Constraining Ensemble Forecasts of Discrete Convective Initiation with Surface Observations

LUKE E. MADAUS[a] AND GREGORY J. HAKIM

University of Washington, Seattle, Washington

(Manuscript received 14 October 2016, in final form 17 February 2017)

ABSTRACT

Predicting when and where individual convective storms will develop remains an elusive challenge. Previous studies have suggested that surface observations can capture convective-scale features relevant to the convective initiation (CI) process, and new surface observing platforms such as crowdsourcing could significantly increase surface observation density in the near future. Here, a series of observing system simulation experiments (OSSEs) are performed to determine the required density of surface observations necessary to constrain storm-scale forecasts of CI. Ensemble simulations of an environment where CI occurs are cycled hourly using the CM1 model while assimilating synthetic surface observations at varying densities. Skillful and reliable storm-scale forecasts of CI are produced when surface observations of at least 4-km—and particularly with 1-km—density are assimilated, but only for forecasts initiated within 1 h of CI. Time scales of forecast improvement in surface variables suggest that hourly cycling is at the upper limit for CI forecast improvement. In addition, the structure of the assimilation increments, ensemble calibration in these experiments, and challenges of convective-scale assimilation are discussed.

1. Background

There is interest in improving storm-scale forecasts of convective initiation (CI) to extend forecast lead times for potential severe weather hazards. Lilly (1990) challenged the meteorological community to gain the ability to explicitly forecast thunderstorms at city and county scales. Since that time, numerous studies have investigated CI in both modeling and observational contexts, for example in a component of the International H_2O Project (IHOP; Weckwerth and Parsons 2006). This work has successfully identified environments that support CI and meso- to synoptic-scale features that focus convective development. The results of this research now allow skillful forecasts of convective events with 0–3-day lead time over regions of 100–1000 km in scale within time windows on the order of hours (e.g., Kain et al. 2013; Duda and Gallus 2013). Nevertheless, storm-scale prediction of individual convective elements remains elusive.

Recently, the Warn-on-Forecast experiment at the National Severe Storms Laboratory (WoF; Stensrud et al. 2009) attempts to produce real-time, storm-scale probabilistic numerical forecasts of individual severe convective storms and their associated hazards. However, WoF has noted that their methods focus on predicting storm evolution once initiation has occurred given the difficulty of predicting storm-scale initiation. They specifically cite the lack of dense observations of the preconvective environment as a significant limitation to their methods, a sentiment echoed by numerous studies of CI (e.g., Stensrud and Fritsch 1994; Roebber et al. 2002; Mass et al. 2002; Fowle and Roebber 2003; Dabberdt et al. 2005; Gallus et al. 2005; Snook et al. 2015; Sobash and Stensrud 2015). Though radar observations have been used successfully in forecasting convective-scale features, their utility prior to and during CI is limited unless sensitive cloud radars are available (e.g., Markowski et al. 2006). Additionally, efforts to assimilate or apply satellite-based observations to CI have had mixed success. For example, Jones et al. (2015) finds that assimilating satellite-based estimates of liquid water path and ice water path can improve convective forecasts when assimilated prior to initiation, but the utility decreases in complex cloud fields and these observations can produce unwanted artifacts. Mecikalski et al. (2013) finds that the satellite characteristics of

[a] Current affiliation: National Center for Atmospheric Research, Boulder, Colorado.

Corresponding author: Luke E. Madaus, lmadaus@atmos.washington.edu

DOI: 10.1175/MWR-D-16-0395.1

developing storms vary substantially from case to case. However, Sieglaff et al. (2011) suggests that assimilating satellite-based cloud-top cooling estimates may be useful for CI forecasting.

Surface observations are one platform whose relevance for CI forecasting remains underexplored. Surface observation networks designed for synoptic-scale forecasting are ill-suited for constraining short-term, convective-scale forecasts (Sun et al. 2014). However, recent studies suggest that new surface observing platforms including crowdsourced observations could dramatically increase surface observation density in the coming years, and these observations have successfully been applied to convective events (e.g., Mass and Madaus 2014; Madaus et al. 2014; Muller et al. 2015; Jacques et al. 2015; Madaus and Hakim 2016, hereafter MH16). Furthermore, Sobash and Stensrud (2015) show that assimilating surface observations from local mesonets can also aid in CI forecasting. Yet, those studies that have connected an increased density of surface observations to improved CI forecasts have mostly focused on the ability of these observations to constrain broader mesoscale features (e.g., fronts or drylines) that can serve as focal points for CI. Investigating how dense surface observations may improve *storm-scale* CI forecasts is timely both for initiatives like WoF and to justify the utility of rapidly expanding surface observation sources.

Recently, MH16 investigated basic characteristics for a surface observing network capable of capturing the CI process. MH16 examined idealized simulations of CI in environments capable of convective initiation due to boundary layer processes alone; that is, without mesoscale or synoptic-scale forcing. They describe coherent patterns of anomalies that standard surface observations measure in the time period prior to CI. These results allow speculation on requirements for a surface observing network that could potentially capture the storm-scale CI process. Specific hypotheses include the following:

1) Correlation length scales for surface variables are approximately 4–5 km during the preconvective period. This suggests a spatial observation density of *no more than 4–5 km separating observations.*
2) The earliest detectable surface anomalies were seen in the 2-m temperature field as early as 60–80 min prior to CI. As such, to ensure that at least one assimilation cycle occurs during the preconvective period when anomalies may be detectable, *a cycling frequency of once per hour or greater is required.*

This study evaluates these hypotheses in observing system simulation experiments (OSSEs) using the MH16 simulation methodology. The primary goal of this study is to ascertain the density of surface observations that can provide skill to storm-scale forecasts of CI in an hourly cycling, kilometer-scale ensemble. By simulating CI in a regime where organized mesoscale forcing is absent, the ability of dense surface observations to capture storm-scale structure may be clearly evaluated. The remainder of this work is organized as follows. The methodology behind these experiments, including the model and data assimilation configuration, and forecast evaluation metrics are described in section 2. In section 3, we describe the environment simulated and the outcomes of experiments. A summary and thoughts on future work follow in section 4.

2. Methods

Here we describe the numerical model and data assimilation methods used to perform OSSEs, specific experiments performed, and the methods employed for measuring performance in these experiments. A description of the background environment used for initial conditions is given in section 3a.

a. Numerical model

The simulations here largely follow the simulation method used in MH16 with some exceptions. Here we use Cloud Model 1, revision 18.3 (CM1; Bryan and Fritsch 2002), as this revision includes modifications to the model output to enable data assimilation (http://www2.mmm.ucar.edu/people/bryan/cm1/CHANGES). From a chosen environmental sounding (section 3a), homogeneous initial conditions are generated for a 128 km × 128 km domain with 1-km horizontal grid spacing. Model vertical grid spacing is 80 m in the lowest 3.2 km and linearly increases to 500 m above 9 km with the model top at 18 km. Rayleigh damping of vertical motion is applied above 15 km. Parameterizations used include the Yonsei University (YSU) planetary boundary layer (PBL) scheme (Hong et al. 2006), NASA Goddard longwave and shortwave radiation (Chou and Suarez 1999, 2001), and Lin–Farley–Orville (LFO)/Goddard microphysics scheme with six microphysical classes (Lin et al. 1983; Tao and Simpson 1993). There is no topography. Land surface properties are uniform and defined by the U.S. Geological Survey (USGS) land-use category that is most prevalent within a 100-km radius of the sounding site. Boundary conditions are doubly periodic and the only external forcing is diurnal radiation. The simulation period focuses on the 10-h period following the time of the sounding launch, which, for the environment chosen, is long enough for CI to occur in the simulations. As in MH16, we only impose random, white-noise, gridpoint perturbations with a maximum magnitude of 0.2 K to the initial

potential temperature field at all model levels to promote random spatial variability.

b. Data assimilation

For data assimilation, we use a version of the Data Assimilation Research Testbed (DART; Anderson et al. 2009) ensemble adjustment Kalman filter (EAKF; Anderson 2001) that has been designed specifically to interface with the CM1 model. Only surface observations are assimilated, and we use horizontal spatial covariance localization (Gaspari and Cohn 1999) with a half-width of 15 km. While this is smaller than typically used for mesoscale data assimilation, results from MH16 indicate that this is three times the typical decorrelation length scale for surface variables in the pre-CI environment and is sufficient to ensure that even at the coarsest observation density examined (16 km), all grid points may be adjusted by at least one observation. Sampling error correction for a finite ensemble size (Anderson 2012) and adaptive covariance inflation (Anderson 2009) are used to maintain ensemble spread.

c. Generation of ensemble forecasts and observations

For these OSSEs, here we describe the generation of the truth or "nature" simulation, the ensemble forecasts being tested and the synthetic observations. Truth is given by a unique, deterministic, 10-h simulation that represents a single simulated evolution of CI in this environment. DART utilities are used to draw synthetic observations from the truth simulation. As the focus here is surface observations, five observation types are drawn, all from their corresponding fields within the CM1 model output:

- Surface pressure (PSFC; hPa)
- 2-m temperature (T2; K)
- 10-m U wind (U10; $\mathrm{m\,s^{-1}}$)
- 10-m V wind (V10; $\mathrm{m\,s^{-1}}$)
- 2-m specific humidity (Q2; $\mathrm{g\,kg^{-1}}$)

Observations are drawn hourly from the 1-km grid spacing truth simulation on a rectangular grid at three horizontal spacings: 16, 4, and 1 km. The observation locations are a subset of the model grid points for all densities, so no spatial interpolation is required and the values from the corresponding grid points in the truth simulation are directly used as the observation values. To simulate observation error, the observation values are randomly perturbed in accordance with a normal distribution of mean zero and variance equal to 1 unit^2. Observation errors of 1 unit^2 are similar to the error magnitudes used in operational data assimilation systems for surface observations (e.g., Burton 2013; Hu et al. 2013).

A 50-member ensemble is generated as the test forecast for the OSSEs. Initial ensemble diversity is provided by random small (maximum 0.2 K) temperature perturbations at the initial time as described above. This ensemble is integrated forward 1 h to the first assimilation time (hour 1). Each experiment begins with the same prior ensemble at hour 1. In the control ensemble, integration of all 50 ensemble members continues to the end of the 10-h period without assimilation. In all other experiments the ensemble is cycled hourly, assimilating observations as described in section 2d. After each cycle, a 6-h ensemble forecast is generated, initialized from the ensemble analysis at that time. These forecasts are compared with the corresponding truth simulation from which the observations were drawn and their skill is quantified as described in section 2e.

d. Experiments

Here we test the hypothesis regarding the minimum required observation density required to add skill to forecasts of CI by conducting four experiments. The control experiment does not assimilate any observations and provides a baseline forecast for comparison with all other experiments. We perform three additional experiments assimilating observations at spatial densities of 16 km (all_16km), 4 km (all_4km) and 1 km (all_1km). To estimate the robustness of the findings, these experiments are repeated 10 times using 10 different realizations of truth. To generate each truth realization, a different pattern of white noise is applied to the initial state as described above.

e. Measuring forecast skill

A variety of metrics are considered for evaluating the performance of the ensemble forecasts. To estimate how rapidly forecasts diverge from the truth simulation following data assimilation, the root-mean-squared difference (RMSD) between fields is computed as follows:

$$\mathrm{RMSD} = \left[\frac{1}{n}\sum_{i=1}^{n}(f_i - o_i)^2\right]^{1/2}, \tag{1}$$

where n is the total number of grid points, f_i is the forecast value at each grid point, and o_i is the value from the truth simulation at the corresponding grid point.

For computing probabilistic forecast skill from the ensembles, the Brier score (BS) is employed (Brier 1950), which is similar to RMSD but for probabilistic forecasts:

$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{n}(P_i - o_i)^2, \tag{2}$$

where n is the total number of locations where forecasts are produced, P_i is the forecast probability of occurrence of a specific event at each location, and o_i is a value of 1

or 0 if the event actually occurred at that location or not, respectively. In our analysis, P_i is determined by the fraction of ensemble members that forecast the event to occur. We specifically use the Brier score in the context of the Brier skill score (BSS) to compare the skill of one forecast against another:

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}_{\mathrm{fcst}}}{\mathrm{BS}_{\mathrm{ref}}}, \tag{3}$$

where $\mathrm{BS}_{\mathrm{fcst}}$ is the Brier score from the test forecast and $\mathrm{BS}_{\mathrm{ref}}$ is the Brier score from some reference forecast. A negative BSS indicates the test forecast has less skill than the reference forecast (here, forecasts from the control experiment), while a positive BSS indicates greater skill than the reference. A perfect forecast would have a BSS of 1.

When computing BSS values in our forecasts, we chiefly use them to evaluate probabilities of exceedance of reflectivity thresholds. To reduce errors associated with the exact timing of CI, we use a ±10-min neighborhood window when computing ensemble probability forecasts. Specifically, the probability of exceedance at a given time and grid point is computed by finding the fraction of ensemble members that exceed the threshold at that grid point within 10 min of the given time.

3. Results

a. *Convective evolution in truth and the control experiment*

For initial conditions, we use an environment based on the observed radiosonde profile at 1200 UTC 27 July 2014 in Lincoln, Illinois (KILX) (Fig. 1), which was one of many considered in MH16. This particular environment is chosen due to the relatively isolated nature of the convection that develops (described below, Fig. 2), which facilitates a clearer analysis of storm-scale CI prediction. In this environment, a nearly isothermal layer is present from the surface to around 900 hPa, with the lapse rate steepening above this level. The environment is nearly saturated from the surface up to 750 hPa, except for a drier layer between 925 and 875 hPa. Above 750 hPa, the profile is dry, and the tropopause level is reached at about 150 hPa.

From the onset of simulations (1200 UTC or 0700 local time), diurnal radiation heats the surface, and the boundary layer deepens as convective mixing becomes more vigorous. Surface-based parcels begin reaching their LCLs around 4–5 h and cumuliform clouds develop (not shown). Figure 2 shows the simulated composite reflectivity (CREF) evolution of one truth simulation as an example of the convective development that occurs

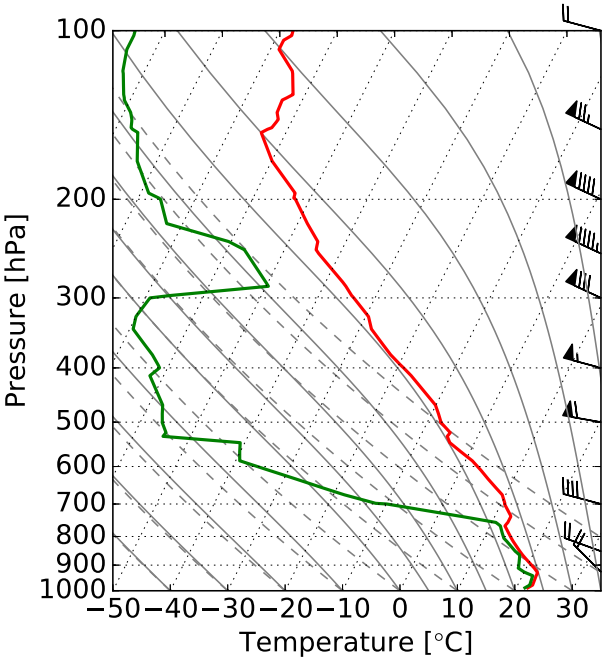

FIG. 1. The Lincoln, Illinois (KILX), 1200 UTC 27 Jul 2014 sounding used to generate initial conditions for the OSSE simulations.

similarly in all simulations. Here, CREF is computed using a diagnostic algorithm described in Stoelinga (2005). This algorithm first estimates the S-band (10 cm) reflectivity factor given the air density and rain, snow, and graupel mixing ratios at each grid point. To form the composite field, each vertical column is considered and the maximum reflectivity factor in each column is retained. This field is used as a proxy to identify individual convective features.

Initial deep cumulus formation is widely scattered. Several clouds reach their levels of free convection (LFCs) and there is a continuous transition to deep, precipitating convection between hours 6 and 7 (Fig. 2). The first precipitation accumulation of greater than 0.1 mm at the surface occurs at 6.8 h in this specific simulation. The storms continue to grow and mature through 8.5 h, while remaining discrete and relatively few in number. As the initial storms begin to decay, convergence on the edges of remnant cold pools sparks a new wave of initiation at 9.2 h into the simulation and these storms persist through the simulation end at hour 10.

The control ensemble simulation does not assimilate observations and, as such, is agnostic to matching truth. Convection develops along a similar timeline to the single realization shown above, but in random spatial locations in different ensemble members (not shown). This random spatial development is uniformly spread across the domain, and because of this, the domain-mean probability

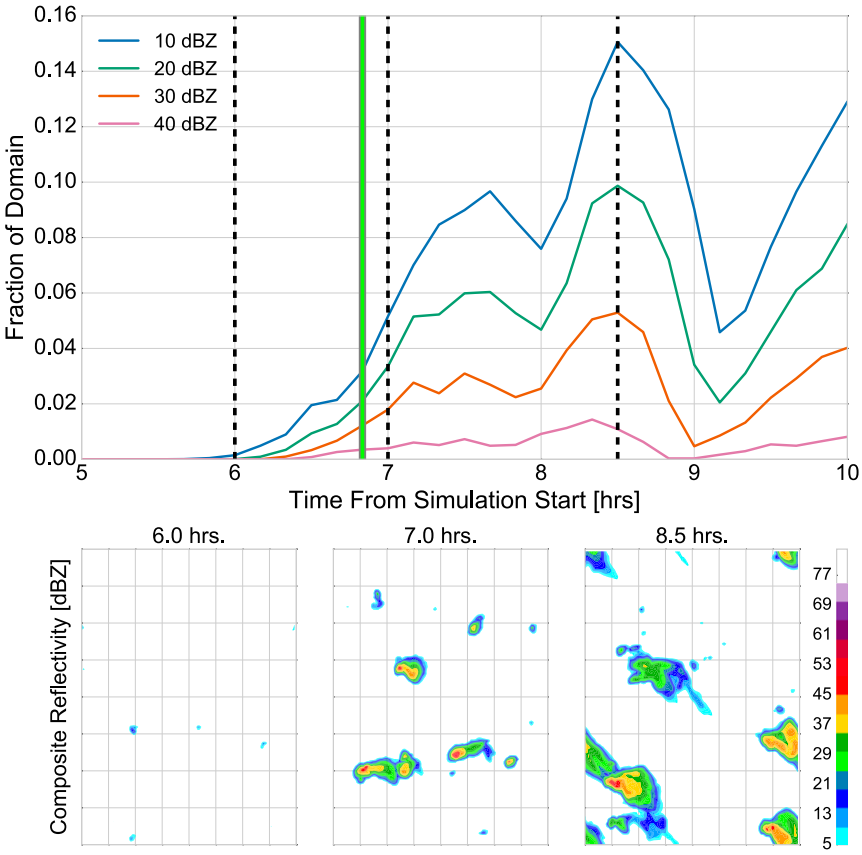

FIG. 2. Evolution of simulated composite reflectivity (CREF) in a single realization of the KILX environment. (top) The fraction of the total domain area where CREF exceeds various threshold values as a function of time is shown. The vertical green line indicates the time when the total accumulated surface precipitation first exceeds 0.1 mm at any point in the domain (CI time). (bottom three panels) Snapshots of the CREF field at the times indicated by the vertical dashed lines in the top panel. Gray lines show a 16-km grid, for scale.

of exceeding a particular CREF threshold over time is similar to the fractional coverage of the domain by points exceeding that CREF threshold (Fig. 2). Thus, the control ensemble serves as a reasonable "random" background forecast against which improvements may be measured.

b. Varying observation density

Three experiments are performed that assimilate all surface observation types at hourly frequency, but with different observation grid spacings: 16 km (all_16km), 4 km (all_4km), and 1 km (all_1km). These experiments are repeated with 10 different truth simulations, or realizations. Here we compare the forecasts from these experiments.

1) FORECAST SKILL IN A SPECIFIC REALIZATION

Figure 3 shows the probabilistic skill (BSS) in ensemble forecasts of CREF ≥ 10 dBZ during the period surrounding CI for a single truth realization. The threshold of 10 dBZ is chosen because, in these experiments, it can indicate the presence of deep clouds even prior to those clouds producing precipitation that reaches the surface. The time of the first appearance of 10 dBZ in the truth simulation is at 5.5 h from the beginning of the simulation, and the time when accumulated precipitation at any point in truth first exceeds 0.1 mm is shown by the vertical black line (0630). In our analysis we will refer to two periods: the CI period (between the appearance of 10 dBZ and the onset of precipitation, or CI time) and the mature convection period (after the onset of precipitation in truth). As seen in Fig. 2, for the first part of the CI period the areas where CREF $\geq$ 10 dBZ are limited to small, transient features. Robust convective clouds with lifespans that persist longer than a few minutes do not begin to occur until about 30 min prior to the onset of precipitation, which is consistent with the typical life cycle of airmass-type convection (e.g., Houze 2014). While the results in Fig. 3 are for a single realization, the patterns

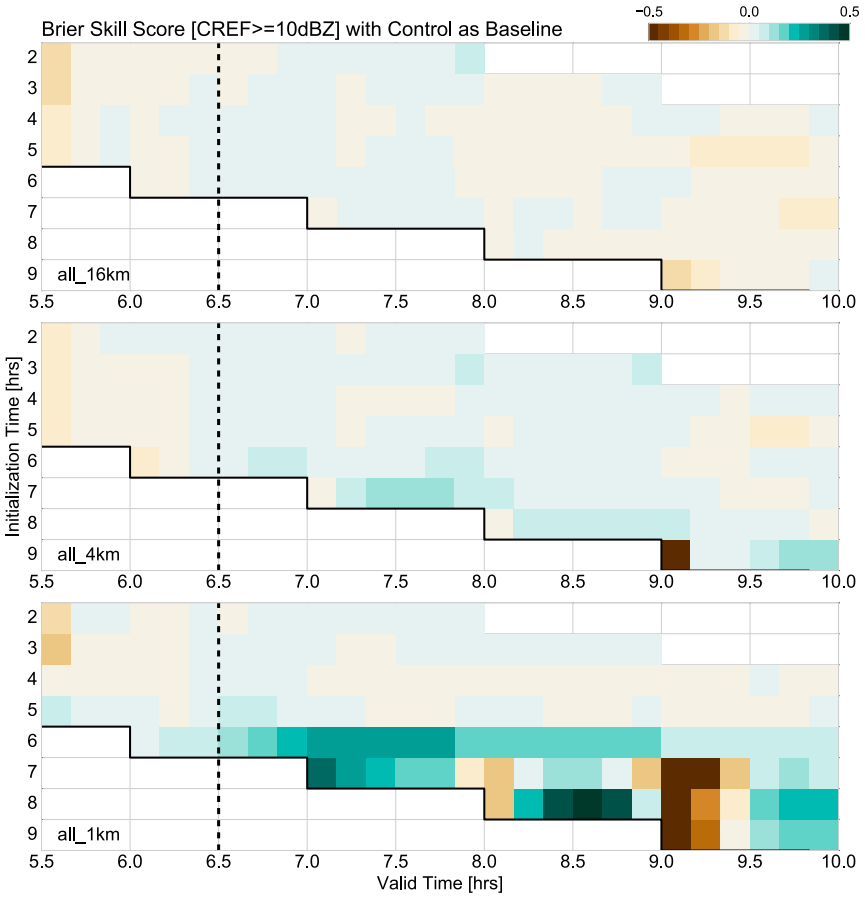

FIG. 3. Brier skill scores (BSSs) from a single realization of the KILX environment showing the skill in probabilistic forecasts of CREF $\geq$ 10 dBZ verified against the truth simulation. (from top to bottom) Forecasts from the experiments with observation grid spacings of 16, 4, and 1 km, respectively. In these panels, each row represents a forecast initiated at a different time and each column is a verification time. Skill scores are measured relative to the control ensemble forecast that assimilates no observations. The vertical black line indicates the time when precipitation first occurs in the truth (CI time; 6 h, 30 min).

shown here are qualitatively similar in all the realizations, as will be summarized later.

We discuss forecast skill for this realization in detail to describe several features of the forecasts. For forecasts valid during the CI period (left of black line, Fig. 3), the 16-km observation grid (all_16km) provides no clear benefit to the forecast, as skill score magnitudes remain close to zero throughout this period for all forecasts. The 4-km observation grid skill (all_4km) is low in magnitude; however, in the forecasts initiated at hour 6 there is a period of slightly positive skill through the time of CI. This suggests that 4-km density observations while the cumulus field is deepening may provide some skill to forecasts of initiation. The 1-km observation grid (all_1km) likewise shows small skill magnitudes during this period for forecasts initiated at hours 2–4. However, for forecasts initiated at hours 5–6, there is positive skill throughout the CI period, with skill increasing toward the time of CI. As such, the 1-km observation grid appears to be able to extend the positive skill of the 4-km grid by at least 1 h in this realization, and provides positive skill at predicting CI locations when observations are assimilated while the cumulus field is deepening.

In the mature convective period after CI has occurred (right of the vertical line, Fig. 3), all_16km still provides no clear improvement over the control forecast. However, all_4km forecasts initiated at or after hour 6 have positive skill, particularly within the first hour of each forecast. In general, much higher skill scores are seen throughout this period in the all_1km experiment. However, there are notable skill "dropouts" at around 8 and 9 h for forecasts initiated after CI has occurred. These correspond to times when the overall coverage of storms is at a relative minimum (Fig. 2). Additionally, the skill in these forecasts

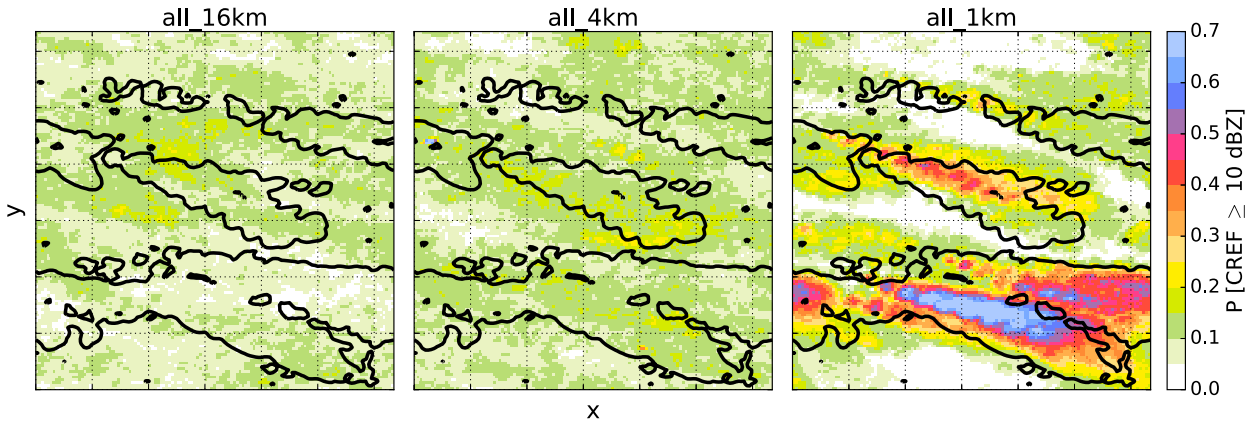

FIG. 4. The ensemble probabilities of exceeding 10 dBZ during the first 2 h of the forecast initiated at hour 6. Shown are maps of the entire simulation domain for a single realization of the (left) all_16km, (middle) all_4km, and (right) all_1km experiments. The black contours outline where the 10-dBZ threshold was exceeded in the truth simulation. The dashed black grid shows a 16 km × 16 km scale.

becomes positive as the forecast progresses. This appears to be related to the spinup time to develop convection in response to the assimilation increments (not shown). We comment further on this and other challenges of convective-scale assimilation below.

To better visualize what these skill scores mean in terms of an actual convective forecast, Fig. 4 compares forecasts initiated at a single time in the CI period—hour 6, 30 min prior to precipitation onset—for each of the grid spacing experiments. The probability "swaths" in Fig. 4 show where the ensemble forecasts place the likely locations of the storm cores (here, CREF ≥ 10 dBZ) as they develop and evolve throughout the first 2 h of this forecast. These may be compared with the true swaths outlined in black. The all_16km ensemble members do not have strong consensus on the locations of storms and there is no clear correspondence between its forecast and the truth. The all_4km ensemble has slightly elevated probabilities (>10%) that loosely correspond with some storms in the truth run, namely in the central part of the domain. However, the forecast poorly captures some storms and completely misses others, which decreases confidence in its indication of storm development. The all_1km ensemble captures the majority of the storm swaths well. There is good agreement on the existence and location of nearly all of the major storm elements in the truth run, with high confidence in the locations of several storm cores. There is some discrepancy over the exact path and evolution of individual storms, but this forecast highlights the likely storm development and evolution well.

2) FORECAST SKILL ACROSS ALL REALIZATIONS

To examine these skill improvements further, we consider the variability in forecast skill with forecast lead time in different realizations of all_16km, all_4km, and all_1km. Considering skill across 10 different realizations helps develop a composite picture of how skill varies with observation density and tests the sensitivity of these results to a particular truth simulation.

For each realization of each experiment, we consider the forecasts initiated each hour as the ensemble cycles. For each forecast, we compute both the lead time of the forecast (the difference between the CI time in truth and the forecast initialization time) and the mean BSS for ensemble probability forecasts of CREF ≥ 10 dBZ within 30 min of the time of CI. Since the onset of precipitation (the time of CI) varies by as much as 40 min in different truth realizations, a broad spectrum of forecast lead times are represented in these different realizations. Figure 5 shows the average forecast skill around CI time as a function of forecast lead time for all forecasts from all realizations of all experiments. We overlay a 1-h moving average of the skill scores from all realizations of each experiment.

For forecasts with lead times greater than 60 min, all experiments have skill scores close to zero. Within 1 h of CI, forecasts in the all_16km experiments continue to show near-zero skill. In the all_4km forecasts, there is a weak trend of increasing skill in this period as the time of CI approaches, but several forecasts still have no skill around CI time. On average, the all_1km forecasts initiated within 1 h of CI time have higher skill than the all_4km or all_16km forecasts. However, this skill improvement is not robust as forecast lead time decreases. Some all_1km forecasts still have similar skill to all_4km or all_16km forecasts even with lead times of less than 1 h. Nevertheless, there is increased potential for skillful forecasts when a 1-km observation grid is assimilated within 1 h of CI. This 1-h window when increased CI

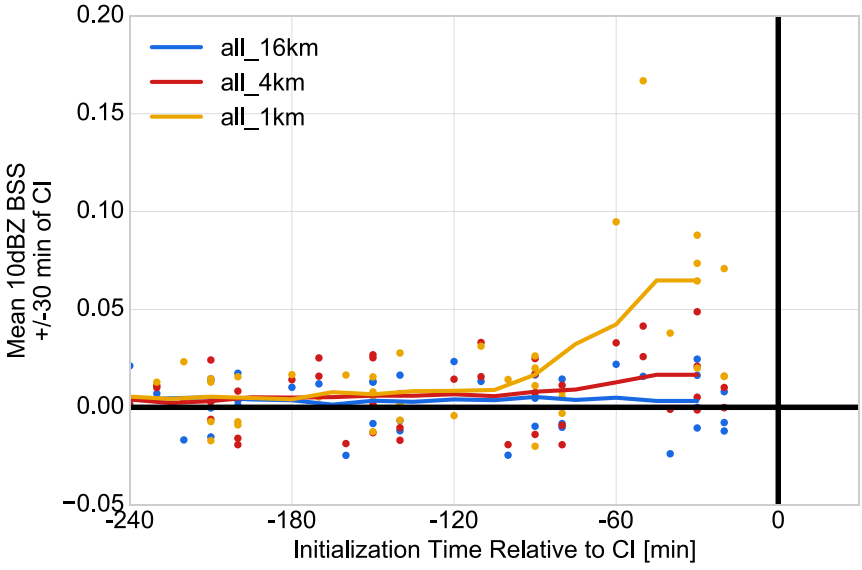

FIG. 5. Mean Brier skill score within 30 min of the time of CI as a function of forecast lead time for each forecast from each realization of each experiment. BSSs are shown from forecasts in all realizations of the all_16km (blue), all_4km (red) and all_1km (gold) experiments. The solid lines represent a 1-h moving average for each experiment.

forecast skill is possible agrees well with previous estimates of the storm-scale predictability of CI (e.g., Droegemeier 1990; Weckwerth 2000; Zhang et al. 2003).

We consider now the reliability for probabilistic forecasts of locations of convective storms. Figure 6 compares the ensemble probability of CREF $\geq$ 20 dBZ to the observed frequency of occurrence for the first 2 h of all forecasts initiated at hour 6, which spans the CI time in all realizations. We use the 20-dBZ threshold here to highlight more developed storm cores in Fig. 6. The all_16km and all_4km ensembles produce only low probabilities in all realizations, and these generally overpredicted the rate of occurrence. This qualitatively agrees with the results in Fig. 4. Going from low (16 and 4 km) to high (1 km) density observations, the maximum probabilities produced by the ensemble increase, which illustrates smaller spread and greater agreement among the ensemble members on potential storm locations when higher density observations are assimilated. The all_1km forecasts show good, but overconfident reliability, particularly for probabilities greater than 25%. The overconfidence in all experiments suggests poor ensemble spread calibration, which will be discussed further in section 3e.

c. *Assimilation increments*

MH16 describe evolving patterns in surface fields prior to and during CI and suggest that increments of surface fields matching these patterns may be able to constrain convective development. To explore this, we examine a vertical cross section through the region with the largest 2-m temperature increment during the hour 6 assimilation (within the CI period) of a single realization of all_1km. Figure 7 shows this cross section of prior and posterior ensemble mean cloud water; increments in cloud water, T2, Q2 and U10; and the cloud water from truth at that time.

In the center of this cross section (at x = 30 km), the prior ensemble mean has a cloud extending from 1 to 5 km in the vertical and approximately 8 km wide (Fig. 7, top panel, shading). In this same area, assimilated surface observations produce a negative T2 increment and opposing wind increments (positive to the left, negative to the right) that would promote increased convergence (or reduction in divergence) (Fig. 7, middle panel). These temperature and wind adjustments agree well with cloud shadowing and low-level convergence anomalies associated with a developing convective cloud at this location, discussed in MH16. In response to the assimilated surface observations, there is positive increment in cloud water in this cloud object (Fig. 7, top panel, contours), which increases cloud water in the posterior ensemble mean (bottom panel, shading). As anticipated, the true state does indeed have a cloud object at this location (bottom panel, black contours) marking a reasonable translation of surface increments to cloud field adjustments. Another combination of negative T2 increment, convergent wind increment, and correctly augmenting a cloud object may be found in the cloud at x = 5 km in the cross section.

Likewise, the converse of these surface increments leading to the *removal* of cloud objects from the ensemble mean may also be seen in Fig. 7. This is most notable for

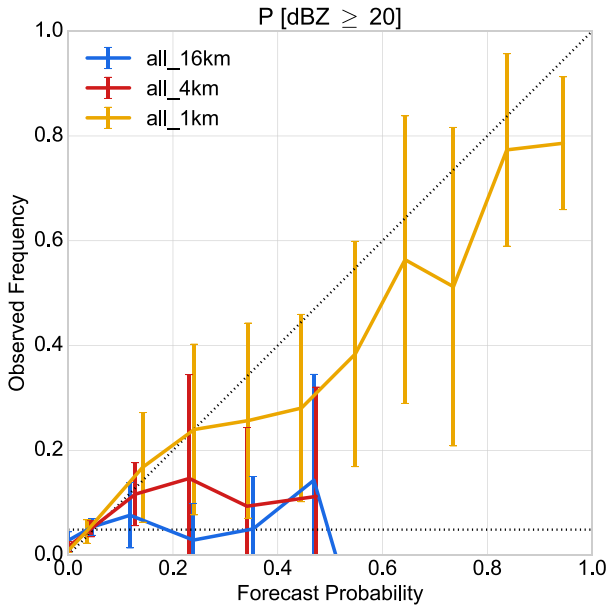

FIG. 6. Reliability diagram comparing the ensemble probability of CREF ≥ 20 dBZ to the observed frequency for the first 2 h of forecasts initiated at hour 6. The mean of all 10 realizations of hour 6 forecasts is shown for the all_1km (gold), all_4km (red), and all_16km (blue) experiments. Vertical bars indicate one standard deviation of the reliability across all realizations. The horizontal dotted line shows the climatological rate of occurrence (0.06) during this time, while the sloped dotted line indicates the ideal, 1-to-1 line.

the cloud object in the prior ensemble mean between 45 and 52 km in the cross section. Here, positive temperature and divergent wind increments underlie a cloud in the prior ensemble mean. The resulting increment in cloud water is negative, and in the posterior mean (bottom panel) this cloud is largely (and correctly) removed from the analysis.

While these adjustments agree with expectations from MH16, there are locations where the surface increment pattern does not correspond exactly as anticipated with the presence (or absence) of a cloud object. For instance, despite a cloud object present in truth between $x = 9$ and 14 km in Fig. 7, there is no clear signal in the temperature increments in that region, while the wind increments show weak divergence. Complicating the response to this cloud is the lack of a cloud object at this location in the background ensemble mean (Fig. 7, top panel). This highlights an additional challenge to storm-scale data assimilation: if no ensemble members have nonzero cloud water at a given location, the assimilation cannot produce a cloud. This can limit assimilation effectiveness when using a finite ensemble to estimate background covariances.

d. Lifetime of assimilation adjustments

Considering ensemble cycling frequency, MH16 hypothesize that at least hourly cycling is required to capture the CI process. Because of computational limitations, we have only tested assimilation for hourly cycling. We assess the sensitivity to hourly cycling by examining the duration of improvements to the forecast following assimilation and how this changes with observation density. Figure 8 shows the RMSD of various surface fields between the ensemble member forecasts initiated at hour 6 (within the CI period) and the truth simulation for a single realization of the all_16km, all_4km, and all_1km experiments. Only a single realization is shown here to highlight specific features of these differences, but the patterns and time scales in Fig. 8 are similar across all realizations (not shown). To evaluate the significance of the differences between experiments, we employ a simple t test of the difference in ensemble mean RMSD using the standard deviation of ensemble member RMSDs. Differences with greater than 95% confidence are considered significant.

In the all_16km and all_4km experiments, the ensemble mean RMSD for T2 (Fig. 8, upper left), Q2 (lower left), and U10 (lower right) is slightly lower in the 4-km experiment. The all_16km − all_4km difference is only significant for the first 30 min for U10 and Q2 and is not significant for T2. The all_1km experiment shows large reductions in RMSD during the initial period of the forecast. Reduction in RMSD from the all_4km to all_1km ensembles remains significant for 40 min in T2 (upper left) and 80 min in U10 (lower right) and Q2 (lower left). The additional information provided by the 1-km density observations contributes positively to surface forecasts for 40–80 min after assimilation. This suggests that hourly cycling is likely at the upper limit of useful cycling frequencies for storm-scale updates.

The RMSD values for the surface pressure field (PSFC) show a different pattern than the other variables. Instead of higher observation density reducing the error in PSFC for the early part of the forecast, an initial spike in PSFC RMSD grows more pronounced as observation density increases. Coherent oscillations among ensemble members in PSFC RMSD suggest gravity wave noise as the source of these errors (Anderson et al. 2005). After assimilation there is a rapid adjustment to balance any nonphysical features introduced by assimilation. These errors are dispersed as gravity waves. Over time, the Rayleigh damping used in these simulations reduces this signal. Though gravity wave noise would also affect the surface wind and temperature fields, PSFC is both highly constrained in this environment (as there is no synoptic or broad mesoscale variability) and is sensitive to perturbations throughout the vertical column. The relatively small pressure perturbations induced by deepening cumulus

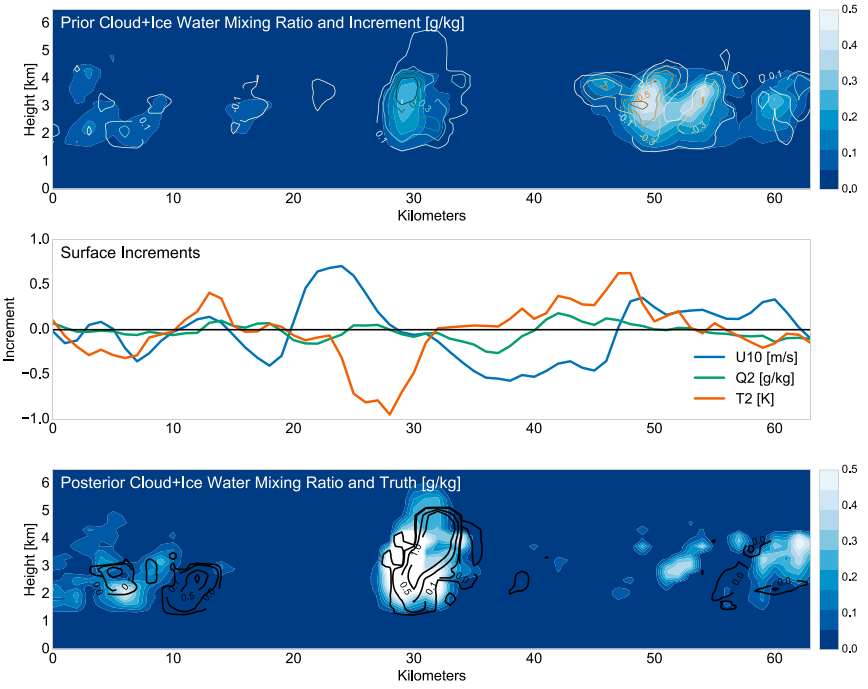

FIG. 7. Cross section of (top) prior (shading) and increment (contours) and (bottom) posterior (shading) and truth (black contours) ensemble mean cloud + ice water mixing ratio values for the assimilation at hour 6 with 1-km observation density. (middle) Increments of U10, Q2, and T2 along the same cross section are shown.

clouds [O(1–10) Pa; MH16] are of similar or smaller magnitude than the PSFC differences seen in Fig. 8 and the gravity waves dominate the signal. In contrast, T2 and U10 observations are mostly insensitive to perturbations above the boundary layer, and the magnitude of signals induced by deepening clouds [T2: O(1) K, U10: O(1–2) m s^{-1}; MH16] is likely larger than the errors induced by the gravity wave noise. As such, this noise is mostly manifested in the PSFC field.

The PSFC gravity wave noise increases with increasing observation density and is coherent among ensemble members. This indicates that as observation density increases, the ensemble fit is increasingly constrained to the (perturbed) observations and deviates more from the balanced background state, amplifying the gravity wave response. Dispersion of these imbalances may interfere with the spinup of convective-scale features following the assimilation. In general, this result highlights a shortcoming of using surface pressure observations to inform storm-scale features in a numerical model—initial condition imbalances can generate gravity wave noise with pressure signals that overwhelm any meteorological pressure signal on the convective scale.

These oscillations in the PSFC RMSD agree with Anderson et al. (2005), who also observe high-frequency gravity wave oscillations when dense surface pressure observations are assimilated on an hourly basis in a general circulation model. They note that these oscillations are in phase in all ensemble members, similar to what is seen here, and these can adversely affect assimilation performance. However, Anderson et al. (2005) observes that assimilating more frequently (not allowing errors to grow significantly in the ensemble) can reduce gravity wave oscillations. This suggests that, despite skillful CI forecasts with hourly cycling in these experiments, further improvement may be possible with more frequent cycling.

e. *Ensemble calibration*

While the primary forecast objective—constraining ensemble forecasts of storm features—appears possible at high surface observation densities, here we examine another performance measure: ensemble calibration. One calibration measure is the inflation factors generated by the time-varying adaptive inflation (Anderson 2009) employed in these experiments. The inflation factor estimates the degree to which the ensemble prior variance must be increased such that the error between the observations and the ensemble mean prior estimate of these observations remains consistent with the sum of ensemble and observation error variance. Figure 9 shows the

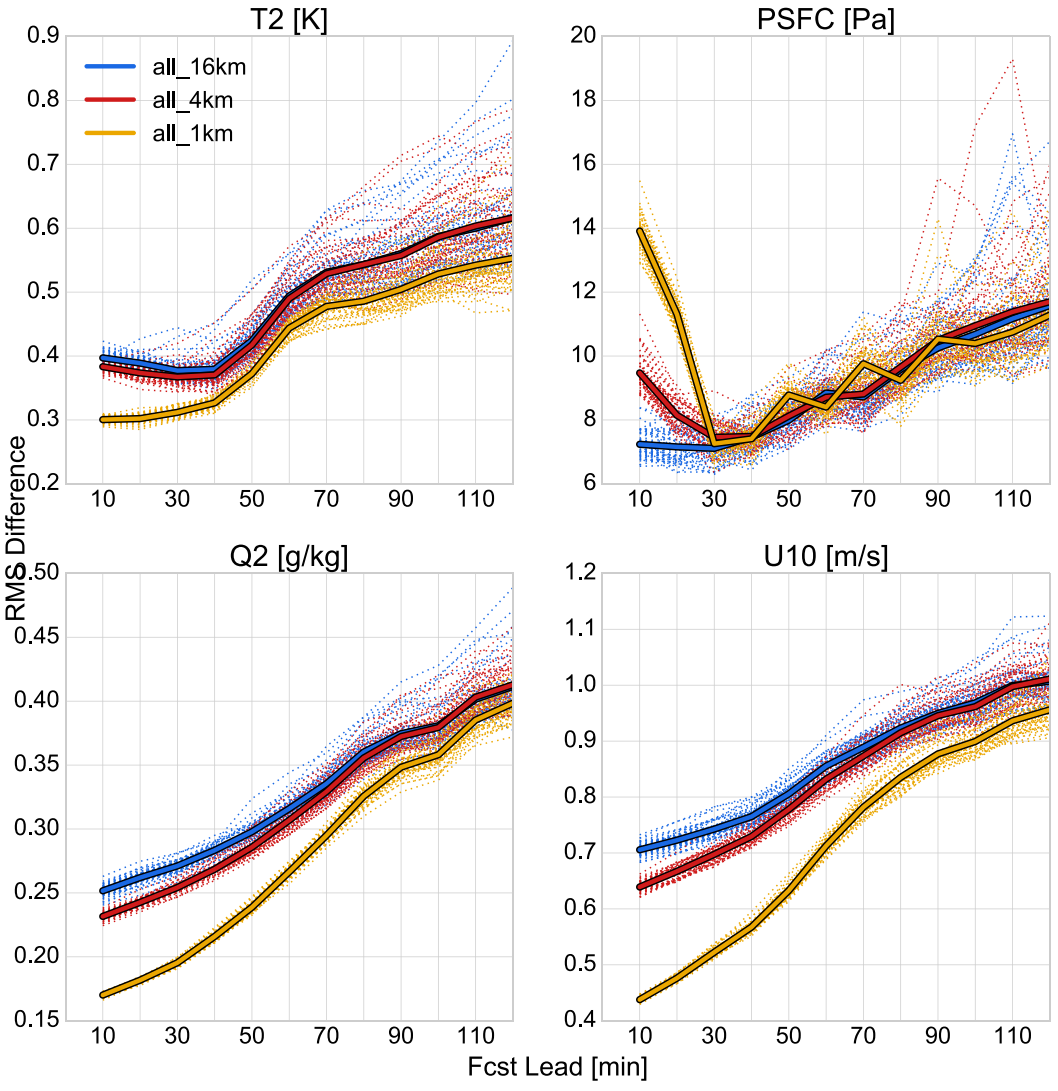

FIG. 8. RMSD between ensemble members and the truth simulation for selected surface variables of the forecast following the hour 6 assimilation for a single realization of truth. The all_16km, all_4km, and all_1km experiments are shown. Individual ensemble members are shown in the dotted lines, while the ensemble means are solid lines.

domain-averaged inflation factor applied to the T2 and PSFC fields as a function of assimilation time for a single realization of the three observation density experiments. In Fig. 9, for both variables, the inflation factors increase as observation densities increase. This indicates that though increased observation density reduces ensemble spread about a particular forecast solution (e.g., Figs. 4 and 8), the error growth in the ensemble about this solution is suboptimal, requiring larger inflation factors to avoid filter divergence. The differences between the necessary inflation factors for each observation density become noticeable by the assimilation at hour 3, but increase approximately exponentially as convective clouds develop and mature. More careful consideration of methods to control ensemble variance (e.g., localization radius, inflation methods) is likely required to maintain good ensemble calibration as convection develops.

4. Summary

These experiments tested the ability of dense surface observations to provide skill to ensemble forecasts of CI in a challenging scenario where organized meso- or synoptic-scale forcing was absent. The OSSEs made a perfect-model assumption and assumed unbiased observations with perfect knowledge of observation error. Under these conditions, experiments testing the impact of observation density supported the hypothesis that observation spacing of 4 km or less is required to

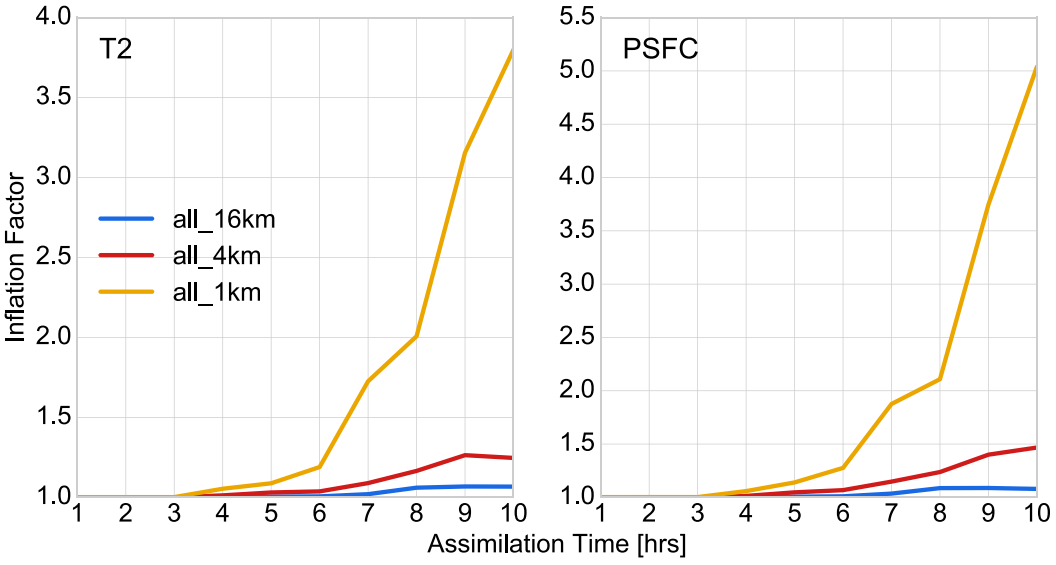

FIG. 9. Domain-mean inflation factors from the 16-km (blue), 4-km (red), and 1-km (gold) experiments as a function of assimilation time for (left) 2-m temperature (T2) and (right) surface pressure (PSFC).

constrain CI forecasts with hourly cycling. Forecasts of CI showed no improvement over control forecasts with 16-km observation grids, marginal improvements with a 4-km grid, and larger improvement with a 1-km grid. However, on a case-by-case basis the relative improvement between 4- and 1-km density was variable, even within an hour of CI. At 1 km observation density CI forecast skill was only improved in forecasts initiated within one hour of CI time; assimilating observations at earlier times had no noticeable impact on convective-scale forecasts of subsequent CI. In the 1-km experiments, ensemble forecasts of convective features during the CI period had good reliability, though were slightly overconfident.

Assimilation increments during the hour prior to CI (when improved skill was possible) were examined to investigate how the surface observations were adjusting the ensemble state. Changes to the cloud field were associated with patterns of surface increments that generally corresponded to those expected from MH16. Specifically, areas with negative 2-m temperature and convergent 10-m wind increments also had positive cloud water increments in the atmosphere above. The lack of CI forecast skill in forecasts initiated before the cloud field developed further suggests that surface observations are most useful once cumulus formation has begun.

Forecasts of surface fields in the 1-km observation density forecasts had lower errors than in the 4- or 16-km observation density forecasts. The lower errors with the 1-km observation grid persisted for 40 min (for temperature) to over an hour (for wind and moisture), suggesting that hourly cycling is at the margin for enabling improvements to CI forecasts when used with a very high density observing network. Gravity wave noise was observed following assimilations during the CI period, which may have reduced the assimilation effectiveness.

Anderson et al. (2005) suggest more frequent assimilation could reduce this noise, but cautions that frequent assimilation (on the orders of tens of minutes or less) can also introduce instabilities in the model that prevent successful simulations. There have been recent examples of surface observing networks being successfully used for 5-min update cycles (e.g., Sobash and Stensrud 2015) (albeit at much lower observation densities than tested here). We attempted to test 5-min update cycles with our experiment setup, however, the ensemble members became numerically unstable following assimilation once the cumulus field had begun to develop and the experiments were unable to be completed. We are currently investigating what adjustments may need to be made to the CM1 configuration or the data assimilation methods to enable more frequent assimilation at these densities. Because higher frequency observations may compensate for reduced spatial density, additional experiments assimilating dense surface observations at high temporal frequency are planned.

While these experiments showed skill in constraining CI forecasts, we found that the ensemble calibration became increasingly suboptimal as convection developed in these experiments. This occurred despite the use of techniques designed to mitigate this (e.g., covariance localization, adaptive inflation, and sampling error correction). Large inflation factors were required

for hourly assimilation of 1-km density observation observations (factors of greater than 5 as the convection matures). Additional experiments (not shown) suggested that a better-calibrated localization radius may improve ensemble calibration. However, to further improve ensemble calibration, techniques like localization and inflation need to be more thoroughly evaluated for rapidly evolving convective scenarios.

It should be emphasized that factors such as model error and observation bias are not considered in these experiments. In operational practice, these effects can detract from forecast performance and, as such, even the high densities of observations required here may be optimistic. However, while there are many environments and scenarios where CI may occur in the absence of organized larger-scale forcing as simulated here (such as warm-season diurnal convection in the southeastern United States), CI often occurs in the relation to mesoscale features such as fronts, drylines, and cold pools (e.g., Owen 1966; Ziegler et al. 1997), which dense surface observations are able to constrain (e.g., Wheatley and Stensrud 2010; Madaus et al. 2014; Sobash and Stensrud 2015). Surface observing networks of the densities considered here would likely improve the representation of these features and, in turn, increase CI forecast skill. Given the findings here, scenarios with organized mesoscale forcing represent the most likely situations for dense surface observations to contribute to CI forecast skill.

Acknowledgments. This work was supported through NOAA CSTAR Award NA13NWS4680006. We would also like to acknowledge high-performance computing support from Yellowstone (ark:/85065/d7wd3xhc) provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation. Cliff Mass and Dale Durran provided useful discussion and suggestions, including comments on an early version of this manuscript. We also are extremely grateful to Jeff Anderson and the Data Assimilation Research Testbed team at NCAR who worked to create the interface between DART and the CM1 model used here. In addition, we thank George Bryan for making several modifications to the CM1 model to enable data assimilation and make this work possible.

REFERENCES

Anderson, J., 2001: An ensemble adjustment Kalman filter for data assimilation. *Mon. Wea. Rev.*, **129**, 2884–2903, doi: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, doi: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, doi:10.1175/MWR-D-11-00013.1.

——, B. Wyman, S. Zhang, and T. Hoar, 2005: Assimilation of surface pressure observations using an ensemble filter in an idealized global atmospheric prediction system. *J. Atmos. Sci.*, **62**, 2925–2938, doi:10.1175/JAS3510.1.

——, 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, doi:10.1175/2009BAMS2618.1.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Bryan, G. H., and J. M. Fritsch, 2002: A benchmark simulation for moist nonhydrostatic numerical models. *Mon. Wea. Rev.*, **130**, 2917–2928, doi:10.1175/1520-0493(2002)130<2917:ABSFMN>2.0.CO;2.

Burton, P., 2013: IFS documention—Cy40r1, Operational implementation 22 November 2013. Part I: Observations. Tech. Rep., European Center for Medium-Range Weather Forecasts, 76 pp. [Available online at http://www.ecmwf.int/sites/default/files/IFS_CY40R1_Part1.pdf.]

Chou, M. D., and M. J. Suarez, 1999: A solar radiation parameterization for atmospheric studies. Tech. Rep. NASA/TM-1999-10460, NASA, 38 pp.

——, and ——, 2001: A thermal infrared radiation parameterization for atmospheric studies. Tech. Rep. NASA/TM-2001-104606, NASA, 55 pp.

Dabberdt, W. F., and Coauthors, 2005: Multifunctional mesoscale observing networks. *Bull. Amer. Meteor. Soc.*, **86**, 961–982, doi:10.1175/BAMS-86-7-961.

Droegemeier, K., 1990: Toward a science of storm-scale prediction. Preprints, *16th Conf. on Severe Local Storms*, Kananaskis, Alberta, Canada, Amer. Meteor. Soc., 256–262.

Duda, J. D., and W. A. Gallus, 2013: The impact of large-scale forcing on skill of simulated convective initiation and upscale evolution with convection-allowing grid spacings in the WRF. *Wea. Forecasting*, **28**, 994–1018, doi:10.1175/WAF-D-13-00005.1.

Fowle, M. A., and P. J. Roebber, 2003: Short-range (0–48 h) numerical prediction of convective occurrence, mode, and location. *Wea. Forecasting*, **18**, 782–794, doi:10.1175/1520-0434(2003)018<0782:SHNPOC>2.0.CO;2.

Gallus, W. A., J. Correia, and I. Jankov, 2005: The 4 June 1999 derecho event: A particularly difficult challenge for numerical weather prediction. *Wea. Forecasting*, **20**, 705–728, doi:10.1175/WAF883.1.

Gaspari, G., and S. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.*, **125**, 723–757, doi:10.1002/qj.49712555417.

Hong, S., Y. Noh, and J. Dudhia, 2006: A new vertical diffusion package with an explicit treatment of entrainment processes. *Mon. Wea. Rev.*, **134**, 2318–2341, doi:10.1175/MWR3199.1.

Houze, R. A., Jr., 2014: *Cloud Dynamics.* 2nd ed. Academic Press, 496 pp.

Hu, M., H. Shao, D. Stark, and K. Newman, Eds., 2013: Gridpoint Statistical Interpolation (GSI) version 3.2 user's guide. Tech. Rep., Developmental Testbed Center, 187 pp. [Available online at http://www.dtcenter.org/com-GSI/users/docs/users_guide/GSIUserGuide_v3.2.pdf; error variances obtained from file nam_errtable.r3dv.]

Jacques, A. A., J. D. Horel, E. T. Crosman, and F. L. Vernon, 2015: Central and eastern U.S. surface pressure variations derived

from the USArray network. *Mon. Wea. Rev.*, **143**, 1472–1493, doi:10.1175/MWR-D-14-00274.1.

Jones, T. A., D. Stensrud, L. Wicker, P. Minnis, and R. Palikonda, 2015: Simultaneous radar and satellite data storm-scale assimilation using an ensemble Kalman filter approach for 24 May 2011. *Mon. Wea. Rev.*, **143**, 165–194, doi:10.1175/MWR-D-14-00180.1.

Kain, J. S., and Coauthors, 2013: A feasibility study for probabilistic convection initiation forecasts based on explicit numerical guidance. *Bull. Amer. Meteor. Soc.*, **94**, 1213–1225, doi:10.1175/BAMS-D-11-00264.1.

Lilly, D. K., 1990: Numerical prediction of thunderstorms—Has its time come? *Quart. J. Roy. Meteor. Soc.*, **116**, 779–798, doi:10.1002/qj.49711649402.

Lin, Y.-L., R. D. Farley, and H. D. Orville, 1983: Bulk parameterization of the snow field in a cloud model. *J. Climate Appl. Meteor.*, **22**, 1065–1092, doi:10.1175/1520-0450(1983)022<1065:BPOTSF>2.0.CO;2.

Madaus, L. E., and G. J. Hakim, 2016: Observable surface anomalies preceding simulated isolated convective initiation. *Mon. Wea. Rev.*, **144**, 2265–2284, doi:10.1175/MWR-D-15-0332.1.

——, ——, and C. F. Mass, 2014: Utility of dense pressure observations for improving mesoscale analyses and forecasts. *Mon. Wea. Rev.*, **142**, 2398–2413, doi:10.1175/MWR-D-13-00269.1.

Markowski, P., C. Hannon, and E. Rasmussen, 2006: Observations of convection initiation failure from the 12 June 2002 IHOP deployment. *Mon. Wea. Rev.*, **134**, 375–405, doi:10.1175/MWR3059.1.

Mass, C. F., and L. E. Madaus, 2014: Surface pressure observations from smartphones: A potential revolution for high-resolution weather prediction? *Bull. Amer. Meteor. Soc.*, **95**, 1343–1349, doi:10.1175/BAMS-D-13-00188.1.

——, D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.*, **83**, 407–430, doi:10.1175/1520-0477(2002)083<0407:DIHRPM>2.3.CO;2.

Mecikalski, J. R., P. Minnis, and R. Palikonda, 2013: Use of satellite derived cloud properties to quantify growing cumulus beneath cirrus clouds. *Atmos. Res.*, **120–121**, 192–201, doi:10.1016/j.atmosres.2012.08.017.

Muller, C. L., L. Chapman, S. Johnston, C. Kidd, S. Illingworth, G. Foody, A. Overeem, and R. R. Leigh, 2015: Crowdsourcing for climate and atmospheric sciences: Current status and future potential. *Int. J. Climatol.*, **35**, 3185–3203, doi:10.1002/joc.4210.

Owen, J., 1966: A study of thunderstorm formation along dry lines. *J. Appl. Meteor.*, **5**, 58–63, doi:10.1175/1520-0450(1966)005<0058:ASOTFA>2.0.CO;2.

Roebber, P. J., D. M. Schultz, and R. Romero, 2002: Synoptic regulation of the 3 May 1999 tornado outbreak. *Wea. Forecasting*, **17**, 399–429, doi:10.1175/1520-0434(2002)017<0399:SROTMT>2.0.CO;2.

Sieglaff, J. M., L. M. Cronce, W. F. Feltz, K. M. Bedka, M. J. Pavolonis, and A. K. Heidinger, 2011: Nowcasting convective storm initiation using satellite-based box-averaged cloud-top cooling and cloud-type trends. *J. Appl. Meteor. Climatol.*, **50**, 110–126, doi:10.1175/2010JAMC2496.1.

Snook, N., M. Xue, and Y. Jung, 2015: Multiscale EnKF assimilation of radar and conventional observations and ensemble forecasting for a tornadic mesoscale convective system. *Mon. Wea. Rev.*, **143**, 1035–1057, doi:10.1175/MWR-D-13-00262.1.

Sobash, R. A., and D. J. Stensrud, 2015: Assimilating surface mesonet observations with the EnKF to improve ensemble forecasts of convection initiation on 29 May 2012. *Mon. Wea. Rev.*, **143**, 3700–3725, doi:10.1175/MWR-D-14-00126.1.

Stensrud, D. J., and J. M. Fritsch, 1994: Mesoscale convective systems in weakly forced large-scale environments. Part III: Numerical simulations and implications for operational forecasting. *Mon. Wea. Rev.*, **122**, 2084–2104, doi:10.1175/1520-0493(1994)122<2084:MCSIWF>2.0.CO;2.

——, and Coauthors, 2009: Convective-scale warn-on-forecast system: A vision for 2020. *Bull. Amer. Meteor. Soc.*, **90**, 1487–1499, doi:10.1175/2009BAMS2795.1.

Stoelinga, M. T., 2005: Simulated equivalent reflectivity factor as currently formulated in RIP: Description and possible improvements. The Pennsylvania State University, 5 pp. [Available online at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.522.925&rep=rep1&type=pdf.]

Sun, J., and Coauthors, 2014: Use of NWP for nowcasting convective precipitation: Recent progress and challenges. *Bull. Amer. Meteor. Soc.*, **95**, 409–426, doi:10.1175/BAMS-D-11-00263.1.

Tao, W.-K., and J. Simpson, 1993: Goddard Cumulus Ensemble Model. Part I: Model description. *Terr. Atmos. Oceanic Sci.*, **4**, 35–71, doi:10.3319/TAO.1993.4.1.35(A).

Weckwerth, T. M., 2000: The effect of small-scale moisture variability on thunderstorm initiation. *Mon. Wea. Rev.*, **128**, 4017–4030, doi:10.1175/1520-0493(2000)129<4017:TEOSSM>2.0.CO;2.

——, and D. B. Parsons, 2006: A review of convection initiation and motivation for IHOP 2002. *Mon. Wea. Rev.*, **134**, 5–22, doi:10.1175/MWR3067.1.

Wheatley, D. M., and D. J. Stensrud, 2010: The impact of assimilating surface pressure observations on severe weather events in a WRF mesoscale ensemble system. *Mon. Wea. Rev.*, **138**, 1673–1694, doi:10.1175/2009MWR3042.1.

Zhang, F., C. Snyder, and R. Rotunno, 2003: Effects of moist convection on mesoscale predictability. *J. Atmos. Sci.*, **60**, 1173–1185, doi:10.1175/1520-0469(2003)060<1173:EOMCOM>2.0.CO;2.

Ziegler, C. L., T. J. Lee, and R. A. Pielke, 1997: Convective initiation at the dryline: A modeling study. *Mon. Wea. Rev.*, **125**, 1001–1026, doi:10.1175/1520-0493(1997)125<1001:CIATDA>2.0.CO;2.