A Framework for Convective-Scale Observing System Simulation Experiments Using Ensembles

JEREMY A. GIBBS,[a] JOSHUA G. GEBAUER,[a,b] LOUIS J. WICKER,[a] MATTHEW B. AMMON,[a,b,c]
AND DEREK R. STRATMAN[a,b]

[a] *NOAA/OAR National Severe Storms Laboratory, Norman, Oklahoma*
[b] *Cooperative Institute for Severe and High-Impact Weather Research and Operations, Norman, Oklahoma*
[c] *School of Meteorology, University of Oklahoma, Norman, Oklahoma*

(Manuscript received 9 August 2024, in final form 25 May 2025, accepted 29 July 2025)

ABSTRACT: Convection-allowing models are used to predict the evolution of severe weather phenomena in the atmosphere. These models are sensitive to errors in the numerical environment used to initialize their forecast integration. Data assimilation methods can help overcome these errors but are often limited by sparse observational coverage. Numerous novel observation platforms are currently available that promise to close these coverage gaps, but they have not yet been widely assimilated into weather models. Observing system simulation experiments are often used to determine how best to assimilate these novel observations in space and time by using synthetic observations from a high-resolution "nature run." However, a single deterministic nature run does not provide a measure of the flow's intrinsic predictability within the chosen modeling system. We present a framework to provide insight into this predictability by using an ensemble of nature runs. The ensemble provides a range of likely outcomes for storm evolution and an upper limit against which forecasts are verified. We applied this framework to two events from Oklahoma in 2023: a quasi-linear convective system in February and a supercell case in April. The intrinsic predictability of the nature-run ensemble was used to calibrate each case and to verify the forecast ensembles. Results showed that the intrinsic predictability suggested by the nature-run ensemble was both field and case dependent. This framework may help guide future studies by giving researchers a better understanding of what is possible for a chosen flow problem and modeling system and how best to include and arrange novel observations.

SIGNIFICANCE STATEMENT: We wanted to place the relative gains in severe weather forecast performance that arise from including observational data in context to what may be reasonably expected from our chosen modeling system. In other words, if our forecast moves an inch toward the "truth," it matters whether that truth is a foot or a mile away from our forecast estimate—and in turn, whether the effort to move that inch is worth the cost. Results demonstrate a proof-of-concept framework to achieve this goal and show that our ability to understand what is possible is both field and case dependent. Our framework may help improve efficiency by providing researchers with a better understanding of the trade-offs involved with their observational experiment design.

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

1. Introduction

Convection-allowing model (CAM) forecasts can skillfully predict the evolution of thunderstorms, but their skill is sensitive to errors in the initial environment. Atmospheric observations are frequently assimilated to improve the representation of the environment. Gaps in observational coverage can reduce the effectiveness of data assimilation (DA) systems and allow error to persist in the environment. Many novel observation platforms are now available in varying stages of development and deployment that promise to improve the observational coverage of the atmosphere but as yet have not been widely assimilated into numerical weather prediction models. Observing system simulation experiments (OSSEs) are frequently used to determine how to optimally assimilate these novel observations (e.g., spatial and temporal density) and understand their impact on forecast performance. Simulated observations extracted from a high-resolution simulation that closely resembles an observed weather phenomenon—a nature run—are assimilated in OSSEs. The impact from various platforms and networks with different spatial and temporal characteristics can be readily tested because simulated observations are created synthetically.

While OSSEs are an effective tool to test the impact of different assimilated observation types, extensive testing is required to optimally tune each DA parameter. To do this, many OSSEs and real-data experiments conduct DA tests to determine which configuration results in the most skilled forecasts. This approach helps to establish an adequate DA configuration but may not result in optimal system performance. Even if forecasts are initialized with accurate conditions, subtle errors in the environment and model physics often cause forecast skill to degrade with time. For instance, the evolution of convective storms is highly sensitive to modest changes in the environment. Many of these features occur at currently unobservable spatial and temporal

Corresponding author: Jeremy A. Gibbs, jeremy.gibbs@noaa.gov

DOI: 10.1175/WAF-D-24-0153.1

scales and thus remain a large source of forecast uncertainty. Importantly, the error growth is highly case and flow dependent (Melhauser and Zhang 2012).

Melhauser and Zhang (2012) outline two forms of predictability associated with forecast errors from a meso- and convective-scale OSSE. Practical predictability is present when a reduction of the initial-condition error via the inclusion of new observations leads to a reduction in forecast error. Practical predictability is missing when small differences in initial conditions result in multiple preferred solutions (Melhauser and Zhang 2012), the details of which are weather and flow dependent (Zhang et al. 2019). Intrinsic predictability is present when a reduction of the initial-condition error between two fields with small differences in their initial state does not reduce the error growth rate between the two forecasts (e.g., Lorenz 1963, 1969). In practice, OSSEs have a mixture of these two types of predictability.

Typical OSSEs create a nature run and assume it to be the "true" solution (i.e., the perfect model assumption). However, the use of a single deterministic nature run cannot adequately measure the predictability of the flow. We introduce the use of an ensemble of nature runs to provide insight into this predictability by adding subtle changes in the environment to see how small forecast errors grow in time. This framework as presented is meant to augment OSSE analyses for individual cases and not to draw conclusions about the efficacy of particular observational strategies across a broad range of events. Henceforth, our use of "intrinsic predictability" describes the growth of differences between the nature run control and ensemble—though we acknowledge that this error growth is some mixture of practical and intrinsic predictability. The nature-run ensemble then provides a range of likely outcomes for the evolution of individual convective cells and an envelope of likely outcomes against which to verify the OSSE forecasts. This insight serves to "calibrate" the OSSE results (i.e., to place forecast error changes that arise from the inclusion of new observations in the context of the intrinsic predictability limit of a given case). This knowledge can help guide the design of the OSSEs and provide realistic expectations for the extent of potential improvements gained from the inclusion of additional observations.

In this study, we used the initial and boundary conditions from the National Severe Storms Laboratory (NSSL) Warn-on-Forecast System (WoFS; Stensrud et al. 2009; Heinselman et al. 2024) for two real-data cases that occurred in central Oklahoma in 2023: an extensive quasi-linear convective system (QLCS) on 26 February and a spatially confined tornado outbreak on 19 April. We conducted forecast ensembles to test the impacts of assimilating simulated observations from radars, surface stations, and profilers. We then used the intrinsic predictability from an ensemble of nature runs to both calibrate each case and to verify the OSSE forecast ensembles. We describe each case in section 2, the experimental design in section 3, and verification methods and results in section 4. Finally, we offer conclusions and future direction in section 5.

2. Case descriptions

We describe the chosen cases here to motivate their use in this study. Our summaries are basic because it is beyond the scope of the paper to provide an in-depth case review. For instance, our actual experimental setup is not completely representative of the corresponding nature run setup because we explore OSSEs. We encourage readers to explore the detailed summaries provided by the National Weather Service (NWS) for more information about our chosen cases (NWS 2023a,b).

a. 26 February 2023 QLCS case

The 26 February 2023 case (C1; see Fig. 1) was an out-of-season humdinger of a severe weather event that occurred in an extreme shear parameter space that was atypical of winter. An intense and compact closed upper-level low moved eastward across the southwestern contiguous United States (CONUS), which was centered over northern New Mexico by 0000 UTC 27 February. This upper-level low caused a surface low with rapid deepening to develop near the western edge of the Oklahoma Panhandle throughout the day, which resulted in the transport of an unseasonably warm and moist, conditionally unstable air mass ahead of a dryline that was located in the western Texas Panhandle. Soundings in central Oklahoma showed that the environment was strongly capped, but forcing from the upper-level low and approaching cold front caused supercells to develop along the dryline, which quickly grew upscale into a QLCS. As the system approached the Texas/Oklahoma border, embedded mesocyclones in the QLCS began to produce tornadoes. The QLCS intermittently produced these tornadoes from 0100 to 0400 UTC as it propagated from the western Oklahoma border into the Oklahoma City metropolitan area. The system produced 12 tornadoes, including three that were rated EF2 by the NWS. Here, "EF" refers to the enhanced Fujita scale, which rates the intensity of tornadoes on a numerical scale between 1 and 5 (WSEC 2006; McDonald et al. 2009). A special sounding launched by the NWS Norman Forecast Office at 0300 UTC 27 February ahead of the QLCS showed that the prestorm environment was still capped but had a 0–1-km storm relative helicity value of 1017 $m^2 s^{-2}$. An EF2 tornado passed just south of the forecast office only 30 min after that sounding was released. The strongly forced nature of the convection and the rapidly changing environment in which it occurred make this event a good case for OSSEs to compare to the more subtly forced supercell case on 19 April 2023.

b. 19 April 2023 supercell case

The 19 April 2023 severe weather event (C2; see Fig. 2) in central Oklahoma was a typical central Plains severe weather setup. That is, there was a broad upper-level trough over the western CONUS and a slow, deepening surface low over central Kansas. A dryline extended south from this surface low through west-central Oklahoma. The upper-level forcing for ascent was located north of Oklahoma, so convection initiation was considered highly conditional on daytime heating and the strength and depth of the dryline circulation. Dewpoint temperatures east of the dryline were in the low to mid-60s (°F). A sounding in Norman, Oklahoma, at 1900 UTC 19 April showed that the capping inversion was weak. However, dry air was present above the boundary layer, which created uncertainty for

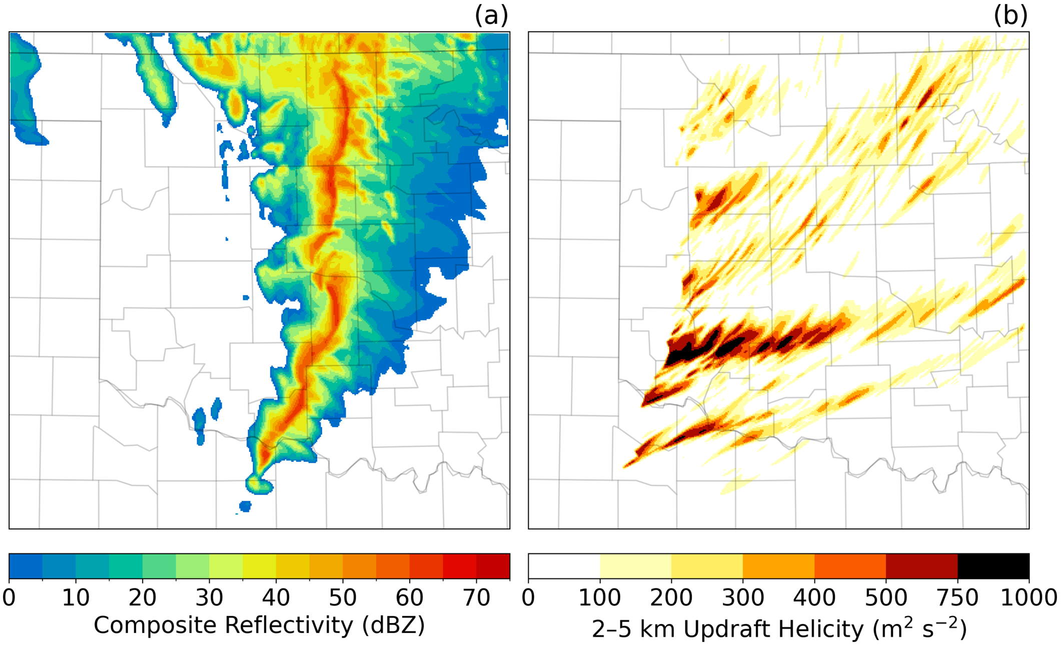
(a)
(b)
0
10
20
30
40
50
60
70
Composite Reflectivity (dBZ)
0
100
200
300
400
500
750
1000
2–5 km Updraft Helicity (m^2 s^{-2})

FIG. 1. (a) CR from the NRC valid at 0300 UTC 27 Feb 2023 and (b) storm-total UH25 swaths for the 26 Feb 2023 case.

how long storms could maintain themselves as they propagated away from the dryline. Convection initiation commenced in west-central Oklahoma at 2100 UTC. These storms produced severe hail but struggled to maintain organization until the low-level hodograph enlarged with the evening transition around 0000 UTC 20 April. From 0000 to 0400 UTC, the supercells in central Oklahoma produced 18 tornadoes, of which the NWS rated two as EF3 and four as EF2. The conditional environment with high-impact outcomes makes this an intriguing case for which to perform OSSEs.

3. Ensembles

We used the NSSL WoFS as the ensemble system for this study. WoFS is a rapidly updating regional convection-allowing ensemble data assimilation and prediction system, originally

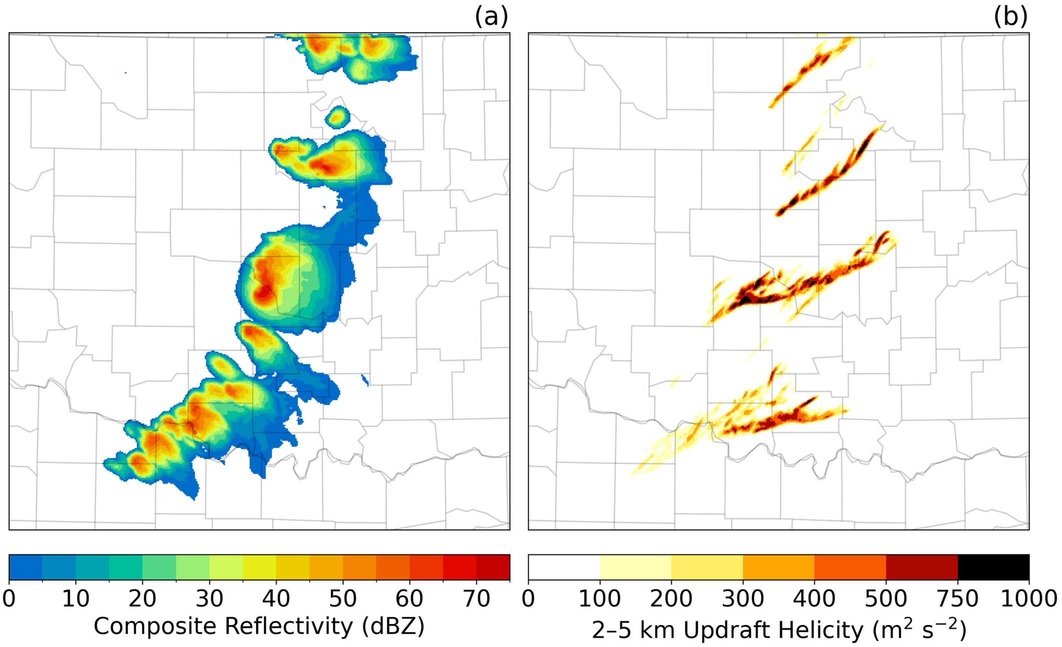
(a)
(b)
0
10
20
30
40
50
60
70
Composite Reflectivity (dBZ)
0
100
200
300
400
500
750
1000
2–5 km Updraft Helicity (m^2 s^{-2})

FIG. 2. (a) CR from the NRC valid at 0000 UTC 20 Apr 2023 and (b) storm-total UH25 swaths for the 19 Apr 2023 case.

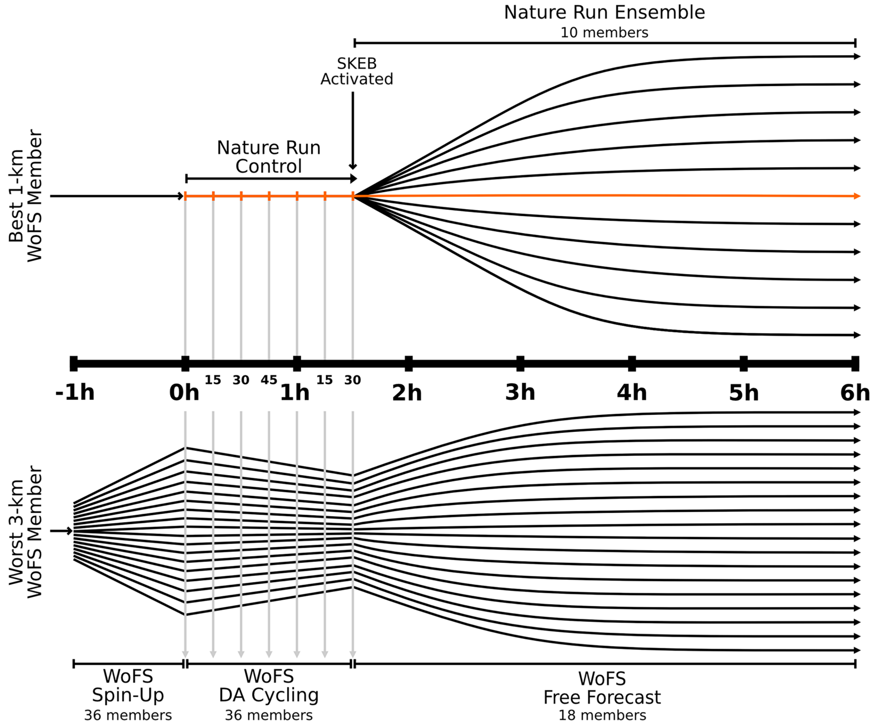

FIG. 3. Conceptual diagram of the multiensemble workflow. The orange line represents the NRC, the upper black curves represent the 10 NRE members, and the lower black curves represent the 18 experimental forecast ensemble members (SR or SRP). Vertical gray lines represent DA cycling times for the experimental ensemble, taken from the NRC.

designed to provide probabilistic forecasts of convective-scale atmospheric processes (Stensrud et al. 2009; Heinselman et al. 2024). The WoFS comprises 36 members based on version 3.9 of the Weather Research and Forecasting (WRF) Model with the Advanced Research version of WRF (ARW) dynamical core (Skamarock et al. 2008). Members are stratified based on particular physical parameterization schemes (see Table 1 in Kerr et al. 2023). The 36-member High-Resolution Rapid Refresh Data Assimilation System (HRRR-DAS; Dowell et al. 2022) provides the initial conditions, while the operational HRRR forecast is combined with large-scale perturbations from 18 members of the Global Ensemble Forecast System (GEFS; Zhou et al. 2022) to provide the lateral boundary conditions (Heinselman et al. 2024). The traditional WoFS (hereafter WoFS-3 km) numerical mesh is 900 × 900 km^2 with uniform horizontal spacing of 3 km and 51 stretched vertical levels. Observations, including radar reflectivity, radial velocity, and GOES-16 cloud water path, are assimilated into WoFS-3 km every 15 min. WoFS-3-km forecasts comprise the first 18 members and are up to 6 h in length, issued every 30 min, with an output frequency of 5 min (Heinselman et al. 2024). In addition to the traditional WoFS-3-km setup, an experimental 1-km version (WoFS-1 km) has been used in recent years (e.g., Wang et al. 2022; Kerr et al. 2023). In WoFS-1 km, a one-way nested domain is placed in the WoFS-3-km domain. This domain is 402 × 402 km^2 with a uniform horizontal spacing of 1 km and identical 51 stretched vertical levels. The WoFS-3-km domain provides lateral boundary conditions to the WoFS-1-km domain, while the HRRR-DAS provides the initial conditions. Our nature-run ensemble is based on WoFS-1 km, while our degraded ensemble is based on WoFS-3 km. The methodology is described below and visualized in Fig. 3.

a. Nature-run ensemble

We first created a nature-run control (NRC) simulation by subjectively choosing the "best" member of the WoFS-1-km ensemble at the time of interest—0000 UTC 27 February for C1, 2100 UTC 19 April for C2—and running a stand-alone 6-h free forecast from that point forward using only that member. The first 90 min of the NRC provided ideal observations to the coarser forecast ensembles during DA cycling (see section 3b), while the remaining four-and-a-half hours were used as the baseline for the nature-run ensemble (NRE).

The goal in creating the NRE was to capture the inherent predictability of the considered flow problem within the limitations of our "best available" modeling framework. In other words, we wanted to put in context the appropriateness of, or gains from, various OSSE configurations. Because we have already determined that the NRC represents the best model configuration to represent the physics of our problem, we added ensemble spread through uncertainty of our chosen physics schemes rather than through a multiphysics approach. One method to introduce such realistic model uncertainty is the stochastic kinetic energy backscatter (SKEB) scheme (Berner et al. 2009, 2011). In short, the

SKEB scheme accounts for subgrid uncertainty by introducing spatially and temporally correlated perturbations to the tendency terms of potential temperature and rotational components of the horizontal wind. In a radar data assimilation study using WoFS-1 km, Stratman et al. (2024) found that SKEB provided one of the best skill improvements and largest ensemble spreads as compared with other perturbation methods. We use the same SKEB parameter settings as listed in Table 1 of Stratman et al. (2024).

The SKEB scheme generates random number streams at model initialization (i.e., it cannot be arbitrarily turned on mid-simulation). For that reason, we ran 10 short one-and-a-half-hour realizations of the NRC with the SKEB scheme activated, each with a different seed supplied to the random number generator. The resulting perturbation fields were saved as SKEB restart files. We then applied the random fields to the stand-alone NRC simulation by pausing the run after 90 min, activating the SKEB scheme, supplying the restart files, and continuing the run for the remaining four-and-a-half hours. This process generated a 10-member NRE to accompany the NRC.

b. *Forecast ensemble*

After the NRC and NRE were created, we generated the experimental forecast ensembles by subjectively choosing the “worst” member of the WoFS-3-km ensemble at one hour prior to the nature-run time of interest—2300 UTC 26 February for C1 and 2000 UTC 19 April for C2. We made this choice to allow for the maximum possible assimilation benefit. The selected member was taken as the new mean of the experimental forecast ensemble. The remaining members of the WoFS-3-km ensemble were recentered on the new mean, resulting in a new 36-member WoFS ensemble. We retained the same settings as those used in the WoFS-3-km ensemble (see section 3a). We then imposed a spinup period to increase spread by freely advancing the forecast ensemble forward 1 h.

After the spinup period, we assimilated simulated observations from the NRC every 15 min for one-and-a-half hours (seven cycles) using the ensemble adjustment Kalman filter (EAKF; Anderson 2001) from the Data Assimilation Research Testbed (DART; Anderson et al. 2009). During this period, we assimilated both simulated radar reflectivity/radial velocity and simulated surface meteorological station observations for all experiments. For radar reflectivity/radial velocity, we 1) superobbed the radar reflectivity/radial velocity to 5-km spacing and 2) filtered observations with reflectivity of less than 20 dBZ and assimilated clear-air reflectivity values of 0 dBZ with 15-km spacing. For simulated surface stations, we assimilated surface pressure (at terrain height), temperature, dewpoint temperature, and horizontal winds. We further made use of prior inflation (Anderson 2009) and additive noise (Dowell and Wicker 2009; Sobash and Wicker 2015) with a 35-dBZ reflectivity threshold and a 10-dBZ innovation threshold to help maintain spread in the ensemble. Additional DA parameters, such as observation error variance, match those listed in Labriola et al. (2023). These settings composed the baseline experiments called surface radar (SR).

In additional experiments, we assimilated perfect (i.e., taken as is from the NRC with zero observation error assumed) vertical profile observations of temperature, dewpoint temperature, and horizontal winds interpolated, interpolated to each of the Oklahoma Mesonet surface stations. The perfect profile observations (see Fig. 4) were intended to identify the maximum improvement expected from assimilating vertical profile information from a profiling network, while also showing the utility of the NRE. We explored three profiling strategies with varying levels of data retention (not presented): full profiles, vertical spacing of 500 m, and vertical spacing of 1500 m. For the presented work, we chose thinned profiles with 500-m spacing for the perfect-profiler experiments, hereafter referred to as surface radar profiler (SRP). We plan future studies to explore the potential for objective thinning strategies and to use realistic simulators for the profiling observations.

Following data assimilation, we ran a four-and-a-half-hour free-forecast ensemble with the first 18 members of the analysis ensemble for comparison with the NRE. An important caveat to our experimental setup is that the nature run and forecast ensembles both used the ARW-WRF dynamical core and shared the same microphysics parameterization. This “identical twin” problem (present in other profiler network OSSEs; see, e.g., Kay et al. 2022) can lead to underrepresented model error and, consequently, an exaggerated observational impact. However, our use of the SKEB scheme in the NRE and the physics diversity in the SR/SRP ensembles somewhat mitigated this issue. While the identical twin problem remains a limitation of this study, we do not believe it undercuts the principles or utility of the presented NRE framework. We plan to reduce this issue in future studies by using a different forecast ensemble modeling framework. For instance, a WoFS ensemble based on the Model for Prediction Across Scales (MPAS; Skamarock et al. 2012) is currently under development at NSSL.

4. Verification

a. *Method*

We validated fields of composite reflectivity (CR), 0–2-km updraft helicity (UH02), and 2–5-km updraft helicity (UH25) using the ensemble fractions skill score (eFSS; Duc et al. 2013). For CR, we used the instantaneous 5-min output as is, while for UH02 and UH25 we took maximum values over a moving window to generate 30-min swaths with 5-min frequency.

$$\text{eFSS} = 1 - \frac{\dfrac{1}{N \times M}\sum_{n=1}^{N}\sum_{m=1}^{M}(O_{n,m} - F_{n,m})^2}{\dfrac{1}{N \times M}\left[\sum_{n=1}^{N}\sum_{m=1}^{M}(O_{n,m})^2 + \sum_{n=1}^{N}\sum_{m=1}^{M}(F_{n,m})^2\right]}, \quad (1)$$

where M is the number of ensemble members, N is the number of neighborhood windows, and O (F) is the observed (forecast) fraction of grid points that exceed a specified threshold within a given neighborhood window. In all cases, O is the NRC, while F is either from the NRE, SR, or SRP. For the latter two experiments, we used a simple bilinear interpolation to move the WoFS-3-km grid to a spacing of 1 km and then only considered

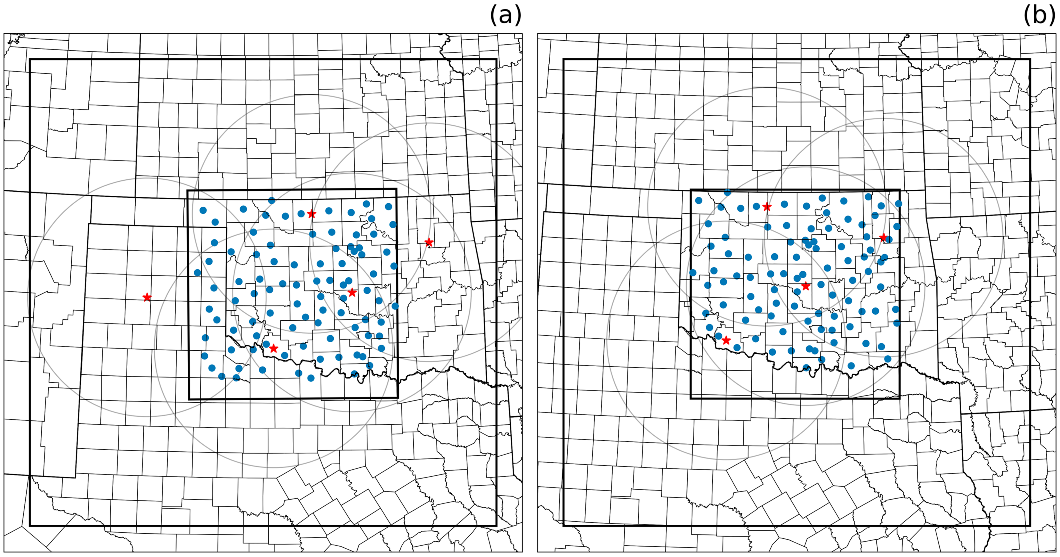

FIG. 4. Model domains and observation locations for the (a) 26 Feb 2024 and (b) 19 Apr 2024 cases. The inner box is the 1-km nature-run domain, and the outer box is the 3-km forecast domain. The red stars show the locations of simulated radars, and the light gray circles show the range of the radar observations. Observations are only generated for portions of the radar range that overlap the 1-km domain. The blue dots show the locations of simulated surface stations and profiler observations.

points where the interpolated grid overlapped the WoFS-1-km grid (i.e., within the inner domain box shown in Fig. 4). We caution that the bilinear interpolation worked here because the 3-km grid points exactly align with points on the 1-km grid. In cases where that is not true, this interpolator can reduce maximum values through smoothing and make threshold-based validations suspect.

For this study, we present eFSS results only for a neighborhood window size of 12 km, which is consistent with size thresholds used for reflectivity and helicity objects in object-based studies (e.g., Skinner et al. 2018). We computed intensity thresholds (see Table 1) across all forecast times and ensemble members for each experiment to remove climatological biases. Following, e.g., Kerr et al. (2023), we took CR thresholds as the 99th percentile and UH02/UH25 as the 99.9th percentile, all with zeroes included. The large differences for UH02 and UH25 threshold values between the NRC/NRE and SR/SRP experiments are due, in part, to the grid spacing of their native numerical meshes (i.e., UH is inversely proportional to horizontal grid spacing through the vertical component of vorticity). A limiting case of the eFSS formulation given by Eq. (1) occurs when both O and F are zero (i.e., there are no neighborhoods that exceed the threshold). Mittermaier (2021) discussed ways to overcome this undefined (perfect null) score, including adding a small noise term to the denominator or making a piecewise version of the formulation. However, as the authors in op. cit. note, these are purely mathematical solutions to get around mathematical inconveniences. These situations do not add physical meaning or aid in our interpretation of the solutions, so we instead opted to mask and omit eFSS for the perfect null case.

TABLE 1. Thresholds to compute eFSS for CR, UH02, and UH25. Values are computed across all forecast times and ensemble members at the 99th percentile for CR and 99.9th percentile for UH02 and UH25.

Case date	Experiment	CR (dBZ)	UH02 (m^2 s^{-2})	UH25 (m^2 s^{-2})
26 Feb 2023	NRC	53	221	546
	NRE	53	221	546
	SR	50	41	84
	SRP	50	46	106
19 Apr 2023	NRC	47	77	483
	NRE	49	71	492
	SR	26	6	66
	SRP	33	11	101

Because the purpose of the NRE is to establish the intrinsic predictability limits of a particular case, it is important that we understand whether the eFSS differences between experiments are meaningful. In other words, we want to know whether improvements or regressions that result from changes in the simulated observing network in the degraded ensembles are valid. At each forecast time, we followed a procedure similar to that described in Hamill (1999) to determine the statistical significance of these differences:

1) Compute the eFSS for SR and SRP ($eFSS_{SR}$ and $eFSS_{SRP}$).
2) Loop 10 000 times:
   (i) Construct two 18-member forecast ensembles (T_1, T_2) randomly composed of SR and SRP members. The T_1 and T_2 members retain their original ensemble membership order, but not their experiment membership itself. That is, both member 1 from SR and member 1 from SRP remain member 1 but are randomly placed

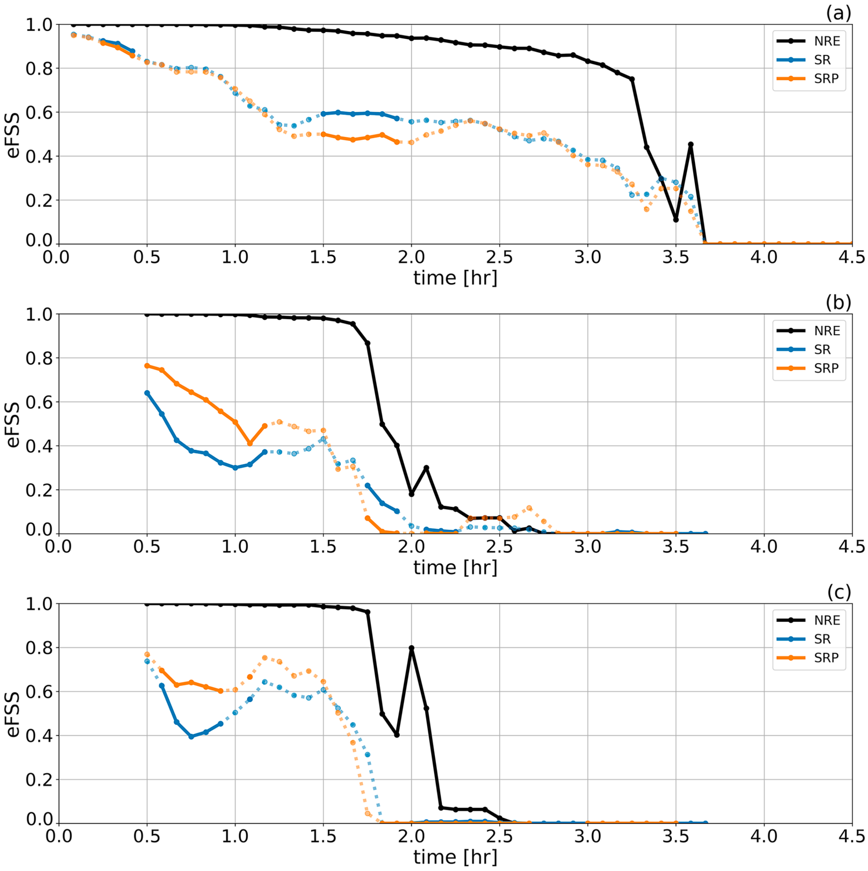

FIG. 5. The eFSS for (a) CR, (b) UH02, and (c) UH25 for the 26 Feb 2023 case. Black, blue, and orange colors correspond to the NRE, SR, and SRP experiments, respectively. The solid (dotted) lines for the SR and SRP cases correspond to times when their differences are (are not) statistically significant at the 95th percentile. Times are relative to the free-forecast period and span from 0130 to 0600 UTC 27 Feb 2023.

in either T_1 or T_2. In this way, the physics diversity is maintained while experimental changes are randomly distributed.

(ii) Compute the eFSS for T_1 and T_2 and save each to a list (eFSS_{T_1} and eFSS_{T_2}).

3) Compute $\Delta_E = |\text{eFSS}_{\text{SR}} - \text{eFSS}_{\text{SRP}}|$ if both SR and SRP have a valid eFSS. Otherwise, skip this time.

4) Compute $\Delta_T = |\text{eFSS}_{T_1} - \text{eFSS}_{T_2}|$ for all jointly valid eFSS values in the distributions. The size of Δ_T is N_T.

5) Compute $N_C = \sum_{n=1}^{N_T} \begin{cases} 1, & \text{if } \Delta_T(n) > \Delta_E, \\ 0, & \text{if } \Delta_T(n) \le \Delta_E. \end{cases}$

6) If $N_C/N_T \le 0.05$, then Δ_E is statistically significant at the 95th percentile.

b. Results and discussion

We present eFSS results for C1 and C2 in Figs. 5 and 6, respectively. In C1, the NRE eFSS curve for CR remains high for the first 3 h of free forecast, after which it drops to zero over the next 30 min. For UH02/UH25, the drop-off in eFSS occurs just shy of 2 h into the free forecast and reaches zero approximately 30–45 min later. In C2, the NRE eFSS curve for CR begins to drop off earlier, after approximately 2 h, but with a much slower decline. It remains high until approximately 3 h and does not reach zero until the end of the free-forecast period. The UH02/UH05 curves exhibit the same behavior as for CR, but with approximately 30 fewer minutes in their intrinsic predictability timeline. The behavior exhibited by the NRE suggests that CR and UH have different error-growth characteristics, resulting in an intrinsic predictability time scale of approximately 3 and 2 h, respectively. Accordingly, forecasters can expect the potential for observational improvements for the first few hours of a free forecast in the presented cases. Forecasts beyond that time frame may be instructive in gauging the broad continued evolution of storms, but the utility of the underlying DA adjustments will have already been exhausted. While the relative differences between the intrinsic predictability of the CR and UH fields are consistent between the two cases, their absolute values are case dependent (e.g., whether the storm mode is continuous or discrete).

Differences in the forecast ensemble eFSS between the SR and SRP experiments in C1 were not consistent for CR and UH

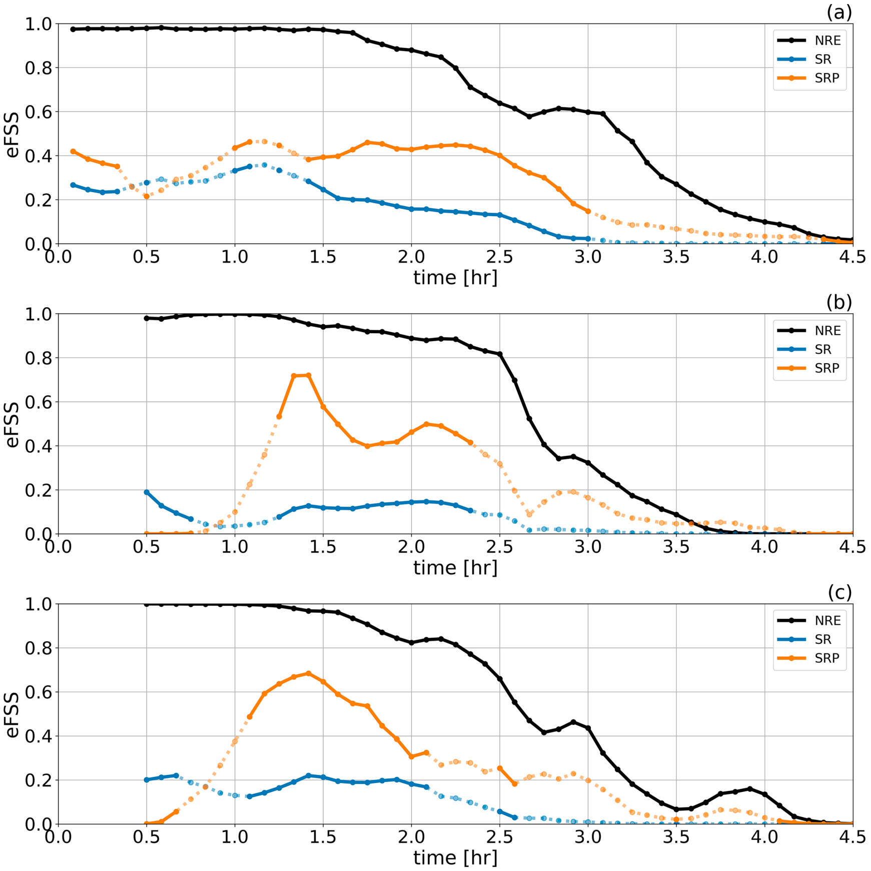

FIG. 6. The eFSS for (a) CR, (b) UH02, and (c) UH25 for the 19 Apr 2023 case. Black, blue, and orange colors correspond to the NRE, SR, and SRP experiments, respectively. The solid (dotted) lines for the SR and SRP cases correspond to times when their differences are (are not) statistically significant at the 95th percentile. Times are relative to the free-forecast period and span from 2230 UTC 19 Apr to 0300 UTC 20 Apr 2023.

fields. For instance, the differences between the experiments for CR were small for most of the free-forecast period. The curves steadily declined over the first 90 min and then flattened over the subsequent 90 min, of which only the differences over the first 30 min were statistically significant. During this time, perhaps counterintuitively, the inclusion of profilers resulted in a reduced eFSS. For the UH swaths, differences in eFSS were larger and were improved by the inclusion of profilers. The statistically significant differences were limited to the first 30–45 min of the free forecast. We show the mean values of eFSS over the first 2 h of the free forecast in Table 2 since that time frame represents the approximate lower limit of the intrinsic predictability of the NRE. We also show the mean values over that time in which the differences between SR and SRP were statistically significant to confirm that the relative behavior is consistent. We see that the NRE eFSS for CR is higher than it is for the UH swaths. Considering the speed and scale of the event, we suggest that the NRE better captured the general shape and characteristics of the convection but struggled in representing the location and intensity of rotation. In treating the NRE as the upper limit of “perfect,” we see that eFSS degrades slightly for CR and improves by a larger amount for UH with the inclusion of profilers. The experiments with profilers had slightly smaller analysis increments with time, but better spread and consistency ratios late in the DA period (not shown). Given that most members captured the size and continuous nature of the QLCS

TABLE 2. Average eFSS values for CR, UH02, and UH25. Values were computed across the first 2 h of free forecast from each experiment. Bolded values in parentheses represent the same averages, but only when taken during times in which the differences between SR and SRP were statistically significant to the 95th percentile.

Case date	Experiment	CR (dBZ)	UH02 (m^2 s^{-2})	UH25 (m^2 s^{-2})
26 Feb 2023	NRE	0.98 (**0.97**)	0.88 (**0.89**)	0.92 (**0.85**)
	SR	0.71 (**0.69**)	0.34 (**0.34**)	0.45 (**0.32**)
	SRP	0.62 (**0.62**)	0.42 (**0.46**)	0.51 (**0.43**)
19 Apr 2023	NRE	0.96 (**0.95**)	0.96 (**0.95**)	0.96 (**0.95**)
	SR	0.26 (**0.25**)	0.10 (**0.12**)	0.18 (**0.19**)
	SRP	0.38 (**0.40**)	0.31 (**0.37**)	0.40 (**0.44**)

event, it is likely that there was not much room for improvement in CR and that the DA process simply added noise. Conversely, improving the environment through the inclusion of profilers may have made the assimilation of radar radial velocity more effective, which in turn improved the representation of the UH fields.

The behavior of eFSS computed from the CR and UH fields is consistent in C2. Unlike in C1, there is no steady decline at the start of the free forecast, and in the case of SRP, we see an increase in eFSS over time within the intrinsic predictability window suggested by the NRE. The regions of statistical significance in differences between experiments are upward of 90 min for CR and 60 min for the UH fields—both longer than in C1. During these times, we observe significant improvement when profilers are included, which is supported in Table 2. We see that the NRE eFSS for CR and UH fields is equal during the first couple of hours of the free forecast, likely indicating that the model was able to represent both the convection and rotation features associated with this discrete event. Again, treating the NRE as the ceiling on what we can hope to reproduce with our system, we see that improving the structure of the environment through profilers led to enhanced forecasts. We also see that, as in C1, there is available overhead to further improve forecasts through, e.g., the addition of novel observations and/or a strategic network design of the existing observations. We also see that the intrinsic predictability is longer in this case and may expect improvements to persist further into the free forecast period than in C1.

5. Summary and conclusions

CAM ensembles are useful tools to provide probabilistic forecasts of severe weather events. Like all models, they suffer from errors associated with numerical approximations, simplifications within the physical parameterizations of subgrid-scale processes, and inaccurate initial environmental conditions. Data assimilation can help improve these errors by including information about the environment from observations. These observations, however, are often limited in scope, both in time and in space. Many new novel, portable observing platforms are designed with the hope of improving these coverage gaps, although many have not been widely assimilated in CAM ensembles to date. The use of OSSEs can help researchers understand the potential impacts of assimilating particular observations using different strategies (i.e., time and space configuration) without having to actually deploy real systems during expensive field campaigns.

Traditionally, a high-resolution simulation, or nature run, of a particular event is used to provide synthetic observations to the DA system and to serve as a validation dataset for the forecast ensemble. In this sense, the nature run is our best attempt to replicate the environment and evolution of a considered flow. However, a single deterministic nature run cannot provide adequate context to the forecast ensemble because its use in this manner attempts to absolve a false prophet of its many sins. That is, using a single nature run gives unrealistic expectations for how well the forecast ensemble can hope to perform given the specifics of a particular event. The use of an ensemble of nature runs can alleviate this issue. By introducing small changes in the initial environment and examining how errors grow in time, we can better understand the case-specific predictability associated with the control run. We showed that the errors from the ensemble saturate after a period of time, marking the bounds of the intrinsic predictability. This intrinsic predictability, in turn, helps calibrate results from a forecast ensemble, guide the design of OSSE experiments, and provide realistic expectations for potential gains expected from the inclusion of additional observations.

In this work, we introduced a framework to accomplish these goals and applied it to two different severe weather events as a proof of concept. We used the NSSL WoFS ensemble forecast system as our basis. We first constructed the NRC by subjectively choosing the best member from the WoFS-1-km ensemble at the time of interest and then running a 6-h free forecast from that point forward. The first 90 min of the NRC was used to provide observations to the DA system in the forecast ensemble. At the end of that period, we introduced small perturbations in the fields by turning on the SKEB scheme and then ran a four-and-a-half-hour free forecast—a process that was repeated 10 times. Next, we subjectively chose the worst WoFS-3-km member from 1 h prior to the time of interest. This member served as the mean of a new forecast ensemble, around which the remaining WoFS-3-km members were centered. After a 1-h free-forecast spinup, we assimilated synthetic observations from the NRC every 15 min for one-and-a-half hours using two experimental setups—both with radar reflectivity/radial velocity and surface meteorological stations, and one with the added inclusion of perfect vertical profiles. Following the DA period, we ran the ensemble to generate four-and-a-half-hour free forecasts. These forecasts were used as comparisons with the NRE.

We used the eFSS to compare the NRE with the NRC and to validate the forecast ensemble for fields of CR, UH02, and UH25. For eFSS, we only considered points within the innermost -nature-run domain and used a window size of 12 km. We took CR intensity thresholds as the 99th percentile and UH02/UH25 as the 99.9th percentile, all with zeroes included. We then used a bootstrap-without-replacement method to identify times in which the differences between the two OSSEs were statistically significant at the 95th percentile. We found that the intrinsic predictability suggested by the NRE was different for CR and UH fields—approximately 3 h for the former and 2 h for the latter. Further, this predictability was flow dependent, with the fast-moving continuous QLCS case C1 shorter than the more discrete supercell case C2. The inclusion of profilers in the SRP OSSE resulted in improved eFSS values for all fields and cases, except for CR in the C1 case. The gains were more modest in C1 and rather substantial in C2, with the eFSS more than doubling when profilers were included.

The framework has flaws. First, it suffered from the identical twin problem because both our nature-run and forecast ensembles used the WRF-ARW dynamical core. This can underrepresent model error and exaggerate observational impact. We attempted to limit this effect through the use of the SKEB scheme in the NRE and the multiphysics approach in the SR/SRP ensembles. Future implementations should use different models to eliminate this issue (e.g., using WRF-ARW for the nature-run ensemble and MPAS for the forecast ensemble).

Further, our framework had a potential "inverse domain" problem because the nature-run domain is smaller and sits inside the forecast ensemble domain. This could shorten the residence time of DA improvements (e.g., as seen in C1). Future implementations should flip the domains to boost the framework's effectiveness (e.g., using a standalone 2/3-CONUS simulation as the base for the NRE). While limiting, we do not believe these issues undercut the purpose of the presented NRE framework because every member was subjected to the same limitations. We suggest that the relative outcomes would remain and that perfect should not be the enemy of good for purposes of a proof of concept.

The use of an ensemble of nature runs was shown to serve as calibration for the forecast ensemble results by placing realistic expectations for a given flow's predictability within our chosen modeling system, while also demonstrating that there was potential room for improvement through the inclusion of different observations or by examining the network design of existing observations. This framework can serve to guide future studies by understanding what is possible for a chosen flow problem and modeling system and how to optimally assimilate novel observing platforms in time and space. This information can then help maximize efficiency and expenses in field studies by providing researchers with a better understanding of the involved trade-offs and expectations for the considered phenomena.

Acknowledgments. Funding was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement NA21OAR4320204, U.S. Department of Commerce, NOAA/OAR Weather Program Office Grant NA23OAR4590406, and the Warn-on-Forecast project. We thank Drs. Chris Kerr, Thomas Jones, and Patrick Skinner for technical help in constructing the simulation system and for many helpful discussions about the experiments and analysis. We thank Dr. Harold Brooks for useful discussions about approaches to statistical significance testing and Catherine Chapman for feedback on figure design. We also thank Benjamin Root for consultation about software tools and design. Analysis was conducted using version 3.10.14 of the Python programming language (Van Rossum and Drake 2009) with substantial use of the following packages: Matplotlib v3.8.4 (Hunter 2007), NumPy v1.26.4 (Harris et al. 2020), and SciPy v1.13.0 (Virtanen et al. 2020).

Data availability statement. The model output and code that we used in our study are not currently available in a publicly accessible repository. The model output and code that we used to generate the results are available upon request.

REFERENCES

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.

——, 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.

——, T. Hoar, K. Raeder, H. Liu, N. Collins, R. Torn, and A. Avellano, 2009: The Data Assimilation Research Testbed: A community facility. *Bull. Amer. Meteor. Soc.*, **90**, 1283–1296, https://doi.org/10.1175/2009BAMS2618.1.

Berner, J., G. J. Shutts, M. Leutbecher, and T. N. Palmer, 2009: A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ECMWF ensemble prediction system. *J. Atmos. Sci.*, **66**, 603–626, https://doi.org/10.1175/2008JAS2677.1.

——, S.-Y. Ha, J. P. Hacker, A. Fournier, and C. Snyder, 2011: Model uncertainty in a mesoscale ensemble prediction system: Stochastic versus multiphysics representations. *Mon. Wea. Rev.*, **139**, 1972–1995, https://doi.org/10.1175/2010MWR3595.1.

Dowell, D. C., and L. J. Wicker, 2009: Additive noise for storm-scale ensemble data assimilation. *J. Atmos. Oceanic Technol.*, **26**, 911–927, https://doi.org/10.1175/2008JTECHA1156.1.

——, and Coauthors, 2022: The High-Resolution Rapid Refresh (HRRR): An hourly updating convection-allowing forecast model. Part I: Motivation and system description. *Wea. Forecasting*, **37**, 1371–1395, https://doi.org/10.1175/WAF-D-21-0151.1.

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.

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.

Harris, C. R., and Coauthors, 2020: Array programming with NumPy. *Nature*, **585**, 357–362, https://doi.org/10.1038/s41586-020-2649-2.

Heinselman, P. L., and Coauthors, 2024: Warn-on-Forecast System: From vision to reality. *Wea. Forecasting*, **39**, 75–95, https://doi.org/10.1175/WAF-D-23-0147.1.

Hunter, J. D., 2007: Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.*, **9**, 90–95, https://doi.org/10.1109/MCSE.2007.55.

Kay, J., T. M. Weckwerth, W.-C. Lee, J. Sun, and G. Romine, 2022: An OSSE study of the impact of micropulse differential absorption lidar (MPD) water vapor profiles on convective weather forecasting. *Mon. Wea. Rev.*, **150**, 2787–2811, https://doi.org/10.1175/MWR-D-21-0284.1.

Kerr, C. A., B. C. Matilla, Y. Wang, D. R. Stratman, T. A. Jones, and N. Yussouf, 2023: Results from a pseudo-real-time next-generation 1-km Warn-on-Forecast System prototype. *Wea. Forecasting*, **38**, 307–319, https://doi.org/10.1175/WAF-D-22-0080.1.

Labriola, J. D., J. A. Gibbs, and L. J. Wicker, 2023: A method for generating a quasi-linear convective system suitable for observing system simulation experiments. *Geosci. Model Dev.*, **16**, 1779–1799, https://doi.org/10.5194/gmd-16-1779-2023.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 130–141, https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

——, 1969: Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.*, **26**, 636–646, https://doi.org/10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2.

McDonald, J. R., K. C. Mehta, D. A. Smith, and J. A. Womble, 2009: The enhanced Fujita scale: Development and implementation. *Forensic Engineering 2009: Pathology of the Built Environment*, The American Society of Civil Engineers, 719–728, https://doi.org/10.1061/41082(362)73.

Melhauser, C., and F. Zhang, 2012: Practical and intrinsic predictability of severe and convective weather at the mesoscales. *J. Atmos. Sci.*, **69**, 3350–3371, https://doi.org/10.1175/JAS-D-11-0315.1.

Mittermaier, M. P., 2021: A "meta" analysis of the fractions skill score: The limiting case and implications for aggregation. *Mon. Wea. Rev.*, **149**, 3491–3504, https://doi.org/10.1175/MWR-D-18-0106.1.

NWS, 2023a: The severe weather and tornado outbreak of February 26, 2023. U.S. National Weather Service, accessed 23 May 2025, https://www.weather.gov/oun/events-20230226.

——, 2023b: The severe weather and tornado outbreak of April 19, 2023. U.S. National Weather Service, accessed 23 May 2025, https://www.weather.gov/oun/events-20230419.

Skamarock, W. C., 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.

——, J. B. Klemp, M. G. Duda, L. D. Fowler, S.-H. Park, and T. D. Ringler, 2012: A multiscale nonhydrostatic atmospheric model using centroidal voronoi tesselations and C-grid staggering. *Mon. Wea. Rev.*, **140**, 3090–3105, https://doi.org/10.1175/MWR-D-11-00215.1.

Skinner, P. S., and Coauthors, 2018: Object-based verification of a prototype Warn-on-Forecast System. *Wea. Forecasting*, **33**, 1225–1250, https://doi.org/10.1175/WAF-D-18-0020.1.

Sobash, R. A., and L. J. Wicker, 2015: On the impact of additive noise in storm-scale EnKF experiments. *Mon. Wea. Rev.*, **143**, 3067–3086, https://doi.org/10.1175/MWR-D-14-00323.1.

Stensrud, D. J., and Coauthors, 2009: Convective-scale warn-on-forecast system: A vision for 2020. *Bull. Amer. Meteor. Soc.*, **90**, 1487–1500, https://doi.org/10.1175/2009BAMS2795.1.

Stratman, D. R., N. Yussouf, C. A. Kerr, B. C. Matilla, J. R. Lawson, and Y. Wang, 2024: Testing stochastic and perturbed parameter methods in an experimental 1-km Warn-on-Forecast System using NSSL's phased-array radar observations. *Mon. Wea. Rev.*, **152**, 433–454, https://doi.org/10.1175/MWR-D-23-0095.1.

Van Rossum, G., and F. L. Drake, 2009: *Python 3 Reference Manual.* CreateSpace, 242 pp.

Virtanen, P., and Coauthors, 2020: SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods*, **17**, 261–272, https://doi.org/10.1038/s41592-019-0686-2.

Wang, Y., N. Yussouf, C. A. Kerr, D. R. Stratman, and B. C. Matilla, 2022: An experimental 1-km Warn-on-Forecast System for hazardous weather events. *Mon. Wea. Rev.*, **150**, 3081–3102, https://doi.org/10.1175/MWR-D-22-0094.1.

WSEC, 2006: A Recommendation for an Enhanced Fujita Scale (EF Scale). The National Weather Service and Other Interested Users Revision 2 Rep., 111 pp., https://www.depts.ttu.edu/nwi/Pubs/EnhancedFujitaScale/EFScale.pdf.

Zhang, F., Y. Q. Sun, L. Magnusson, R. Buizza, S.-J. Lin, J.-H. Chen, and K. Emanuel, 2019: What is the predictability limit of midlatitude weather? *J. Atmos. Sci.*, **76**, 1077–1091, https://doi.org/10.1175/JAS-D-18-0269.1.

Zhou, X., and Coauthors, 2022: The development of the NCEP Global Ensemble Forecast System version 12. *Wea. Forecasting*, **37**, 1069–1084, https://doi.org/10.1175/WAF-D-21-0112.1.