICPs spun up precipitation much faster than the other CAEs but grossly overshot observed domain-total precipitation (Fig. 9a). While the CAEs with EnKF and SREF ICPs had broadly similar mean spinups, distinct trifurcation of GFS_{SREF} members occurred based on dynamic core (Fig. 9b), consistent with Johnson et al. (2011) and indicating how ICPs

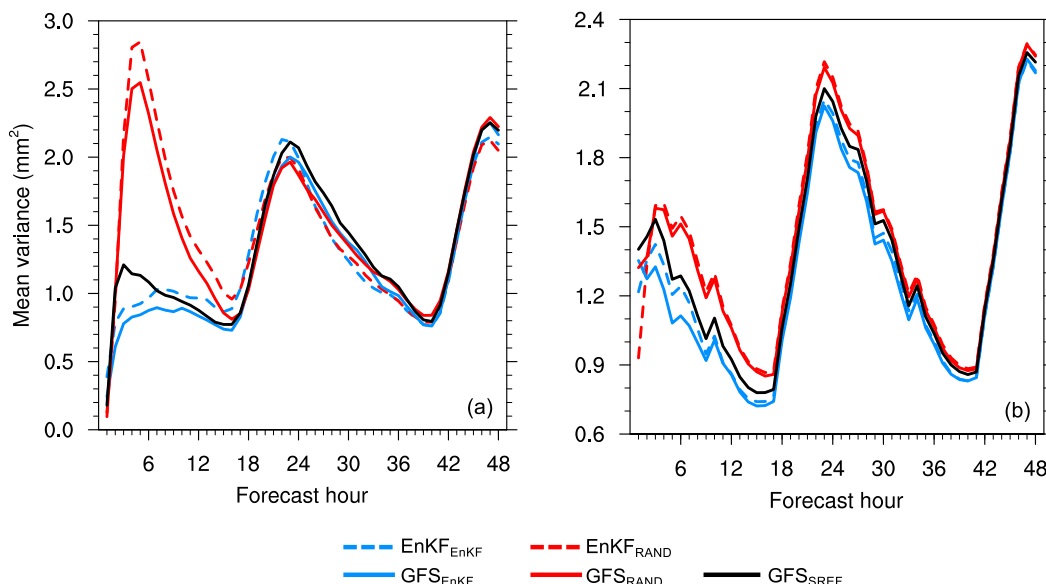


FIG. 6. Average ensemble variance (mm^2) over the verification region (CONUS east of 105°W) and all 31 3-km forecasts of 1-h accumulated precipitation as a function of forecast hour computed from (a) raw, native grid data and (b) bias-corrected precipitation interpolated onto the ST4 grid (see section 4b). 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).

reflecting multiple models can lead to clustering, which is undesirable (e.g., Gowan et al. 2018; Schwartz et al. 2019). In general, spinup appeared more sensitive to ICPs than IC center, even though initial center determined whether the CAE had initial hydrometeors (Table 3). This finding suggests ICP characteristics influence spinup more than initial hydrometeor state for 0000 UTC–initialized forecasts over the central-eastern CONUS.

Despite varied spinups, all CAEs generally well-represented diurnal cycle timing after 18 h, where GFS_{EnKF} and GFS_{SREF} members with NMM dynamic core ICPs had domain-total precipitation typically best matching observations, including the observed peak around 24 h (Fig. 9). Conversely, the two ensembles with random ICPs had less mean precipitation than observations between ~ 24 – 42 h, while at the maximum around 24 h $\text{EnKF}_{\text{EnKF}}$ produced too much precipitation (Fig. 9a). Interestingly, despite overpredicting at 24 h, $\text{EnKF}_{\text{EnKF}}$ precipitation dramatically decreased and underpredicted between ~ 26 – 42 h, perhaps due to insufficient upscale convection growth (e.g., Schwartz et al. 2015b).

GFS_{SREF} clearly had the widest range throughout the forecast, reflecting its ICPs with multimodel diversity (Fig. 9a). Additionally, the two CAEs with random ICPs typically had wider ranges than those with EnKF ICPs, particularly over the first 12 h, essentially a

manifestation of randomness. However, after 12 h, except for GFS_{SREF} , the four other CAEs had fairly similar ranges.

2) PRECIPITATION DISTRIBUTIONS

Average areal coverages of 1-h accumulated precipitation meeting or exceeding selected accumulation thresholds (e.g., 10.0 mm h^{-1}) were calculated over the verification region on native grids to assess precipitation distributions (Fig. 10). The CAEs generally well-represented diurnal cycle timing after the spinup, although there were sometimes biases, particularly for thresholds $\leq 1.0 \text{ mm h}^{-1}$, where all CAEs usually had mean coverages lower than those observed (Figs. 10a,b).

Areal coverage characteristics for thresholds $\geq 2.5 \text{ mm h}^{-1}$ (Figs. 10c–f) were broadly consistent with domain-total precipitation statistics, and GFS_{SREF} members again clustered based on dynamic core represented in the ICPs (not shown). Specifically, the two ensembles with random ICPs had lower mean coverages than observations between ~ 24 – 42 h for thresholds $\geq 2.5 \text{ mm h}^{-1}$ but clearly overpredicted during the spinup, which contributed to their excessive total precipitation during this period (e.g., Fig. 9a). Before and during the first observed precipitation peak (18–24 h), GFS_{EnKF} and GFS_{SREF} typically had ensemble mean coverages closest to observations, while $\text{EnKF}_{\text{EnKF}}$ overpredicted for thresholds between 2.5 and 10.0 mm h^{-1} (Figs. 10c–e).

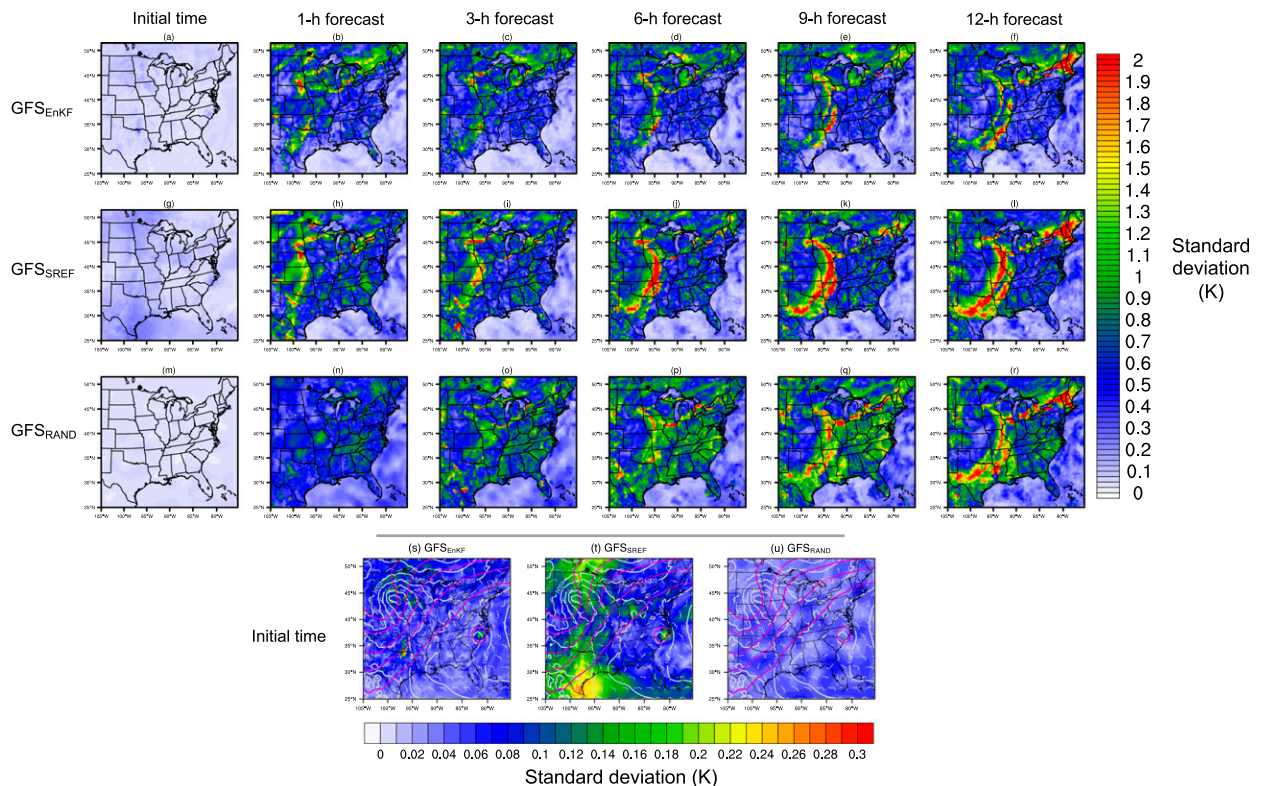


FIG. 7. Standard deviation of 2-m temperature (K) at the (a),(g),(m) initial time and for (b),(h),(n) 1-, (c),(i),(o) 3-, (d),(j),(p) 6-, (e),(k),(q) 9-, and (f),(l),(r) 12-h 3-km forecasts initialized at 0000 UTC 11 May 2015 for (a)–(f) GFS_{EnKF} , (g)–(l) GFS_{SREF} , and (m)–(r) GFS_{RAND} . (s)–(u) As in (a),(g), and (m), respectively, but with a different color scale to more easily see structural features, and 10-m winds (barbs; kts), sea level pressures (hPa; gray lines), and 500-hPa heights (m; magenta lines) from the 0000 UTC 11 May 2015 GFS analysis are overlaid.

Ensemble ranges of areal coverages (lines with circles in Fig. 10) also resembled those for domain-total precipitation, with GFS_{SREF} having the widest ranges and the CAEs with random ICPs possessing relatively large ranges for the first ~ 12 h.

Probability density functions (PDFs) further revealed the different spinups engendered by EnKF and random ICPs (Fig. 11). At 1 h, while finescale structures were still developing, CAEs with EnKF ICPs had more heavy precipitation than those with random ICPs, although none of the forecasts could yet reproduce the observed heavy rainfall frequency (Fig. 11a). However, between 1 and 3 h, heavy precipitation rapidly developed in the CAEs with random ICPs (Fig. 11b), with slower development in the CAEs with EnKF ICPs, and by 5–7 h, the CAEs with random ICPs produced too much rainfall $>40.0 \text{ mm h}^{-1}$ while PDFs of the CAEs with EnKF ICPs gradually aligned with those observed (Figs. 11c,d).

Collective findings clearly suggested issues when initializing CAEs with random noise for short-term precipitation forecasts (Figs. 8–11), possibly due to

gross imbalances in random initial states. Moreover, our results are consistent with Johnson et al. (2014), who found initializing CAEs with correlated random noise led to “spurious precipitation that formed over large areas on many cases” at short forecast ranges, and as Johnson et al. (2014) constructed random ICPs with smaller length scales than those used here, it appears using random noise to initialize CAEs may be challenging regardless of its correlation scale.

b. Ensemble precipitation verification

Areal coverages sometimes indicated biases (Fig. 10), which can hamper interpretation of verification metrics designed to quantify spatial errors (e.g., Baldwin and Kain 2006; Roberts and Lean 2008). Thus, forecasts were bias corrected before assessing measures of probabilistic forecast quality with a “probability-matching” approach that forced each ensemble member’s distribution to the ST4 distribution by replacing the model grid point containing the most precipitation within the verification region with the highest ST4 amount within the verification region, and so on, thus eliminating bias

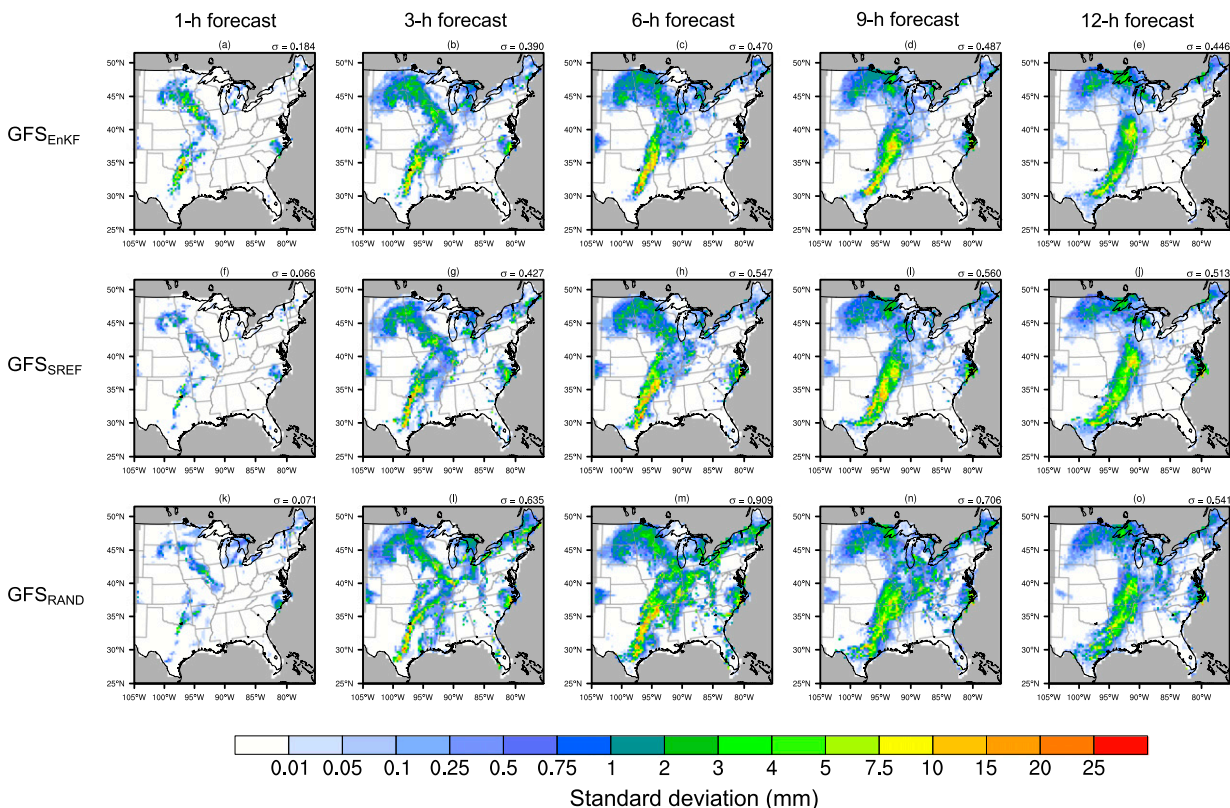


FIG. 8. Standard deviation of 1-h accumulated precipitation (mm) for (a),(f),(k) 1-, (b),(g),(l) 3-, (c),(h),(m) 6-, (d),(i),(n) 9-, and (e),(j),(o) 12-h 3-km forecasts initialized at 0000 UTC 11 May 2015 for (a)–(e) GFS_{ENKF}, (f)–(j) GFS_{SREF}, and (k)–(o) GFS_{RAND} over the verification region (CONUS east of 105°W). The mean standard deviation (σ) over the verification region is annotated above each panel.

(e.g., Ebert 2001; Clark et al. 2009, 2010a, b; Schwartz et al. 2015a; Loken et al. 2019; Pyle and Brill 2019). Despite replacing model values with observations, this method preserves forecast spatial patterns.

After interpolating precipitation forecasts to the ST4 grid and bias correcting, a “neighborhood” approach (e.g., Theis et al. 2005; Ebert 2008, 2009) was employed to derive probabilistic fields suitable for verification following Schwartz and Sobash (2017). First, ensemble probabilities (EPs) at a particular grid point were determined as the fraction of ensemble members predicting an event at that point, where an event was defined as precipitation meeting or exceeding an accumulation threshold (e.g., 5.0 mm h⁻¹). Then, “neighborhood ensemble probabilities” (NEPs; Schwartz et al. 2010; Schwartz and Sobash 2017) were computed by choosing a neighborhood length scale (r) to define a spatial neighborhood and averaging EPs over all grid points in the neighborhood. NEPs are probabilities of event occurrence at a point given a neighborhood length scale (e.g., Schwartz and Sobash 2017) and are more appropriate for verifying CAEs than point-based probabilities (i.e., EPs) because they incorporate

spatial uncertainty and acknowledge that CAEs are inherently inaccurate at the grid scale.

NEPs were produced from all CAEs with r between 5 and 150 km, which represented radii of circular neighborhoods. Following Schwartz and Sobash (2017), NEPs at the i th point were verified against corresponding observations (i.e., ST4) at the i th point, where the i th observed value could either be binary (i.e., 0 or 1) or fractional depending on what the metric required; fractional observations (e.g., Roberts and Lean 2008) at the i th point were obtained by determining the fraction of observed events within its neighborhood, analogously to NEPs.

For brevity, results are shown solely for $r = 100$ km, but overall findings were unchanged using different r . Additionally, a maximum event threshold of 10.0 mm h⁻¹ was used, as metrics computed at higher thresholds were noisy due to small sample sizes (e.g., Fig. 10).

Statistical significance testing followed Schwartz (2019), who examined performance of several ensembles, and the following text parallels from there. Specifically, statistical significance was determined with a bootstrap technique by randomly drawing paired samples (10000

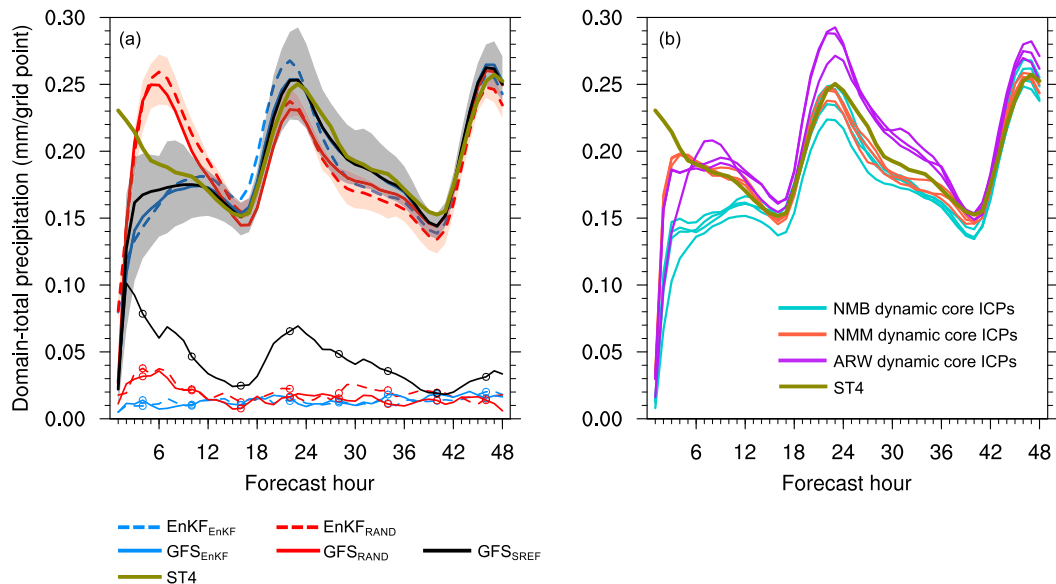


FIG. 9. (a) Average 1-h accumulated precipitation (mm) per grid point over all 31 3-km forecasts and the verification region (CONUS east of 105°W), computed on native grids, as a function of forecast hour. These statistics were computed for all 10 ensemble members, but for most ensembles, only the ensemble mean values (lines) and ranges (maximum minus minimum; lines with circle markers near the x axis) are shown. The red and gray shadings represent envelopes of the 10 members comprising $\text{EnKF}_{\text{RAND}}$ and GFS_{SREF} , respectively, and darker shadings indicate intersections of their two envelopes. 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). (b) As in (a), but the curves are for individual members from the GFS_{SREF} ensemble, with colors corresponding to the dynamic core each SREF member possessed. These different dynamic cores were reflected solely in the ICPs; for the WRF Model forecasts, all members had common configurations.

times) of daily statistics from two ensembles over all forecast cases to calculate resampled distributions of aggregate differences between two ensembles (e.g., Hamill 1999; Wolff et al. 2014). This procedure assumed individual forecasts, initialized 24 h apart, were independent (e.g., Hamill 1999). Bounds of 90% bootstrap confidence intervals (CIs) were obtained from the distribution of resampled aggregate differences using the bias corrected and accelerated method (e.g., Gilleland 2010). If bounds of a 90% bootstrap CI did not encompass zero, using a one-tailed interpretation, differences between two ensembles were statistically significant at the 95% level or higher.

1) FRACTIONS SKILL SCORES

The fractions skill score [FSS; Roberts and Lean (2008)] was used to evaluate spatial placement, where $\text{FSS} = 1$ means a perfect forecast and $\text{FSS} = 0$ indicates no skill. For fixed initial centers, CAEs with flow-dependent ICPs usually had higher FSSs than those with random ICPs, while differences between GFS_{SREF} and GFS_{EnKF} FSSs were usually small and not statistically significant (Fig. 12). These results indicated value of flow-dependent ICPs compared to

random ICPs and minimal benefits of multimodel ICPs. However, regardless of ICPs, the three CAEs initially centered on GFS analyses typically had higher FSSs than the two CAEs initially centered about EnKF mean analyses, demonstrating GFS analyses were generally better than EnKF mean analyses and suggesting initial center is more important than ICPs for achieving high FSSs.

2) RANK HISTOGRAMS

Rank histograms (e.g., Hamill 2001) based on domain-total precipitation were constructed as in Schwartz et al. (2014). Although rank histograms are sensitive to observation errors (e.g., Hacker et al. 2011), ST4 observation errors are not well-known and were not included. The reliability index (RI; Delle Monache et al. 2006) was used to summarize rank histogram flatness; lower values are preferable.

Observations fell within the ensemble more regularly and more optimal values were achieved in most bins when CAEs had random or SREF ICPs rather than EnKF ICPs (Fig. 13). GFS_{SREF} and GFS_{RAND} RIs were fairly similar and much smaller than GFS_{EnKF} RIs (Figs. 13a,c), and differences between the CAEs were

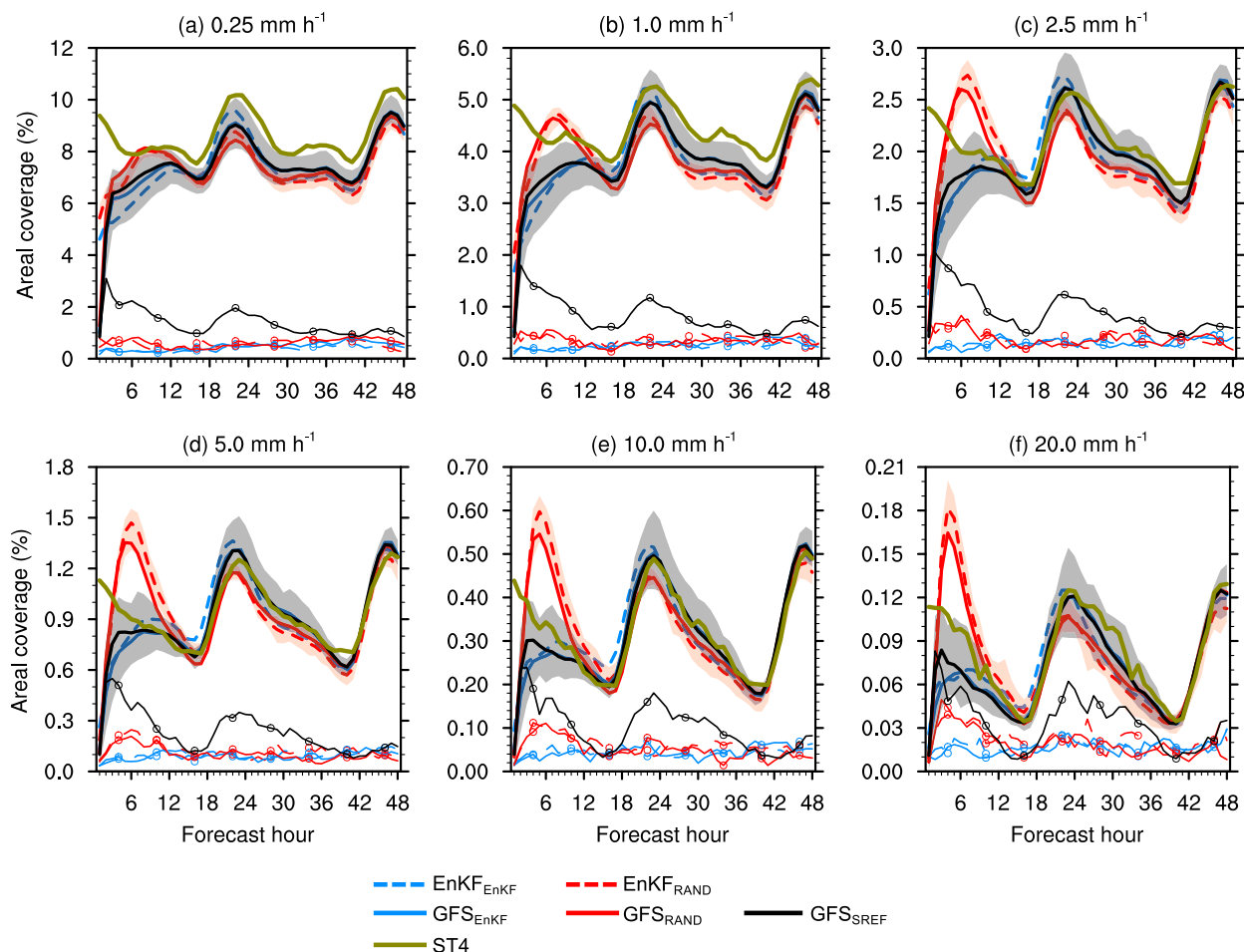


FIG. 10. Fractional areal coverage (%) of 1-h accumulated precipitation meeting or exceeding (a) 0.25, (b) 1.0, (c) 2.5, (d) 5.0, (e) 10.0, and (f) 20.0 mm h⁻¹ over the verification region (CONUS east of 105°W), computed on native grids and aggregated over all 31 3-km forecasts as a function of forecast hour. These statistics were computed for all 10 ensemble members, but for most ensembles, only the ensemble mean values (lines) and ranges (maximum minus minimum; lines with circle markers near the x axis) are shown. The red and gray shadings represent envelopes of the 10 members comprising EnKF_{RAND} and GFS_{SREF}, respectively, and darker shadings indicate intersections of their two envelopes. 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).

comparatively small after 18 h (cf. Fig. 13a and Fig. 13c and Fig. 13b and Fig. 13d), reflecting generally converging spread with time (e.g., Figs. 4, 6). Nonetheless, results at all forecast ranges suggest that enhanced precipitation spread engendered by random ICPs (e.g., Fig. 6) led to better dispersion characteristics than flow-dependent EnKF ICPs, even though this improved spread was also a manifestation of spinup issues (e.g., Figs. 7–11) and using random ICPs degraded forecast skill as measured by FSSs (e.g., Fig. 12).

3) ROC AREAS

Ability to discriminate events from climatology was quantified by area under the relative operating characteristic (ROC) curve (Mason 1982; Mason and

Graham 2002), which was computed using decision thresholds of 1%, 2%, 3%, 4%, 5%, 10%, 15%, . . . , 95%, and 100% and a trapezoidal approximation. ROC area > 0.5 indicates better discriminating ability than random forecasts.

As with FSSs, all three CAEs initially centered on GFS analyses usually had higher ROC areas than the two CAEs initially centered about EnKF mean analyses (Figs. 14a–d), again suggesting GFS analysis superiority to EnKF mean analyses and greater importance of initial center than ICPs. Between ~6–18 h, for fixed initial centers, the CAEs with random ICPs had statistically significantly higher ROC areas than those with EnKF ICPs, while before 6 h and after 18 h, EnKF and random ICPs yielded similar ROC

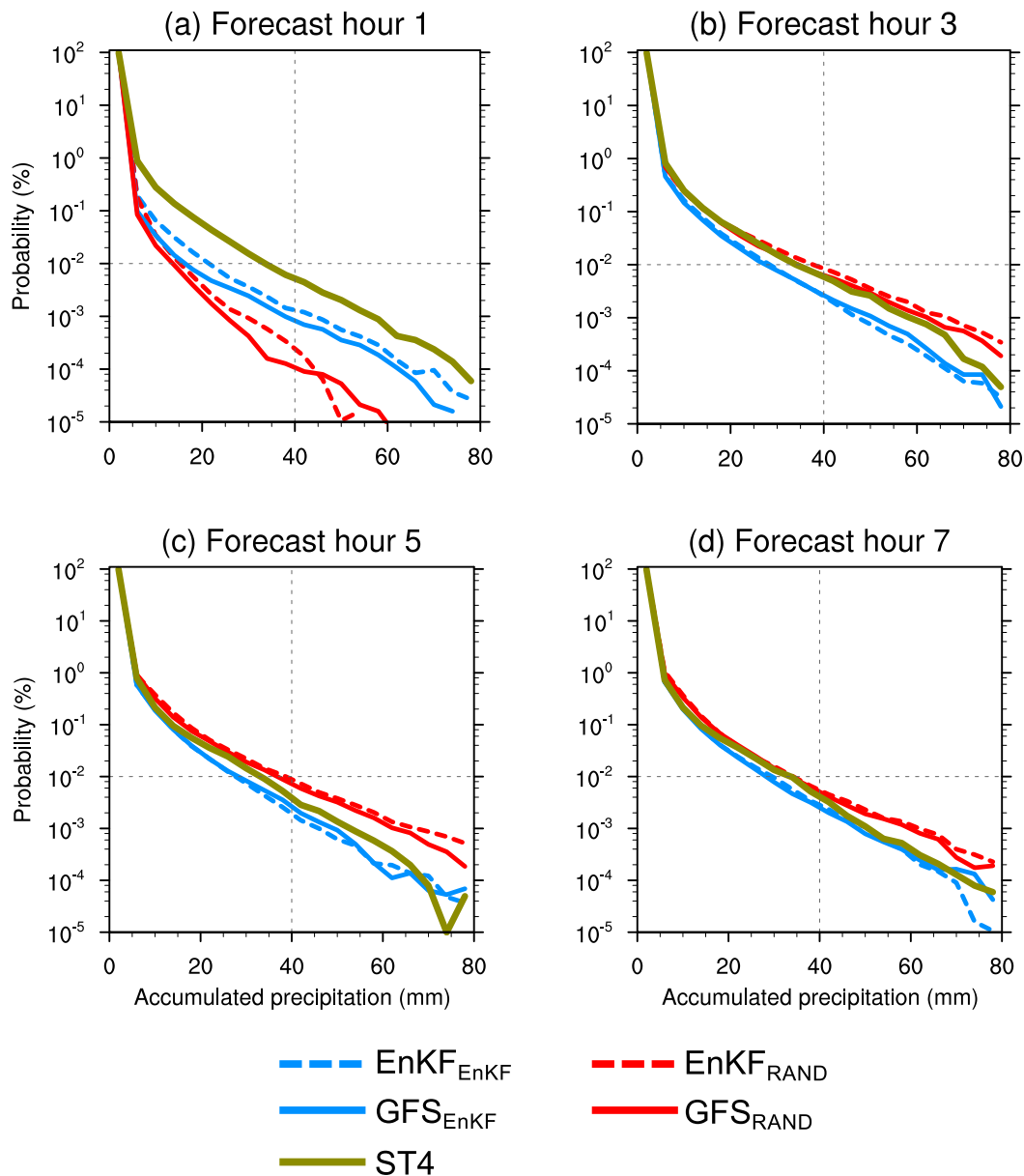


FIG. 11. Probability density functions (PDFs; %) constructed from all points within the verification region (CONUS east of 105°W) over all 31 (a) 1-, (b) 3-, (c) 5-, and (d) 7-h 3-km forecasts of 1-h accumulated precipitation (mm) for member 1 from various ensembles, computed on native grids. The corresponding observed (ST4) PDFs are also shown. Dashed lines are for reference to help visualize changes across forecast hours.

areas (Figs. 14a–d). These results differed from FSSs (Fig. 12) that clearly indicated EnKF ICPs were preferable to random ICPs. Outside the 6–18-h period, GFS_{SREF} had the highest ROC areas among the three CAEs with GFS initial centers that were often statistically significantly higher than GFS_{EnKF} ROC areas, suggesting benefits of incorporating multimodel diversity within CAE ICPs and contrasting the similar GFS_{SREF} and GFS_{EnKF} FSSs.

Further investigation revealed the relatively poor 6–18-h GFS_{EnKF} ROC areas compared to GFS_{RAND} and GFS_{SREF} were primarily due to insufficient contributions from NEPs < 25%. Specifically, GFS_{RAND} and GFS_{SREF} were less sharp than GFS_{EnKF}, with higher coverages of NEPs for $r = 100$ km between 5% and 25% and lower coverages of NEPs $\geq 45\%$ at most thresholds (Figs. 15a–d). In general, the GFS_{RAND} distribution differed more from the GFS_{EnKF} distribution than

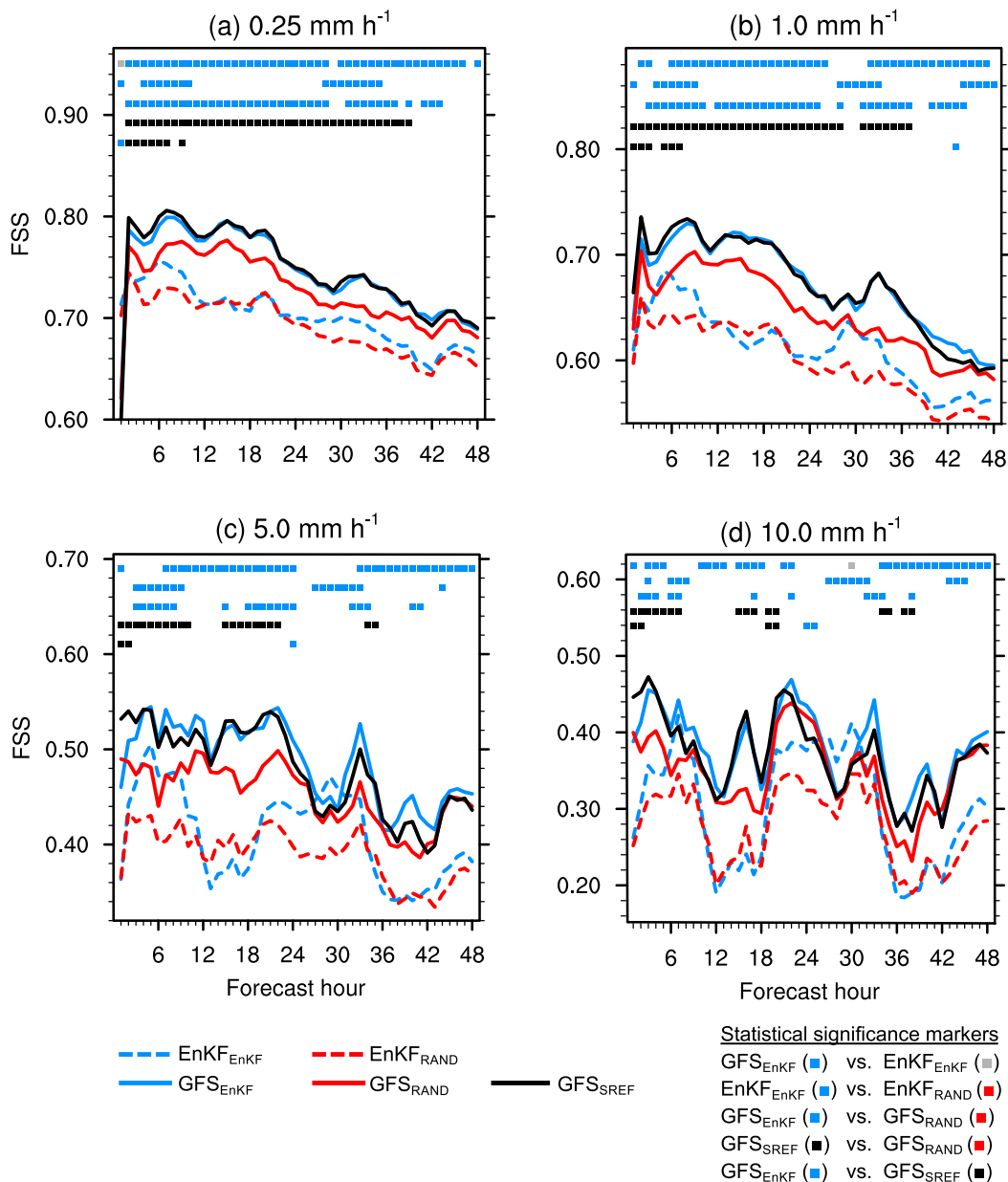


FIG. 12. Fractions skill score (FSS) over the verification region (CONUS east of 105°W) with a 100-km neighborhood length scale for the (a) 0.25, (b) 1.0, (c) 5.0, and (d) 10.0 mm h⁻¹ thresholds aggregated over all 31 3-km forecasts of 1-h accumulated precipitation as a function of forecast hour. Values 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). Note that the y-axis scales are different in each panel. Symbols along the top axis denote forecast hours when differences between two ensembles were statistically significant at the 95% level. As indicated in the legend, in order from top to bottom, the rows indicate differences between GFS_{EnKF} and EnKF_{EnKF}, EnKF_{EnKF} and EnKF_{RAND}, GFS_{EnKF} and GFS_{RAND}, GFS_{SREF} and GFS_{RAND}, and GFS_{EnKF} and GFS_{SREF}, with symbols corresponding to those in the legend that denote which ensemble had statistically significantly higher FSSs. For example, in the top row, blue symbols indicate GFS_{EnKF} had statistically significantly higher FSSs than EnKF_{EnKF}, while gray symbols indicate EnKF_{EnKF} had statistically significantly higher FSSs than GFS_{EnKF}. Similarly, in the bottom row, blue symbols indicate GFS_{EnKF} had statistically significantly higher FSSs than GFS_{SREF}, while black symbols indicate GFS_{SREF} had statistically significantly higher FSSs than GFS_{EnKF}. Absence of a symbol means the difference was not statistically significant at the 95% level.

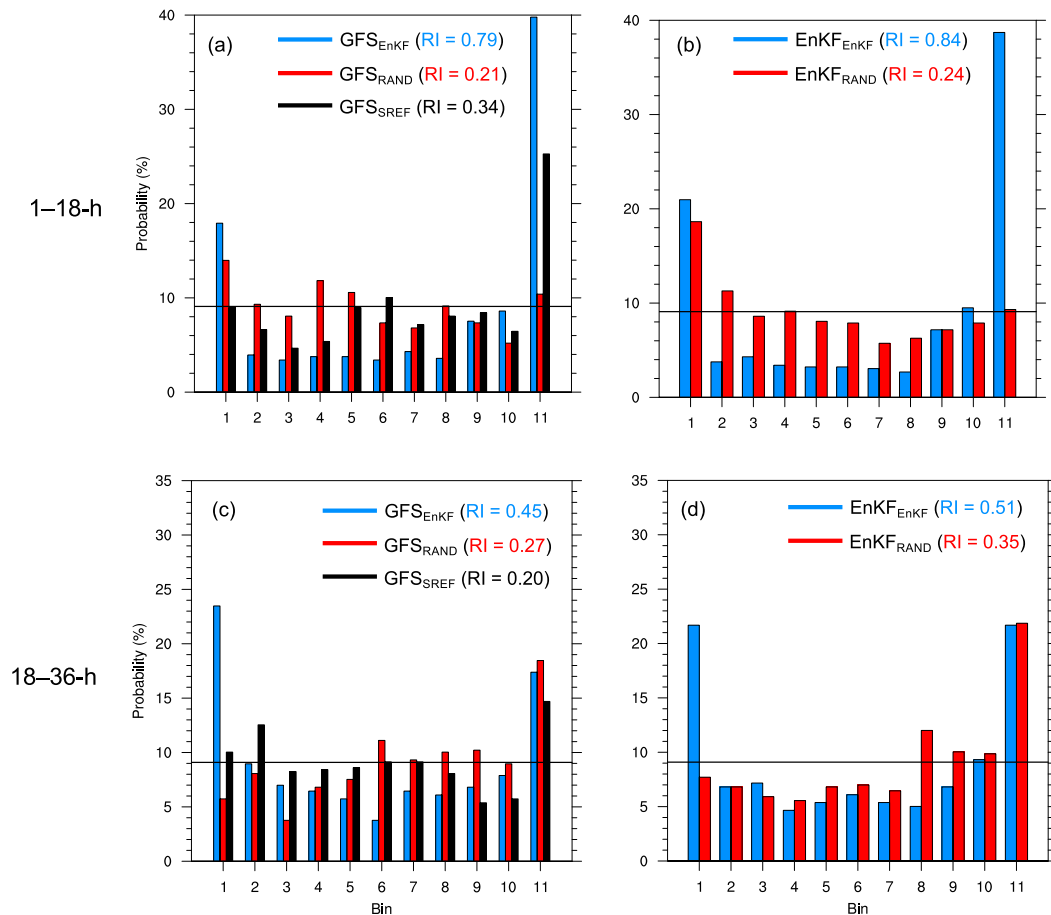


FIG. 13. Rank histograms containing all 31 3-km (a),(b) 1-18- and (c),(d) 18-36-h forecasts of non-bias-corrected domain-total 1-h accumulated precipitation on the ST4 grid over the verification region (CONUS east of 105°W) for ensembles initially centered about (a),(c) GFS and (b),(d) EnKF mean analyses. Horizontal lines are optimal values, and the reliability index (RI; Delle Monache et al. 2006) is annotated for each ensemble in the legend; lower values are better and indicate flatter rank histograms.

GFS_{SREF}; for example, within the 5%–25% bin for the 0.25 and 1.0 mm h⁻¹ thresholds, GFS_{RAND} had ~50% more NEPs than GFS_{EnKF} while the difference between GFS_{SREF} and GFS_{EnKF} was smaller (~15%; Figs. 15a,b). These enhanced low-probability coverages in GFS_{SREF} and GFS_{RAND} reflected their greater spreads relative to GFS_{EnKF} between 6 and 18 h (Fig. 6) that were beneficial (e.g., Fig. 13a) and enabled better detection of low-probability events while not appreciably increasing false alarm rates, boosting ROC areas.

Between 18 and 36 h, even though GFS_{RAND} again had more low probabilities than GFS_{EnKF} (Figs. 15e-h), differences between GFS_{RAND} and GFS_{EnKF} coverages of 5%–25% NEPs were smaller than between 6 and 18 h. Although greater GFS_{RAND} spread was beneficial from a dispersion perspective (e.g., Fig. 13c), GFS_{RAND} spatial placement was significantly poorer than GFS_{EnKF} (Fig. 12), counteracting benefits from

enhanced spread and likely leading to comparable GFS_{RAND} and GFS_{EnKF} ROC areas outside of 6–18 h (Figs. 14a-d). Conversely, differences between GFS_{EnKF} and GFS_{SREF} NEP distributions were similar across both forecast intervals (Fig. 15), and as GFS_{SREF} and GFS_{EnKF} FSSs were similar, the combination of good GFS_{SREF} placement and more GFS_{SREF} spread translated into higher GFS_{SREF} ROC areas for most of the forecast relative to GFS_{EnKF}.

Overall, ROC areas indicated more benefits of using both random and multimodel ICPs than FSSs. However, higher ROC areas from these techniques appear related solely to enhanced spread and greater low probability coverages. In fact, when ROC areas were computed with decision thresholds of 0%, 25%, 30%, 35%, ..., 95%, and 100% to explicitly exclude contributions from NEPs < 25%, although ROC areas plummeted, CAEs with EnKF ICPs had higher ROC areas than

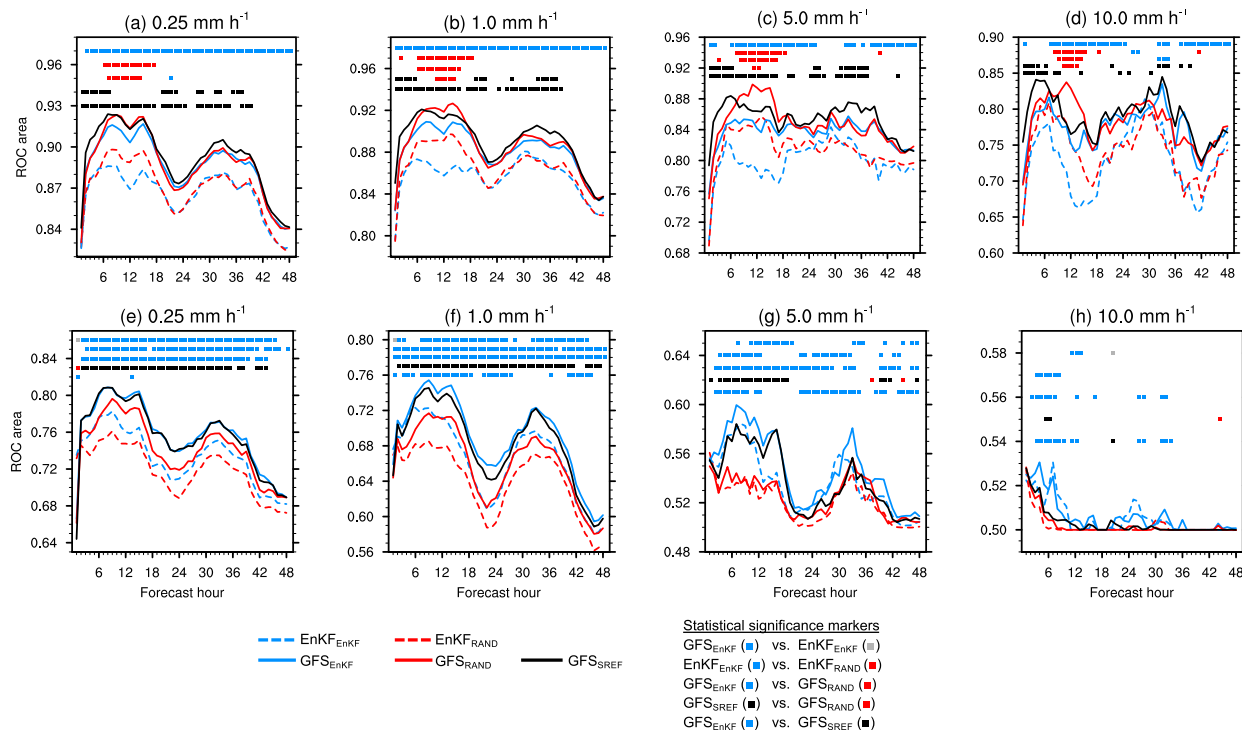


FIG. 14. Area under the ROC curve over the verification region (CONUS east of 105°W) with a 100-km neighborhood length scale for the (a),(e) 0.25, (b),(f) 1.0, (c),(g) 5.0, and (d),(h) 10.0 mm h⁻¹ thresholds aggregated over all 31 3-km forecasts of 1-h accumulated precipitation as a function of forecast hour computed with (a)–(d) the full range of decision thresholds and (e)–(h) a truncated set of decision thresholds with a lowest nonzero threshold of 25% (see text). 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). Note that the y-axis scales are different in each panel. Statistical significance is denoted along the top axis as in Fig. 12.

CAEs with the same initial center but random ICPs, and GFS_{EnKF} had comparable or higher ROC areas than GFS_{SREF} (Figs. 14e–h). These truncated ROC areas provided similar conclusions as FSSs regarding benefits of flow-dependent EnKF ICPs, and it appears that employing multimodel ICPs may be unnecessary for users unconcerned with low-probability decision thresholds.

4) ATTRIBUTES STATISTICS

Attributes diagrams (Wilks 2011) were constructed with forecast probability bins of 0%–5%, 5%–15%, 15%–25%, . . . , 85%–95%, and 95%–100% (Fig. 16) to assess calibration, with curves falling on the diagonal indicating perfect reliability. Over the first 18 h, for fixed IC centers, the CAEs with random and SREF ICPs were more reliable than those with EnKF ICPs for most thresholds and probability bins, and GFS_{RAND} was sometimes more reliable than GFS_{SREF} (Figs. 16a–d). The better GFS_{RAND} and GFS_{SREF} reliabilities compared to GFS_{EnKF} were aided by less sharp distributions with fewer high-probability forecasts (e.g., Figs. 15a–d) that diminished overconfidence, again reflecting their

greater spreads. Nonetheless, relatively poor GFS_{RAND} FSSs suggest many low probabilities did not correspond well with observations. Initial center again mattered, as for fixed ICPs, the CAEs with GFS initial centers typically had better reliabilities than those with EnKF mean initial centers.

Similar conclusions generally held at later times (18–36 h; Figs. 16e–h), although GFS_{SREF} and GFS_{RAND} had closer reliabilities than at earlier times. Over both periods, most ensembles were overconfident and all CAEs had little or no skill with respect to forecasts of climatology at the 10.0 mm h⁻¹ threshold, indicating challenges persist for making reliable predictions of highly localized events like heavy rainfall.

5. Summary and conclusions

Five sets of 48-h, 10-member, 3-km CAE forecasts were initialized at 0000 UTC each day in May 2015 over the CONUS with various configurations designed to isolate forecast sensitivity to ICPs and central initial state. Sensitivity to ICs extended throughout the 48-h forecasts, contrasting many European studies showing

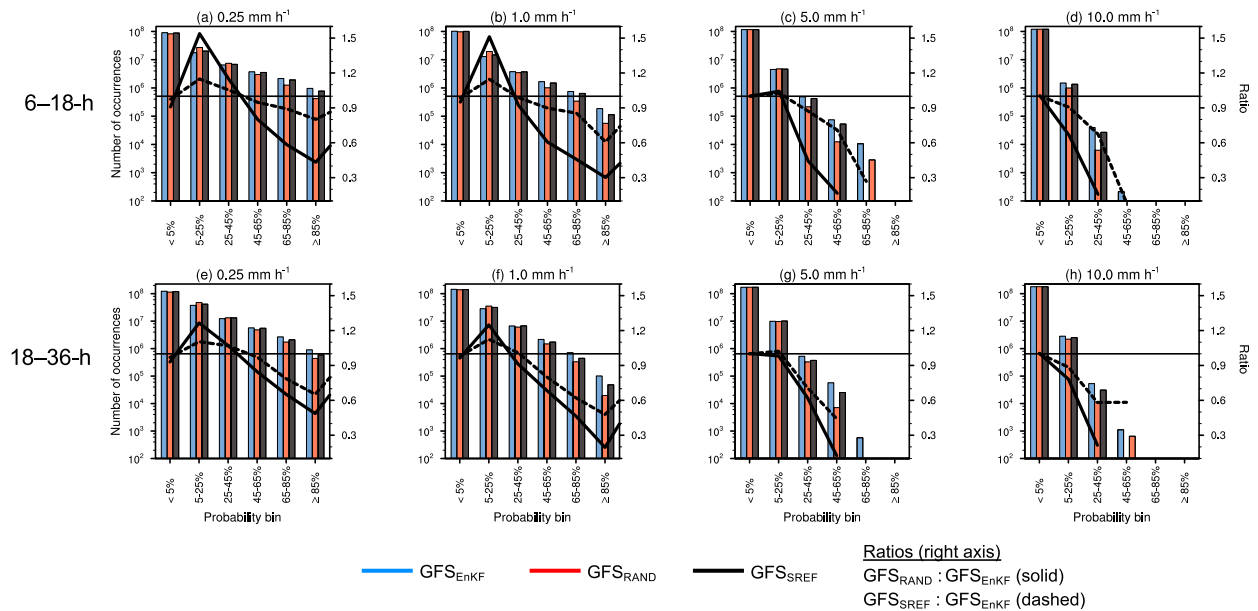


FIG. 15. Sharpness diagrams depicting how often bias-corrected NEPs of 1-h accumulated precipitation computed with $r = 100$ km fell into various probabilistic bins over the verification region (CONUS east of 105°W) and all 31 (a)–(d) 6–18- and (e)–(h) 18–36-h forecasts for event thresholds of (a),(e) 0.25, (b),(f) 1.0, (c),(g) 5.0, and (d),(h) 10.0 mm h^{-1} . Scale along the right axis is for ratios of number of occurrences between GFS_{RAND} to GFS_{ENKF} (solid) and GFS_{SREF} to GFS_{ENKF} (dashed), with a horizontal line at 1.

IC impacts through only 6–12 h (e.g., Hohengger et al. 2008; Vié et al. 2011; Kühnlein et al. 2014; RB16); this disparity is probably due to the much bigger computational domain used here, and our findings suggest enhanced importance of ICs for large domains.

Specifically, using random mesoscale ICPs yielded undesirable spinup characteristics and relatively poor FSSs compared to employing flow-dependent ICPs provided by both single-physics, single-dynamics 15-km limited-area continuously cycling EnKF analyses and 3-h multimodel SREF forecasts. However, these deleterious characteristics from random ICPs increased spread, leading to less overconfidence and broader low-probability coverages that improved ROC areas, rank histogram flatness, and attributes statistics compared to EnKF—and sometimes SREF—ICPs. Therefore, it appears random ICPs engendered some beneficial properties despite lack of flow dependence, but substantial work is needed to further understand and remedy detrimental impacts of random noise on model spinup.

Compared to EnKF ICPs, SREF ICPs yielded comparable FSSs but improved performance for spread-sensitive metrics. Yet, individual members of the SREF-initialized CAE had different climatologies that undesirably clustered by dynamic core reflected in its ICPs. Thus, although SREF-based and random ICPs often provided improvements over EnKF ICPs, given the challenges associated with multimodel and

random ICPs, collective results suggest obtaining “good spread” in CAEs remains elusive, and within future operational CAEs like those being developed under NOAA’s UFS, it may be more fruitful to attempt to recover the helpful, spread-inducing aspects from random and multimodel ICPs by instead using stochastic physics schemes in association with single-physics, single-dynamics, flow-dependent ICPs (e.g., Bouttier et al. 2012; Romine et al. 2014; Jankov et al. 2019).

Additionally, our findings stress the importance of CAE initial center, which was more important than ICPs for achieving high ROC areas and FSSs. Moreover, CAEs initially centered about operational GFS analyses were unequivocally superior to those initially centered on our experimental EnKF mean analyses. These results strongly suggest relative superiority of GFS analyses and lend credence to the “partial cycling” strategy currently employed by NOAA’s limited-area DA systems over the CONUS that periodically discards cycled states and replaces them with fields from a global model (e.g., Benjamin et al. 2016; Wu et al. 2017).

Despite our seemingly discouraging EnKF-based results, continuously cycling EnKFs over large regional domains can potentially be enhanced by decreasing the cycling period (e.g., using 1-h cycles), assimilating more observations, and likely most importantly, improving the limited-area NWP model (e.g., Romine et al. 2013). In addition, our results documenting very slow perturbation

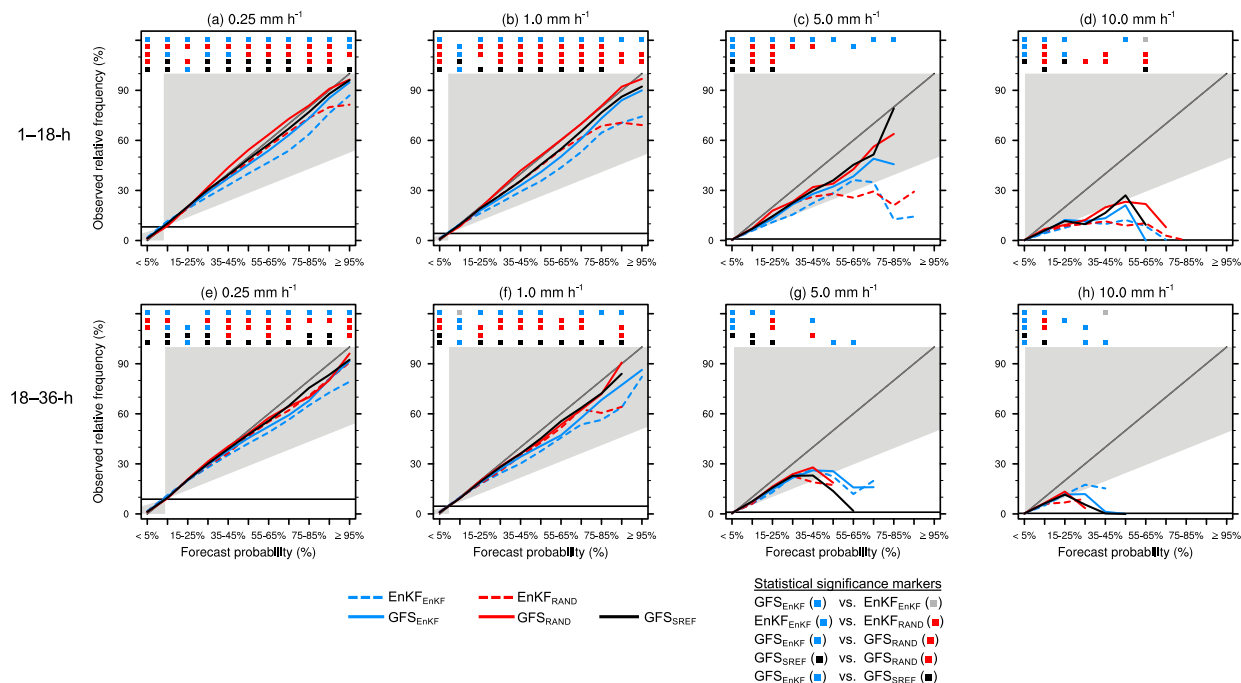


FIG. 16. Attributes statistics computed over the verification region (CONUS east of 105°W) with a 100-km neighborhood length scale aggregated over all 31 (a)–(d) 1–18- and (e)–(h) 18–36-h 3-km forecasts of 1-h accumulated precipitation for the (a),(e) 0.25, (b),(f) 1.0, (c),(g) 5.0, and (d),(h) 10.0 mm h⁻¹ thresholds. Horizontal lines near the x axis represent observed frequencies of the event and diagonal lines are lines of perfect reliability. Points lying in gray-shaded regions had skill compared to climatological forecasts as measured by the Brier skill score (Brier 1950; Wilks 2011). Statistical significance is denoted along the top axis as in Fig. 12, where the symbols denote the ensemble statistically significantly closest to perfect reliability. Values were not plotted for a particular bin if fewer than 500 grid points had forecast probabilities in that bin over the verification region and all 31 forecasts. Note that the attributes diagrams themselves stop at 100%; area above 100% was added to make room for statistical significance markers.

growth over the first 12 h from EnKF ICPs compared to random ICPs suggest efforts toward understanding and accelerating these slow growths should be undertaken to improve short-term forecast spread from EnKF ICPs. While increasing EnKF resolution may also help, especially for nowcasting purposes, finer EnKF resolution is likely not a panacea, and it is entirely possible that continuously cycling limited-area EnKFs, despite their many attractive properties, may not currently be optimal for initializing large-domain regional CAEs, particularly for next-day forecasts that are less impacted by spinup. Nonetheless, ongoing research at NCAR is attempting to improve limited-area NWP models (e.g., Wong et al. 2020), with hopes that these efforts will translate into better continuously cycling DA systems over large regional domains.

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 and NA17OAR4590122. Thanks to Adam Clark and two anonymous reviewers

for their constructive comments. All forecasts were produced on NCAR’s Yellowstone supercomputer (Computational and Information Systems Laboratory 2016). The National Center for Atmospheric Research is sponsored by the National Science Foundation.

REFERENCES

Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2003: Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Wea. Forecasting*, **18**, 918–932, [https://doi.org/10.1175/1520-0434\(2003\)018<0918:SOPFSS>2.0.CO;2](https://doi.org/10.1175/1520-0434(2003)018<0918:SOPFSS>2.0.CO;2).

Anderson, J. L., 2001: An ensemble adjustment Kalman filter for data assimilation. *Mon. Wea. Rev.*, **129**, 2884–2903, [https://doi.org/10.1175/1520-0493\(2001\)129<2884:AEAKFF>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<2884:AEAKFF>2.0.CO;2).

—, 2003: A local least squares framework for ensemble filtering. *Mon. Wea. Rev.*, **131**, 634–642, [https://doi.org/10.1175/1520-0493\(2003\)131<0634:ALLSFF>2.0.CO;2](https://doi.org/10.1175/1520-0493(2003)131<0634:ALLSFF>2.0.CO;2).

—, 2009: Spatially and temporally varying adaptive covariance inflation for ensemble filters. *Tellus*, **61A**, 72–83, <https://doi.org/10.1111/j.1600-0870.2008.00361.x>.

—, 2012: Localization and sampling error correction in ensemble Kalman filter data assimilation. *Mon. Wea. Rev.*, **140**, 2359–2371, <https://doi.org/10.1175/MWR-D-11-00013.1>.

- , and N. Collins, 2007: Scalable implementations of ensemble filter algorithms for data assimilation. *J. Atmos. Oceanic Technol.*, **24**, 1452–1463, <https://doi.org/10.1175/JTECH2049.1>.
- , T. Hoar, K. Raeder, H. Liu, N. Collins, R. Torn, and A. Arellano, 2009: The Data Assimilation Research Testbed: A community facility. *Bull. Amer. Meteor. Soc.*, **90**, 1283–1296, <https://doi.org/10.1175/2009BAMS2618.1>.
- Baldwin, M. E., and J. S. Kain, 2006: Sensitivity of several performance measures to displacement error, bias, and event frequency. *Wea. Forecasting*, **21**, 636–648, <https://doi.org/10.1175/WAF933.1>.
- Barker, D. M., 2005: Southern high-latitude ensemble data assimilation in the Antarctic Mesoscale Prediction System. *Mon. Wea. Rev.*, **133**, 3431–3449, <https://doi.org/10.1175/MWR3042.1>.
- , and Coauthors, 2012: The Weather Research and Forecasting Model's Community Variational/Ensemble Data Assimilation System: WRFDA. *Bull. Amer. Meteor. Soc.*, **93**, 831–843, <https://doi.org/10.1175/BAMS-D-11-00167.1>.
- Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, <https://doi.org/10.1175/MWR-D-15-0242.1>.
- Bishop, C. H., and D. Hodyss, 2009a: Ensemble covariances adaptively localized with ECO-RAP. Part I: Tests on simple error models. *Tellus*, **61A**, 84–96, <https://doi.org/10.1111/j.1600-0870.2008.00371.x>.
- , and —, 2009b: Ensemble covariances adaptively localized with ECO-RAP. Part II: A strategy for the atmosphere. *Tellus*, **61A**, 97–111, <https://doi.org/10.1111/j.1600-0870.2008.00372.x>.
- Blunden, J., and D. S. Arndt, 2016: State of the Climate in 2015. *Bull. Amer. Meteor. Soc.*, **97**, SI–S275, <https://doi.org/10.1175/2016BAMSSTATEOFTHECLIMATE.1>.
- Bouttier, F., B. Vieu, O. Nuissier, and L. Raynaud, 2012: Impact of stochastic physics in a convection-permitting ensemble. *Mon. Wea. Rev.*, **140**, 3706–3721, <https://doi.org/10.1175/MWR-D-12-00031.1>.
- Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2).
- Cafaro, C., T. H. A. Frame, J. Methven, N. Roberts, and J. Bröcker, 2019: The added value of convection-permitting ensemble forecasts of sea breeze compared to a Bayesian forecast driven by the global ensemble. *Quart. J. Roy. Meteor. Soc.*, **145**, 1780–1798, <https://doi.org/10.1002/qj.3531>.
- Chen, F., and J. Dudhia, 2001: Coupling an advanced land surface–hydrology model with the Penn State–NCAR MM5 modeling system. Part I: Model implementation and sensitivity. *Mon. Wea. Rev.*, **129**, 569–585, [https://doi.org/10.1175/1520-0493\(2001\)129<0569:CAALSH>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<0569:CAALSH>2.0.CO;2).
- Clark, A. J., 2017: Generation of ensemble mean precipitation forecasts from convection-allowing ensembles. *Wea. Forecasting*, **32**, 1569–1583, <https://doi.org/10.1175/WAF-D-16-0199.1>.
- , W. A. Gallus Jr., M. Xue, and F. Kong, 2009: A comparison of precipitation forecast skill between small convection-allowing and large convection-parameterizing ensembles. *Wea. Forecasting*, **24**, 1121–1140, <https://doi.org/10.1175/2009WAF222222.1>.
- , —, and M. L. Weisman, 2010a: Neighborhood-based verification of precipitation forecasts from convection-allowing NCAR WRF Model simulations and the operational NAM. *Wea. Forecasting*, **25**, 1495–1509, <https://doi.org/10.1175/2010WAF2222404.1>.
- , —, M. Xue, and F. Kong, 2010b: Growth of spread in convection-allowing and convection-parameterizing ensembles. *Wea. Forecasting*, **25**, 594–612, <https://doi.org/10.1175/2009WAF2222318.1>.
- , and Coauthors, 2011: Probabilistic precipitation forecast skill as a function of ensemble size and spatial scale in a convection-allowing ensemble. *Mon. Wea. Rev.*, **139**, 1410–1418, <https://doi.org/10.1175/2010MWR3624.1>.
- , and Coauthors, 2012: An overview of the 2010 Hazardous Weather Testbed Experimental Forecast Program Spring Experiment. *Bull. Amer. Meteor. Soc.*, **93**, 55–74, <https://doi.org/10.1175/BAMS-D-11-00040.1>.
- , and Coauthors, 2018: The Community Leveraged Unified Ensemble (CLUE) in the 2016 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Bull. Amer. Meteor. Soc.*, **99**, 1433–1448, <https://doi.org/10.1175/BAMS-D-16-0309.1>.
- , and Coauthors, 2020: A real-time, simulated forecasting experiment for advancing the prediction of hazardous convective weather. *Bull. Amer. Meteor. Soc.*, <https://doi.org/10.1175/BAMS-D-19-0298.1>, in press.
- Clayton, A. M., A. C. Lorenc, and D. M. Barker, 2013: Operational implementation of a hybrid ensemble/4D-Var global data assimilation system at the Met Office. *Quart. J. Roy. Meteor. Soc.*, **139**, 1445–1461, <https://doi.org/10.1002/qj.2054>.
- Computational and Information Systems Laboratory, 2016: Yellowstone: IBM iDataPlex System (NCAR Community Computing). National Center for Atmospheric Research, Boulder, CO, accessed 19 May 2020, <http://n2t.net/ark:/85065/d7wd3xhc>.
- Delle Monache, L., J. P. Hacker, Y. Zhou, X. Deng, and R. B. Stull, 2006: Probabilistic aspects of meteorological and ozone regional ensemble forecasts. *J. Geophys. Res.*, **111**, D24307, <https://doi.org/10.1029/2005JD006917>.
- Du, J., and Coauthors, 2014: NCEP regional ensemble update: Current systems and planned storm-scale ensembles. *26th Conf. on Weather Analysis and Forecasting/22nd Conf. on Numerical Weather Prediction*, Atlanta, GA, Amer. Meteor. Soc., J1.4, <https://ams.confex.com/ams/94Annual/webprogram/Paper239030.html>.
- , G. DiMego, D. Jovic, B. Ferrier, B. Yang, and B. Zhou, 2015: Short Range Ensemble Forecast (SREF) system at NCEP: Recent development and future transition. *27th Conf. on Weather Analysis and Forecasting/23rd Conf. on Numerical Weather Prediction*, Chicago, IL, Amer. Meteor. Soc., 2A.5, <https://ams.confex.com/ams/27WAF23NWP/webprogram/Paper273421.html>.
- Duc, L., K. Saito, and H. Seko, 2013: Spatial–temporal fractions verification for high-resolution ensemble forecasts. *Tellus*, **65A**, 18171, <https://doi.org/10.3402/tellusa.v65i0.18171>.
- Durran, D. R., and M. Gingrich, 2014: Atmospheric predictability: Why butterflies are not important. *J. Atmos. Sci.*, **71**, 2476–2488, <https://doi.org/10.1175/JAS-D-14-0007.1>.
- Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480, [https://doi.org/10.1175/1520-0493\(2001\)129<2461:AOAPMS>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2).
- , 2008: Fuzzy verification of high resolution gridded forecasts: A review and proposed framework. *Meteor. Appl.*, **15**, 51–64, <https://doi.org/10.1002/met.25>.
- , 2009: Neighborhood verification: A strategy for rewarding close forecasts. *Wea. Forecasting*, **24**, 1498–1510, <https://doi.org/10.1175/2009WAF2222251.1>.

- Evans, C., D. F. Van Dyke, and T. Lericos, 2014: How do forecasters utilize output from a convection-permitting ensemble forecast system? Case study of a high-impact precipitation event. *Wea. Forecasting*, **29**, 466–486, <https://doi.org/10.1175/WAF-D-13-00064.1>.
- Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, **99**, 10 143–10 162, <https://doi.org/10.1029/94JC00572>.
- Gallo, B. T., and Coauthors, 2017: Breaking new ground in severe weather prediction: The 2015 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Wea. Forecasting*, **32**, 1541–1568, <https://doi.org/10.1175/WAF-D-16-0178.1>.
- Gaspari, G., and S. E. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.*, **125**, 723–757, <https://doi.org/10.1002/qj.49712555417>.
- Gebhardt, C., S. E. Theis, M. Paulat, and Z. Ben Bouallègue, 2011: Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries. *Atmos. Res.*, **100**, 168–177, <https://doi.org/10.1016/j.atmosres.2010.12.008>.
- Gilleland, E., 2010: Confidence intervals for forecast verification. NCAR Tech. Note NCAR/TN-479+STR, 71 pp., <https://doi.org/10.5065/D6WD3XJM>.
- Gowan, T. M., W. J. Steenburgh, and C. S. Schwartz, 2018: Validation of mountain precipitation forecasts from the convection-permitting NCAR ensemble and operational forecast systems over the western United States. *Wea. Forecasting*, **33**, 739–765, <https://doi.org/10.1175/WAF-D-17-0144.1>.
- Hacker, J., and Coauthors, 2011: The U.S. Air Force Weather Agency's mesoscale ensemble: Scientific description and performance results. *Tellus*, **63A**, 625–641, <https://doi.org/10.1111/j.1600-0870.2010.00497.x>.
- Hagelin, S., J. Son, R. Swinbank, A. McCabe, N. Roberts, and W. Tennant, 2017: The Met Office convective-scale ensemble, MOGREPS-UK. *Quart. J. Roy. Meteor. Soc.*, **143**, 2846–2861, <https://doi.org/10.1002/qj.3135>.
- Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, [https://doi.org/10.1175/1520-0434\(1999\)014<0155:HTFENP>2.0.CO;2](https://doi.org/10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2).
- , 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560, [https://doi.org/10.1175/1520-0493\(2001\)129<0550:IORHFV>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2).
- Harnisch, F., and C. Keil, 2015: Initial conditions for convective-scale ensemble forecasting provided by ensemble data assimilation. *Mon. Wea. Rev.*, **143**, 1583–1600, <https://doi.org/10.1175/MWR-D-14-00209.1>.
- Harris, F. J., 1978: On the use of windows for harmonic analysis with the discrete Fourier transform. *Proc. IEEE*, **66**, 51–83, <https://doi.org/10.1109/PROC.1978.10837>.
- Hohenegger, C., and C. Schär, 2007: Predictability and error growth dynamics in cloud-resolving models. *J. Atmos. Sci.*, **64**, 4467–4478, <https://doi.org/10.1175/2007JAS2143.1>.
- , A. Walser, W. Langhans, and C. Schär, 2008: Cloud-resolving ensemble simulations of the August 2005 Alpine flood. *Quart. J. Roy. Meteor. Soc.*, **134**, 889–904, <https://doi.org/10.1002/qj.252>.
- Houtekamer, P. L., and F. Zhang, 2016: Review of the ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, **144**, 4489–4532, <https://doi.org/10.1175/MWR-D-15-0440.1>.
- , H. L. Mitchell, G. Pellerin, M. Buehner, M. Charron, L. Spacek, and B. Hansen, 2005: Atmospheric data assimilation with an ensemble Kalman filter: Results with real observations. *Mon. Wea. Rev.*, **133**, 604–620, <https://doi.org/10.1175/MWR-2864.1>.
- , X. Deng, H. L. Mitchell, S.-J. Baek, and N. Gagnon, 2014: Higher resolution in an operational ensemble Kalman filter. *Mon. Wea. Rev.*, **142**, 1143–1162, <https://doi.org/10.1175/MWR-D-13-00138.1>.
- Hsiao, L.-F., D.-S. Chen, Y.-H. Kuo, Y.-R. Guo, T.-C. Yeh, J.-S. Hong, C.-T. Fong, and C.-S. Lee, 2012: Application of WRF 3DVAR to operational typhoon prediction in Taiwan: Impact of outer loop and partial cycling approaches. *Wea. Forecasting*, **27**, 1249–1263, <https://doi.org/10.1175/WAF-D-11-00131.1>.
- Iacono, M. J., J. S. Delamere, E. J. Mlawer, M. W. Shephard, S. A. Clough, and W. D. Collins, 2008: Radiative forcing by long-lived greenhouse gases: Calculations with the AER radiative transfer models. *J. Geophys. Res.*, **113**, D13103, <https://doi.org/10.1029/2008JD009944>.
- Janjić, Z. I., 1994: The step-mountain eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. *Mon. Wea. Rev.*, **122**, 927–945, [https://doi.org/10.1175/1520-0493\(1994\)122<0927:TSMECM>2.0.CO;2](https://doi.org/10.1175/1520-0493(1994)122<0927:TSMECM>2.0.CO;2).
- , 2002: Nonsingular implementation of the Mellor–Yamada level 2.5 scheme in the NCEP Meso model. NCEP Office Note 437, 61 pp., <http://www.emc.ncep.noaa.gov/officenotes/newernotes/on437.pdf>.
- Jankov, I., J. Beck, J. Wolff, M. Harrold, J. B. Olson, T. Smirnova, C. Alexander, and J. Berner, 2019: Stochastically perturbed parameterizations in an HRRR-based ensemble. *Mon. Wea. Rev.*, **147**, 153–173, <https://doi.org/10.1175/MWR-D-18-0092.1>.
- Jirak, I. L., A. J. Clark, B. Roberts, B. T. Gallo, and S. J. Weiss, 2018: Exploring the optimal configuration of the High Resolution Ensemble Forecast System. *25th Conf. on Numerical Weather Prediction*, Denver, CO, Amer. Meteor. Soc., 14B.6, <https://ams.confex.com/ams/29WAF25NWP/webprogram/Paper345640.html>.
- Johnson, A., and X. Wang, 2016: A study of multiscale initial condition perturbation methods for convection-permitting ensemble forecasts. *Mon. Wea. Rev.*, **144**, 2579–2604, <https://doi.org/10.1175/MWR-D-16-0056.1>.
- , M. Xue, and F. Kong, 2011: Hierarchical cluster analysis of a convection-allowing ensemble during the Hazardous Weather Testbed 2009 Spring Experiment. Part II: Ensemble clustering over the whole experiment period. *Mon. Wea. Rev.*, **139**, 3694–3710, <https://doi.org/10.1175/MWR-D-11-00016.1>.
- , and Coauthors, 2014: Multiscale characteristics and evolution of perturbations for warm season convection-allowing precipitation forecasts: Dependence on background flow and method of perturbation. *Mon. Wea. Rev.*, **142**, 1053–1073, <https://doi.org/10.1175/MWR-D-13-00204.1>.
- Jones, T. A., and D. J. Stensrud, 2012: Assimilating AIRS temperature and mixing ratio profiles using an ensemble Kalman filter approach for convective-scale forecasts. *Wea. Forecasting*, **27**, 541–564, <https://doi.org/10.1175/WAF-D-11-00090.1>.
- Keresturi, E., Y. Wang, F. Meier, F. Weidle, C. Wittmann, and A. Atencia, 2019: Improving initial condition perturbations in a convection-permitting ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **145**, 993–1012, <https://doi.org/10.1002/qj.3473>.
- Klasa, C., M. Arpagaus, A. Walser, and H. Wernli, 2018: An evaluation of the convection-permitting ensemble COSMO-E

- for three contrasting precipitation events in Switzerland. *Quart. J. Roy. Meteor. Soc.*, **144**, 744–764, <https://doi.org/10.1002/qj.3245>.
- Kong, F., and Coauthors, 2008: Real-time storm-scale ensemble forecast experiment—Analysis of 2008 spring experiment data. *24th Conf. on Severe Local Storms*, Savannah, GA, Amer. Meteor. Soc., 12.3, <https://ams.confex.com/ams/pdfpapers/141827.pdf>.
- , and Coauthors, 2009: A real-time storm-scale ensemble forecast system: 2009 spring experiment. *23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction*, Omaha, NE, Amer. Meteor. Soc., 16A.3, <https://ams.confex.com/ams/pdfpapers/154118.pdf>.
- Kühnlein, C., C. Keil, G. C. Craig, and C. Gebhardt, 2014: The impact of downscaled initial condition perturbations on convective-scale ensemble forecasts of precipitation. *Quart. J. Roy. Meteor. Soc.*, **140**, 1552–1562, <https://doi.org/10.1002/qj.2238>.
- Lang, S. T. K., M. Bonavita, and M. Leutbecher, 2015: On the impact of re-centring initial conditions for ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **141**, 2571–2581, <https://doi.org/10.1002/qj.2543>.
- Lin, Y., and K. E. Mitchell, 2005: The NCEP stage II/IV hourly precipitation analyses: Development and applications. *19th Conf. on Hydrology*, San Diego, CA, Amer. Meteor. Soc., 1.2, <http://ams.confex.com/ams/pdfpapers/83847.pdf>.
- Loken, E. D., A. J. Clark, M. Xue, and F. Kong, 2019: Spread and skill in mixed- and single-physics convection-allowing ensembles. *Wea. Forecasting*, **34**, 305–330, <https://doi.org/10.1175/WAF-D-18-0078.1>.
- Mason, I. B., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291–303.
- Mason, S. J., and N. E. Graham, 2002: Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quart. J. Roy. Meteor. Soc.*, **128**, 2145–2166, <https://doi.org/10.1256/003590002320603584>.
- Mellor, G. L., and T. Yamada, 1982: Development of a turbulence closure model for geophysical fluid problems. *Rev. Geophys. Space Phys.*, **20**, 851–875, <https://doi.org/10.1029/RG020i004p00851>.
- Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *J. Geophys. Res.*, **102**, 16 663–16 682, <https://doi.org/10.1029/97JD00237>.
- Nelson, B. R., O. P. Prat, D.-J. Seo, and E. Habib, 2016: Assessment and implications of NCEP Stage IV quantitative precipitation estimates for product intercomparisons. *Wea. Forecasting*, **31**, 371–394, <https://doi.org/10.1175/WAF-D-14-00112.1>.
- Pan, Y., K. Zhu, M. Xue, X. Wang, M. Hu, S. G. Benjamin, S. S. Weygandt, and J. S. Whitaker, 2014: A GSI-based coupled EnSRF-En3DVar hybrid data assimilation system for the operational Rapid Refresh Model: Tests at a reduced resolution. *Mon. Wea. Rev.*, **142**, 3756–3780, <https://doi.org/10.1175/MWR-D-13-00242.1>.
- Pantillon, F., S. Lerch, P. Knippertz, and U. Corsmeier, 2018: Forecasting wind gusts in winter storms using a calibrated convection-permitting ensemble. *Quart. J. Roy. Meteor. Soc.*, **144**, 1864–1881, <https://doi.org/10.1002/qj.3380>.
- Parrish, D. F., and J. C. Derber, 1992: The National Meteorological Center's spectral statistical interpolation analysis system. *Mon. Wea. Rev.*, **120**, 1747–1763, [https://doi.org/10.1175/1520-0493\(1992\)120<1747:TNMCS>2.0.CO;2](https://doi.org/10.1175/1520-0493(1992)120<1747:TNMCS>2.0.CO;2).
- Peralta, C., Z. B. Bouallègue, S. E. Theis, C. Gebhardt, and M. Buchhold, 2012: Accounting for initial condition uncertainties in COSMO-DE-EPS. *J. Geophys. Res.*, **117**, D07108, <https://doi.org/10.1029/2011JD016581>.
- Porson, A. N., S. Hagelin, D. F. A. Boyd, N. M. Roberts, R. North, S. Webster, and J. C.-F. Lo, 2019: Extreme rainfall sensitivity in convective-scale ensemble modelling over Singapore. *Quart. J. Roy. Meteor. Soc.*, **145**, 3004–3022, <https://doi.org/10.1002/qj.3601>.
- Powers, J. G., and Coauthors, 2017: The Weather Research and Forecasting Model: Overview, system efforts, and future directions. *Bull. Amer. Meteor. Soc.*, **98**, 1717–1737, <https://doi.org/10.1175/BAMS-D-15-00308.1>.
- Pyle, M. E., and K. F. Brill, 2019: A comparison of two methods for bias correcting precipitation skill scores. *Wea. Forecasting*, **34**, 3–13, <https://doi.org/10.1175/WAF-D-18-0109.1>.
- Raynaud, L., and F. Bouttier, 2016: Comparison of initial perturbation methods for ensemble prediction at convective scale. *Quart. J. Roy. Meteor. Soc.*, **142**, 854–866, <https://doi.org/10.1002/qj.2686>.
- , and —, 2017: The impact of horizontal resolution and ensemble size for convective-scale probabilistic forecasts. *Quart. J. Roy. Meteor. Soc.*, **143**, 3037–3047, <https://doi.org/10.1002/qj.3159>.
- Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, <https://doi.org/10.1175/2007MWR2123.1>.
- Romine, G. S., C. S. Schwartz, C. Snyder, J. Anderson, and M. Weisman, 2013: Model bias in a continuously cycled assimilation system and its influence on convection-permitting forecasts. *Mon. Wea. Rev.*, **141**, 1263–1284, <https://doi.org/10.1175/MWR-D-12-00112.1>.
- , —, J. Berner, K. R. Fossell, C. Snyder, J. L. Anderson, and M. L. Weisman, 2014: Representing forecast error in a convection-permitting ensemble system. *Mon. Wea. Rev.*, **142**, 4519–4541, <https://doi.org/10.1175/MWR-D-14-00100.1>.
- Schellander-Gorgas, T., Y. Wang, F. Meier, F. Weidle, C. Wittmann, and A. Kann, 2017: On the forecast skills of a convection-permitting ensemble. *Geosci. Model Dev.*, **10**, 35–56, <https://doi.org/10.5194/gmd-10-35-2017>.
- Schraff, C., H. Reich, A. Rhodin, A. Schomburg, K. Stephan, A. Perriñez, and R. Potthast, 2016: Kilometre-scale ensemble data assimilation for the COSMO model (KENDA). *Quart. J. Roy. Meteor. Soc.*, **142**, 1453–1472, <https://doi.org/10.1002/qj.2748>.
- Schumacher, R. S., and A. J. Clark, 2014: Evaluation of ensemble configurations for the analysis and prediction of heavy-rain-producing mesoscale convective systems. *Mon. Wea. Rev.*, **142**, 4108–4138, <https://doi.org/10.1175/MWR-D-13-00357.1>.
- Schwartz, C. S., 2016: Improving large-domain convection-allowing forecasts with high-resolution analyses and ensemble data assimilation. *Mon. Wea. Rev.*, **144**, 1777–1803, <https://doi.org/10.1175/MWR-D-15-0286.1>.
- , 2019: Medium-range convection-allowing ensemble forecasts with a variable-resolution global model. *Mon. Wea. Rev.*, **147**, 2997–3023, <https://doi.org/10.1175/MWR-D-18-0452.1>.
- , and Z. Liu, 2014: Convection-permitting forecasts initialized with continuously cycling limited-area 3DVAR, ensemble Kalman filter, and “hybrid” variational-ensemble data assimilation systems. *Mon. Wea. Rev.*, **142**, 716–738, <https://doi.org/10.1175/MWR-D-13-00100.1>.

- , and R. A. Sobash, 2017: Generating probabilistic forecasts from convection-allowing ensembles using neighborhood approaches: A review and recommendations. *Mon. Wea. Rev.*, **145**, 3397–3418, <https://doi.org/10.1175/MWR-D-16-0400.1>.
- , and Coauthors, 2010: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*, **25**, 263–280, <https://doi.org/10.1175/2009WAF2222267.1>.
- , G. S. Romine, K. R. Smith, and M. L. Weisman, 2014: Characterizing and optimizing precipitation forecasts from a convection-permitting ensemble initialized by a mesoscale ensemble Kalman filter. *Wea. Forecasting*, **29**, 1295–1318, <https://doi.org/10.1175/WAF-D-13-00145.1>.
- , —, M. L. Weisman, R. A. Sobash, K. R. Fossell, K. W. Manning, and S. B. Trier, 2015a: A real-time convection-allowing ensemble prediction system initialized by mesoscale ensemble Kalman filter analyses. *Wea. Forecasting*, **30**, 1158–1181, <https://doi.org/10.1175/WAF-D-15-0013.1>.
- , —, R. A. Sobash, K. R. Fossell, and M. L. Weisman, 2015b: NCAR's experimental real-time convection-allowing ensemble prediction system. *Wea. Forecasting*, **30**, 1645–1654, <https://doi.org/10.1175/WAF-D-15-0103.1>.
- , Z. Liu, and X.-Y. Huang, 2015c: Sensitivity of limited-area hybrid variational-ensemble analyses and forecasts to ensemble perturbation resolution. *Mon. Wea. Rev.*, **143**, 3454–3477, <https://doi.org/10.1175/MWR-D-14-00259.1>.
- , G. S. Romine, R. A. Sobash, K. R. Fossell, and M. L. Weisman, 2019: NCAR's real-time convection-allowing ensemble project. *Bull. Amer. Meteor. Soc.*, **100**, 321–343, <https://doi.org/10.1175/BAMS-D-17-0297.1>.
- Skamarock, W. C., 2004: Evaluating mesoscale NWP models using kinetic energy spectra. *Mon. Wea. Rev.*, **132**, 3019–3032, <https://doi.org/10.1175/MWR2830.1>.
- , and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp., <https://doi.org/10.5065/D68S4MVH>.
- Tegen, I., P. Hollrig, M. Chin, I. Fung, D. Jacob, and J. Penner, 1997: Contribution of different aerosol species to the global aerosol extinction optical thickness: Estimates from model results. *J. Geophys. Res.*, **102**, 23 895–23 915, <https://doi.org/10.1029/97JD01864>.
- Tennant, W., 2015: Improving initial condition perturbations for MOGREPS-UK. *Quart. J. Roy. Meteor. Soc.*, **141**, 2324–2336, <https://doi.org/10.1002/qj.2524>.
- Theis, S. E., A. Hense, and U. Damrath, 2005: Probabilistic precipitation forecasts from a deterministic model: A pragmatic approach. *Meteor. Appl.*, **12**, 257–268, <https://doi.org/10.1017/S1350482705001763>.
- Thompson, G., P. R. Field, R. M. Rasmussen, and W. D. Hall, 2008: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Mon. Wea. Rev.*, **136**, 5095–5115, <https://doi.org/10.1175/2008MWR2387.1>.
- Tiedtke, M., 1989: A comprehensive mass flux scheme for cumulus parameterization in large-scale models. *Mon. Wea. Rev.*, **117**, 1779–1800, [https://doi.org/10.1175/1520-0493\(1989\)117<1779:ACMFSF>2.0.CO;2](https://doi.org/10.1175/1520-0493(1989)117<1779:ACMFSF>2.0.CO;2).
- Torn, R. D., and C. A. Davis, 2012: The influence of shallow convection on tropical cyclone track forecasts. *Mon. Wea. Rev.*, **140**, 2188–2197, <https://doi.org/10.1175/MWR-D-11-00246.1>.
- , G. J. Hakim, and C. Snyder, 2006: Boundary conditions for limited-area ensemble Kalman filters. *Mon. Wea. Rev.*, **134**, 2490–2502, <https://doi.org/10.1175/MWR3187.1>.
- Vié, B., O. Nuissier, and V. Ducrocq, 2011: Cloud-resolving ensemble simulations of Mediterranean heavy precipitation events: Uncertainty on initial conditions and lateral boundary conditions. *Mon. Wea. Rev.*, **139**, 403–423, <https://doi.org/10.1175/2010MWR3487.1>.
- Wang, X., D. F. Parrish, D. T. Kleist, and J. S. Whitaker, 2013: GSI 3DVAR-based ensemble-variational hybrid data assimilation for NCEP Global Forecast System: Single-resolution experiments. *Mon. Wea. Rev.*, **141**, 4098–4117, <https://doi.org/10.1175/MWR-D-12-00141.1>.
- Warner, T. T., R. A. Peterson, and R. E. Treadon, 1997: A tutorial on lateral boundary conditions as a basic and potentially serious limitation to regional numerical weather prediction. *Bull. Amer. Meteor. Soc.*, **78**, 2599–2617, [https://doi.org/10.1175/1520-0477\(1997\)078<2599:ATOLBC>2.0.CO;2](https://doi.org/10.1175/1520-0477(1997)078<2599:ATOLBC>2.0.CO;2).
- Wheatley, D. M., K. H. Knopfmeier, T. A. Jones, and G. J. Creager, 2015: Storm-scale data assimilation and ensemble forecasting with the NSSL experimental Warn-on-Forecast system. Part I: Radar data experiments. *Wea. Forecasting*, **30**, 1795–1817, <https://doi.org/10.1175/WAF-D-15-0043.1>.
- Whitaker, J. S., and T. M. Hamill, 2012: Evaluating methods to account for system errors in ensemble data assimilation. *Mon. Wea. Rev.*, **140**, 3078–3089, <https://doi.org/10.1175/MWR-D-11-00276.1>.
- Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.
- Wolff, J. K., M. Harrold, T. Fowler, J. H. Gotway, L. Nance, and B. G. Brown, 2014: Beyond the basics: Evaluating model-based precipitation forecasts using traditional, spatial, and object-based methods. *Wea. Forecasting*, **29**, 1451–1472, <https://doi.org/10.1175/WAF-D-13-00135.1>.
- Wong, M., G. Romine, and C. Snyder, 2020: Model improvement via systematic investigation of physics tendencies. *Mon. Wea. Rev.*, **148**, 671–688, <https://doi.org/10.1175/MWR-D-19-0255.1>.
- Wu, W.-S., D. F. Parrish, E. Rogers, and Y. Lin, 2017: Regional ensemble-variational data assimilation using global ensemble forecasts. *Wea. Forecasting*, **32**, 83–96, <https://doi.org/10.1175/WAF-D-16-0045.1>.
- Xue, M., and Coauthors, 2007: CAPS real-time storm-scale ensemble and high-resolution forecasts as part of the NOAA Hazardous Weather Testbed 2007 Spring Experiment. *22nd Conf. on Weather Analysis and Forecasting/18th Conf. on Numerical Weather Prediction*, Park City, UT, Amer. Meteor. Soc., 3B.1, <http://ams.confex.com/ams/pdfpapers/124587.pdf>.
- Zhang, C., Y. Wang, and K. Hamilton, 2011: Improved representation of boundary layer clouds over the southeast Pacific in ARW-WRF using a modified Tiedtke cumulus parameterization scheme. *Mon. Wea. Rev.*, **139**, 3489–3513, <https://doi.org/10.1175/MWR-D-10-05091.1>.
- Zhang, F., M. Zhang, and J. Poterjoy, 2013: E3DVar: Coupling an ensemble Kalman filter with three-dimensional variational data assimilation in a limited-area weather prediction model and comparison to E4DVar. *Mon. Wea. Rev.*, **141**, 900–917, <https://doi.org/10.1175/MWR-D-12-00075.1>.
- Zhang, X., 2018: Application of a convection-permitting ensemble prediction system to quantitative precipitation forecasts over southern China: Preliminary results during SCMREX. *Quart. J. Roy. Meteor. Soc.*, **144**, 2842–2862, <https://doi.org/10.1002/qj.3411>.
- , 2019: Multiscale characteristics of different-source perturbations and their interactions for convection-permitting ensemble forecasting during SCMREX. *Mon. Wea. Rev.*, **147**, 291–310, <https://doi.org/10.1175/MWR-D-18-0218.1>.