Testing the Feature Alignment Technique (FAT) in an Ensemble-Based Data Assimilation and Forecast System with Multiple-Storm Scenarios

Derek R. Stratman[a,b] and Corey K. Potvin[b,c]

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

ABSTRACT: Storm displacement errors can arise from a number of potential sources of error within a data assimilation (DA) and forecast system. Conversely, storm displacement errors can cause issues for storm-scale, ensemble-based systems using an ensemble Kalman filter (EnKF), such as NSSL's Warn-on-Forecast System (WoFS). A previous study developed a fully grid-based feature alignment technique (FAT) to mitigate these phase errors and their impacts. However, that study developed and tested the FAT for single-storm cases. This study advances that work by implementing an object-based merging and matching technique into the FAT and tests the updated FAT in more complex scenarios of multiple storms. Ensemble-based experiments are conducted with and without the FAT for each of the scenarios. The experiments' analyses and forecasts of storm-related fields are then evaluated using subjective and objective methods. Results from these idealized multiple-storm experiments continue to reveal the potential benefits of correcting storm displacement errors. For example, running the FAT even once can mitigate the "spinup" period experienced by the no-FAT experiments. The new results also show that running the FAT prior to *every* DA cycling step generally leads to more skillful forecasts at the smaller scales, especially in earlier-initialized forecasts. However, repeatedly running the FAT prior to every DA step can eventually lead to deterioration in analyses and forecasts. Potential solutions to this problem include using longer cycling intervals and running the FAT prior to DA less often. Additional ways to improve the FAT along with other results are presented and discussed.

SIGNIFICANCE STATEMENT: The purpose of this work is to explore the impact of correcting storm displacements on analyses and forecasts of storms using an ensemble-based data assimilation and forecast system in an idealized framework. Storm displacement errors are a common problem in current operational and experimental storm-scale forecast systems, so understanding their impact on these systems and providing a method to help mitigate them is important. Results from this study indicate that correcting storm displacement errors with the feature alignment technique can greatly improve analyses and forecasts in multiple-storm scenarios. Future work will focus on exploring the impact of correcting storm displacement errors in a real-data, storm-scale data assimilation and forecast system.

KEYWORDS: Kalman filters; Ensembles; Forecast verification/skill; Short-range prediction; Data assimilation; Idealized models; Numerical weather prediction/forecasting

---

1. Introduction

A goal of storm-scale, ensemble-based data assimilation (DA) and forecast systems, such as the National Severe Storms Laboratory's Warn-on-Forecast System (WoFS; Stensrud et al. 2009, 2013; Skinner et al. 2018; Jones et al. 2020), is to provide probabilistic location and timing guidance about severe storms and their associated hazards. The experimental WoFS has had many successes at providing accurate and reliable probabilistic forecasts of severe convective weather (e.g., Skinner et al. 2018; Jones et al. 2019; Flora et al. 2019; Yussouf et al. 2020). However, forecast storm motion biases are still a common issue for the WoFS (Wheatley et al. 2015; Skinner et al. 2018; Flora et al. 2019). These storm motion biases can develop due to numerous sources of errors within the data assimilation and model systems (Houtekamer and Zhang 2016). For example, storm motion can develop due to coarse model grid

spacing (e.g., Potvin and Flora 2015) and parameterized model physics (e.g., Stratman and Brewster 2017; Potvin et al. 2020). Conversely, the storm motion biases can lead to storm displacement errors, which can cause issues for DA-forecast systems.

A common DA method used for storm-scale, ensemble-based forecast systems, including the WoFS, is the ensemble Kalman filter (EnKF; Evensen 1994). EnKF performs best when the model is linear and the prior distribution is Gaussian. While neither is true in systems like the WoFS, storm displacement errors still add another layer of nonlinearity/non-Gaussianity, which can prevent EnKF from performing more optimally. Stratman et al. (2018, hereafter S18) showed storm displacement errors can lead to model imbalances and subsequent issues with forecast storm development and propagation. To minimize these errors, S18 developed a grid-based feature alignment technique (FAT) for storm-scale, limited-size domains based on similar techniques developed by Nehrkorn et al. (2014) and Nehrkorn et al. (2015) to correct for the storm displacement errors. The FAT solves for a 2D field of displacement vectors by

Corresponding author: Derek R. Stratman, derek.stratman@noaa.gov

variationally minimizing the difference between observations (e.g., composite reflectivity) and a background field. The displacement vectors are then used to update the model state variables by adjusting the fields at all model levels. Because the FAT is run prior to DA, the observations are more optimally assimilated by the EnKF due to reducing the spatial errors.

This study expands on S18, which only focused on experiments with correcting the location of a single storm, by modifying the FAT to handle multiple storms. S18 used a first-guess field of displacement vectors to reduce the influence of non-global minima (Brewster 2003). However, to form the first guess field, S18 computed the centers of mass for thresholded observation and forecast fields valid over the entire domain since only one object existed in each field. While this method works well in single-storm scenarios, it can fail when multiple storms are present. Thus, an object-based merging and matching method is developed to account for the presence of multiple storms—and potentially differing numbers of observed and forecast storms—when forming the first-guess field of displacement vectors.

The new FAT is tested in various multiple-storm scenarios in an idealized framework using observing system simulation experiments (OSSEs). An overview of the FAT along with the modifications to the process can be found in section 2. Details about the design of the OSSEs are described in section 3. Results from the experiments are presented in section 4. Finally, a summary and discussion of the results and potential future work can be found in section 5.

2. Updated FAT

a. Overview of the FAT

As described in S18, the FAT solves for a 2D field of displacement vectors by minimizing a cost function between a background forecast and observations. The minimization is performed using a nonlinear conjugate gradient method based on Polak and Ribière (1969). The cost function,

$$J = J_r + \lambda_s J_s + \lambda_d J_d + \lambda_m J_m + \lambda_b J_b, \qquad (1)$$

is composed of a residual error term J_r and four constraint terms. The residual error function,

$$J_r = \sum \frac{\left[y(i,j) - x(i + a, j + b)\right]^2}{\sigma_o^2}, \qquad (2)$$

is the summation of squared differences between the background forecasts x and observations y, weighted by the observational error variance σ_o^2. The function is minimized by solving for the i- and j-direction displacement components a and b, respectively. The smoothness constraint,

$$J_s = \sum \left(\frac{\partial^2 a}{\partial i^2} + \frac{\partial^2 a}{\partial j^2}\right)^2 + \sum \left(\frac{\partial^2 b}{\partial i^2} + \frac{\partial^2 b}{\partial j^2}\right)^2, \qquad (3)$$

smooths out the gradient of displacement components. The divergence constraint,

$$J_d = \sum \left(\frac{\partial a}{\partial i} + \frac{\partial b}{\partial j}\right)^2, \qquad (4)$$

prevents the convergence and divergence of displacement components. The magnitude constraint,

$$J_m = \sum \left(\frac{a}{S}\right)^2 + \sum \left(\frac{b}{S}\right)^2, \qquad (5)$$

increasingly penalizes displacement components as $|a|$ and $|b|$ approach the displacement magnitude limit S and acts to suppress displacement components in data-void areas. The barrier constraint,

$$J_b = \sum \left(\frac{a}{S}\right)^{20} + \sum \left(\frac{b}{S}\right)^{20}, \qquad (6)$$

severely penalizes displacement components with $|a|$ and $|b|$ greater than S. In this study, the barrier constraint does not have an impact on the displacement components due to the maximum $|a|$ and $|b|$ being much smaller than S, but this constraint term is retained since it could be useful in future real-data applications. The constraint functions are weighted with coefficients λ_s, λ_d, λ_m, and λ_b, respectively. As in S18, the 2D field of displacement vectors are computed using forecast and observed column-maximum reflectivity, which have been placed on the same 2D grid of i, j indices.

As in Nehrkorn et al. (2014) and S18, the FAT runs as a standalone program prior to data assimilation for each ensemble member and is composed of the following general steps:

1) Read in background forecast and observed column-maximum reflectivity fields.
2) Form a first-guess field of displacement vectors using an object-based merging and matching method.
3) Run the nonlinear conjugate gradient minimization function to solve for the final 2D field of displacement vectors.
4) Update all of the model state variables using the 2D field of displacement vectors and output to new background forecast files.

The object-based merging and matching technique will be described in the following section. The nonlinear conjugate gradient minimization is performed by the "fmin_cg" function in SciPy's "optimize" module. The final 2D fields of a and b are added to the original i, j grid indices, respectively, to yield an adjusted field of grid indices. The "RegularGrid Interpolator" function in SciPy's "interpolate" module is then used to interpolate the model state variables at all vertical levels from the original grid indices to the new adjusted grid indices.

b. FAT modifications

A majority of the changes to the FAT are related to the creation of the first-guess field of displacement vectors. As in S18, the updated FAT initializes the 2D grid of displacement vectors with small random values with a distribution centered on zero. However, the rest of the first guess generation process has been modified from S18 to account for multiple
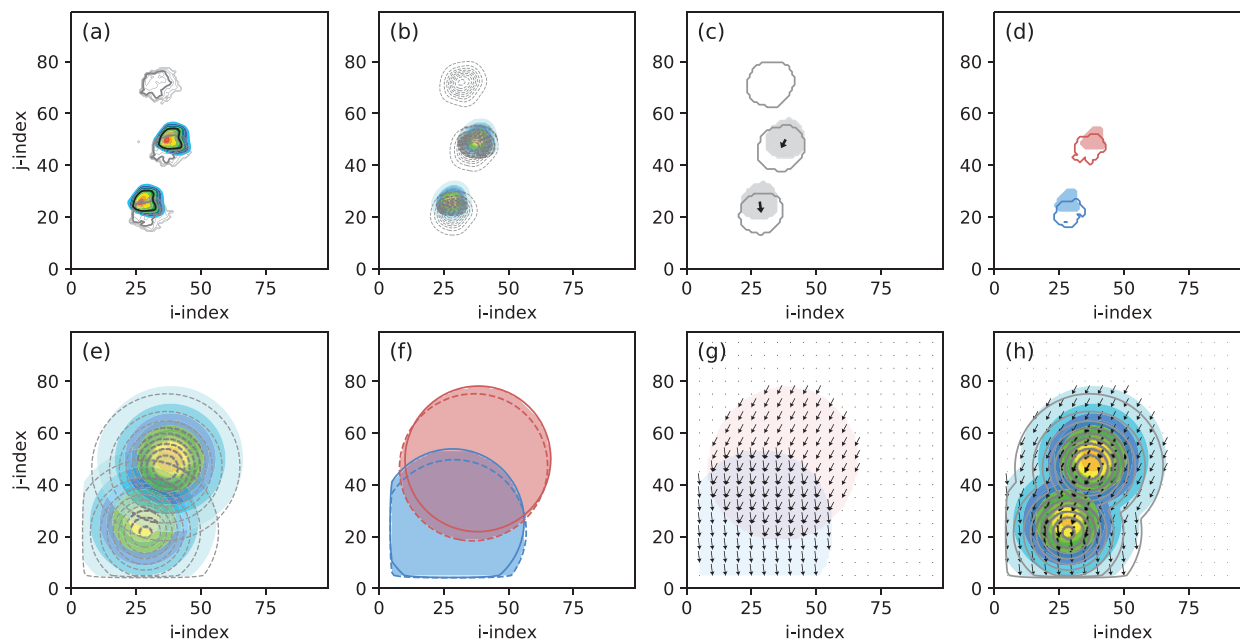
FIG. 1. Schematic showing the process of creating the first-guess field of displacement vectors. (a) Original forecast (color shading) and observed (gray contours) reflectivity fields. Thicker black and gray contours highlight the var_{min} threshold for both fields, respectively. (b) Smoothed forecast (color shading) and observed (gray dashed contours) reflectivity fields using a Gaussian filter with the standard deviation σ_1. (c) Forecast (gray shading) and observed (gray contours) binary fields, which are created from thresholding the smoothed fields using $smth_{min}$. Black arrows represent the displacements between the forecast and observed centroids. (d) Remaining forecast–observed object pairs. (e) Smoothed forecast (color shading) and observed (gray dashed contours) reflectivity fields using a Gaussian filter with the standard deviation σ_2. (f) Forecast (solid contours) and observed (dashed contours) binary fields, which are created from thresholding the smoothed fields using $smth_{min}$. The convex hull area is shaded for each of the matched-object pairs. (g) Displacement vectors and convex hull areas (color shading) for each of the matched-object pairs. (h) Final field of first-guess displacement vectors and the smoothed forecast (color shading) and observed (gray contours) reflectivity fields.

observed and forecast storms and improve the final displacement vector field. S18 thresholded the original observed and forecast fields and found their centers of mass. The observed and forecast fields were then smoothed using a Gaussian smoother, and the vector between the centers of mass was used to fill the area covered by the smoothed fields. This method works well for a single pair of observed and forecast storms, but is inappropriate when multiple storms exist in the observed and/or forecast fields. For example, given a scenario in which there are three storms in the observed field but only one storm in the forecast field, the three storms in the observed field would be considered one object. Thus, the old FAT would have unrealistically stretched the forecast storm to fill the combined areas of the observed storms.

To account for multiple-storm scenarios, we develop an object-based merging and matching technique—similar to those used in object-based verification methods (e.g., Skinner et al. 2018)—to form matched observed-forecast storm pairs, then compute the initial displacement vectors for each pair. In the first step of the object-based merging and matching procedure, the observed and forecast fields (Fig. 1a) are masked using a threshold, var_{min}, and smoothed using a Gaussian smoother with a standard deviation σ_1 (Fig. 1b). The smoothed observed and forecast fields are then normalized using the maximum value in the original observed field,

as in S18. This first smoothing step merges storms that are close together into a single object for the object matching step, which is next.

Observed and forecast objects are identified and matched from the thresholded smoothed fields as follows. First, binary and thresholded fields are created from the smoothed observed and forecast fields using a different masking threshold, $smth_{min}$ (Fig. 1c). The number of distinct binary objects and their labels are determined using the "label" function in scikit-image's "measure" module. The field of labels for the observed and forecast binary fields along with their thresholded smoothed fields are then input to the measure module's "regionprops" function to determine a plethora of object attributes, such as area and centroid location. Next, observed and forecast objects are matched together. Only objects with areas greater than a threshold, $area_{min}$, are considered for matching to avoid matching small, possibly spurious storms. Remaining observed and forecast objects with the smallest distances between their centroid locations are matched together if their distance is less than the maximum allowable distance, $dist_{max}$.

Now that we have our matched objects, the following process is carried out for each pair of matched objects to determine their respective displacement vector fields and their weights, which are used when the individual fields of

displacement vectors are merged together. First, objects in the original observed and forecast fields other than the matched object pair are masked out using the object labels generated during the object identification (Fig. 1d). Then, the masked fields are thresholded using var_{min} and smoothed using a Gaussian smoother with a larger standard deviation σ_2, than used in the previous smoothing step (Fig. 1e). The new smoothed observed and forecast fields are next input into the object identification and matching routine to determine the new object centroid locations and object label fields. Observed and forecast binary fields are created from the smoothed fields using the threshold $smth_{min}$, and then merged together using the object label fields as masks. Next, the convex hull field of the merged binary field is determined using the "convex_hull_image" function in scikit-image's "morphology" module (Fig. 1f). The displacement vector from the forecast centroid to the observed centroid is then applied to the grid points within the convex hull area (Fig. 1g).

Finally, the first-guess fields of displacement vectors for each pair of matched objects are blended to form the final first-guess field (Fig. 1h). In areas where displacement vectors exist for only one pair of matched objects, those displacement vectors are assigned to the final first-guess field. In areas where displacement vectors were computed for multiple observed–forecast object pairs, the displacement vector fields are averaged together using a Gaussian weighting function, which gives more weight to the displacement vectors closest to their associated forecast object's centroid. The individual smoothed object fields are merged together to form the final observed and forecast fields for use in the cost function minimization (Fig. 1h). The initial small random displacement vectors are maintained for grid points outside of any pair of matched objects' convex hull area.

S18 found that displacement vectors undesirably extended beyond the edge of the domain. They corrected the issue by reducing the magnitudes of the vectors that extended beyond the edge of the domain. However, we subsequently attributed the undesirable extensions of displacement vectors outside the domain to the cost function excluding vectors less than three grid points from the domain edge, which in turn arose from the finite differencing stencil used. To mitigate this issue in the updated FAT, the domain over which the cost function is minimized is expanded based on the gridpoint thinning factor. For example, if the original domain is thinned by a factor of 2 for the minimization as in S18, the domain used in the cost function minimization would need to be expanded by 4 grid points on all sides—i.e., a 100 × 100 gridpoint domain would become a 108 × 108 gridpoint domain prior to thinning the domain by a factor of 2 so that the cost function includes displacement vectors out to the boundaries of the original domain. Using the expanded cost function minimization domain smooths the gradients of the displacement vectors near the domain edges, thereby helping prevent displacement vectors from extending beyond the domain edges.

In addition to the changes made to the formation of the first-guess field of displacement vectors, a couple of checks were added to prevent the FAT from applying unwanted updates to the model state variables. First, the FAT will not update the model state variables if the maximum vector in the final field of displacement vectors is smaller than the minimum vector threshold, $vect_{min}$. If set, this threshold can prevent unnecessary and possibly problematic minor adjustments to the model state variables. For another check, the model state variables are not updated if the root-mean-squared error (RMSE) between the observed and forecast fields increases after applying the displacement vectors. This occurrence is unlikely, but if there are, for example, 2 matched forecast storms and 1 unmatched forecast storm, the final displacement vectors from the matched storms could erroneously alter the unmatched storm. In this scenario, the RMSE could undesirably increase.

3. Experiment design

a. Truth run(s)

Prior to running OSSEs, high-resolution simulations are performed to provide synthetic observations for data assimilation and verification fields for the analyses and forecasts. These Truth simulations are generated using version 3.9.1.1 of the Advanced Research version of the Weather Research and Forecasting (WRF) Model (WRF-ARW; Skamarock et al. 2008) in an idealized setting on a 1201 × 1201 × 67 gridpoint domain with a horizontal grid spacing of 250 m. The vertical grid spacing is 100 m below 1 km and is stretched from 100 m at 1 km AGL to 800 m at the model top of ~25 km. The simulations are initialized from a homogenous environment provided by the same 24 May 2011 "El Reno tornado" thermodynamic profile used in S18 (Fig. 2a), but the default wind profile in this study has been modified from S18 to yield more easterly storm motions for all experiment scenarios, which will be described in section 3d (Figs. 2b–d). To generate multiple storms within each simulation, 5-K warm bubbles with horizontal radii of 10 km and vertical radii of 1.5 km are added to the initial fields at 1.5 km AGL. The Truth runs are performed for 3 h of simulation time using a 1-s time step. The NSSL three-moment microphysics scheme (Mansell et al. 2020) with a base cloud condensation nuclei concentration of 1000 cm^{-3} is used to provide more realistic reflectivity observations.

Synthetic reflectivity and radial velocity observations are created from the Truth runs using observation converter programs in National Center for Atmospheric Research's (NCAR) Manhattan release of the Data Assimilation Research Testbed (DART; Anderson et al. 2009). The reflectivity and radial velocity observations are interpolated onto a "grid-tilt" domain defined by a horizontal grid spacing of 3 km, which is the same grid spacing used for the OSSEs, and 14 elevation angles from 0.5° to 19.5°, which radiate from a simulated radar located at the center of the domain. Observations are only generated below 10 km AGL and within 150-km horizontal distance of the simulated radar. Areas where reflectivity < 10 dBZ—referred to herein as clear-air reflectivity—are set to 0 dBZ and thinned to every third grid point. Also, radial velocity observations are omitted where reflectivity < 10 dBZ. The remaining reflectivity and radial velocity observations are included in DART-formatted observation sequence files for assimilation. Drawing from
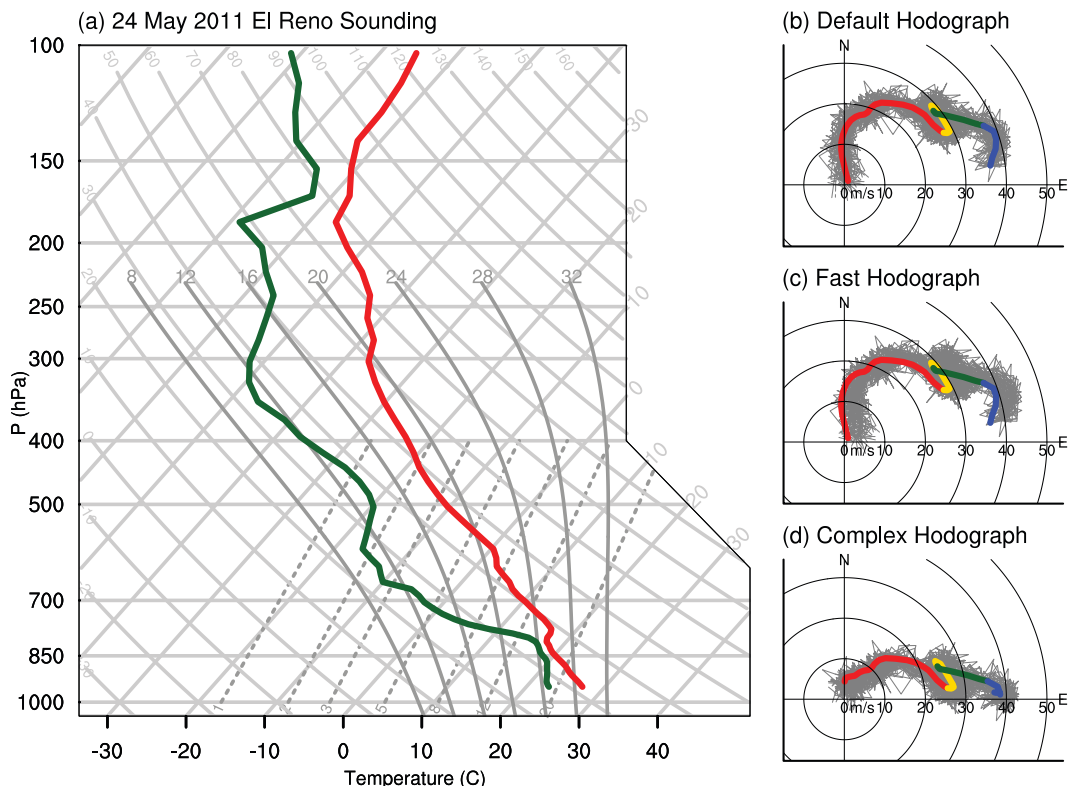
(a) 24 May 2011 El Reno Sounding

(b) Default Hodograph

(c) Fast Hodograph

(d) Complex Hodograph

FIG. 2. (a) Skew T–log$p$ of the temperature and moisture profiles and plots of the (b) default, (c) fast, and (d) complex hodographs used to provide the initial conditions for the Truth and ensemble simulations. In the hodographs, the thick colored lines depict the Truth's wind profiles for the layers 0–3000 (red), 3000–6000 (yellow), 6000–9000 (green), and 9000–12 000 m (blue) AGL, and the ensemble member's initial wind profiles are represented by the thin gray lines.

Gaussian distributions with zero means, observation error standard deviations of 2 dBZ for reflectivity and 2 m s^{-1} for radial velocity (Yussouf and Stensrud 2010; Sobash and Stensrud 2013; Kerr and Wang 2020) are added to the synthetic observations and are assumed in the data assimilation.

b. Data assimilation and forecast system

Experiments with and without the FAT are conducted with a 50-member ensemble data assimilation and forecast system using DART's ensemble adjustment Kalman filter (EAKF; Anderson 2001) and WRF-ARW (version 3.9.1.1) for the initial conditions and forecasts. The OSSEs are completed on a $101 \times 101 \times 67$ gridpoint domain with a horizontal grid spacing of 3 km. Similar to the Truth simulations, 5-K warm bubbles with horizontal radii of 15 km and vertical radii of 2.5 km are added to the ensemble member's initial model fields 2.5 km AGL to generate storms. To create ensemble diversity, each storm's individual ensemble members' warm bubbles are randomly displaced up to 15 km in the i and j directions away from the Truth's warm bubbles following a discrete uniform distribution. To further diversify the ensemble, vertically uncorrelated Gaussian perturbations with a standard deviation of 2 m s^{-1} are added to the u and v wind profiles up to ~11 km AGL (Figs. 2b–d), which is similar to how other OSSE studies (e.g., Kerr and Wang 2020) diversified their ensemble members. The ensemble forecasts are then integrated forward from the

initial conditions until $t = 45$ min to give time for the warm bubbles to develop into storms. To emulate the current experimental WoFS, the 3-km ensemble forecasts use the NSSL two-moment microphysics scheme (Mansell et al. 2010) with a base cloud condensation nuclei concentration of 1250 cm^{-3} and a time step of 10 s. All other WRF-ARW namelist settings are the same as the Truth's namelist settings.

Using DART's EAKF, the Truth's synthetic reflectivity and radial velocity observations are assimilated every 5 min from $t = 45$ to 120 min. An adaptive prior covariance inflation scheme (Anderson 2009) is used during DA cycling to help maintain sufficient ensemble spread and prevent ensemble collapse. The horizontal localization radius for experiments using the FAT is 6 km, but for experiments *not* using the FAT, the horizontal localization radius is 12 km. These horizontal localization radii produce the best results for each set of experiments. All experiments and observation types use a 3-km vertical localization radius. The Gaspari–Cohn function (Gaspari and Cohn 1999) is used for the localization weighting function. To see how the impact of the FAT changes with subsequent DA cycles, ensemble forecasts are initialized every 15 min from $t = 45$ min until $t = 120$ min and are run until $t = 180$ min.

c. FAT configuration

As mentioned before, the FAT is run prior to each DA step. The same FAT configuration is used for all FAT

TABLE 1. FAT tunable parameter labels, values, and descriptions; "gp" is the abbreviation of grid point(s).

Parameter	Value	Description
σ_o	2 dBZ	Observational error standard deviation
S	50 gp	Displacement magnitude limit
λ_s	6.0	Smoothness penalty function weight
λ_d	3.0	Divergence penalty function weight
λ_m	0.1	Magnitude penalty function weight
λ_b	1.0	Barrier penalty function weight
σ_1	0.5 gp	Object-merging step's Gaussian smoother standard deviation
σ_2	10.0 gp	Gaussian standard deviation used to form smoothed fields for first-guess field of displacement vectors
var_{min}	25 dBZ	Threshold used on original forecast and observed fields
$smth_{min}$	1 dBZ	Threshold used on the smoothed forecast and observed fields
$dist_{max}$	30 gp	Maximum allowable distance between object centroids
$area_{min}$	24 gp	Minimum area threshold for objects
$vect_{min}$	0.05 gp	State variables are updated only if the maximum final displacement vector is greater than this threshold

experiments. Just as in S18, the FAT parameters were empirically tuned through brute force to determine the parameter configuration that 1) produced some of the smallest RMSEs for column-maximum reflectivity for a variety of individual ensemble members' forecasts in different scenarios without substantially altering forecast storm sizes and 2) generally resulted in the best analyses and forecasts. Relatively small adjustments to these parameters do not substantially change the results. For example, using $\sigma_2 = 11.0$ produces similar results as with using $\sigma_2 = 10.0$, but using a much larger σ_2, such as 20.0, results in substantially different results (not shown). Parameter values and their descriptions used in the FAT process are listed in Table 1. Also, we thin the forecast and observed grids by using every fifth grid point when performing the cost function minimization to decrease the computational time needed to run the FAT. This reduction in the number of grid points does not substantially affect the quality of the displacement vectors. This quality assessment is based on comparing the RMSEs between the observed and adjusted forecast fields of column-maximum reflectivity for different thinning factors. The thinned field of displacement vectors are interpolated back to the full grid before adjusting the model state variables. By using the thinned grids, the total time of the entire FAT process, including updating and writing out the new state variables, is relatively small, only taking ~1 min for the entire ensemble on a modern HPC system since each member can be processed in parallel.

d. Experiments

Six multiple-storm scenarios are designed to explore the impact of using the FAT to correct various sources of storm displacement errors (Table 2). For each scenario, a set of five experiments are completed (Table 3). First, a no-data-assimilation ensemble of simulations (NoDA) is integrated out to $t = 180$ min to provide the background forecasts at $t = 45$ min for the first analysis cycle of the DA experiments and to provide a baseline for forecast assessment where data assimilation is not used to correct the background forecast. To directly assess the impact of the FAT in each scenario, DA experiments are performed without the FAT (NoFAT). In S18, various configurations of the FAT were tested, but for this study, the same configuration is used for all FAT experiments. Instead, three FAT experiments are run for each scenario to explore how the impact of the FAT evolves with subsequent DA cycles. The three FAT experiments include running the FAT for the first DA time only (1FAT), the first five DA times only (5FAT), and all DA times (AllFAT).

The first scenario tests the FAT in a situation where there are three Truth storms and three ensemble storms with no initial displacement of the ensemble mean warm bubbles' locations from the Truth's warm bubbles' locations (NoDisp; Fig. 3a). For those experiments, the only storm displacements come from the spread in the initial ensemble members' warm bubble placements and from any displacements that develop due to the noise in the wind profiles (Fig. 2b). The next scenario builds on the NoDisp experiments by adding an initial 30-km eastward displacement to the warm bubble locations (Disp; Fig. 3b). The Disp experiments are meant to simulate cases where large phase errors exist in the initial conditions. Another scenario uses a modified wind profile (Fig. 2c) that has a 2.78 m s^{-1} storm motion bias (Fast; Fig. 3c), which results in a storm displacement error of ~30 km after 3 h of simulation. Interestingly, the

TABLE 2. Scenario labels and their descriptions.

Scenarios	Description
NoDisp	No initial displacement of the mean warm bubble locations with 5-min cycling interval
Disp	Initial 30-km eastward displacement of the mean warm bubble locations with 5-min cycling interval
Fast	As in NoDisp, but with a fast storm motion bias
Cyc15	As in NoDisp, but with 15-min cycling interval
Miss	As in NoDisp, but with 30% of initial warm bubbles randomly removed
Complex	Combination of scenarios with 5-min cycling interval

TABLE 3. Experiment labels and their descriptions.

Experiments	Description
Truth	250-m truth simulations
NoDA	3-km ensemble with no data assimilation
NoFAT	3-km ensemble with only EAKF DA
1FAT	3-km ensemble with EAKF DA and the FAT run prior to the first DA time
5FAT	3-km ensemble with EAKF DA and the FAT run prior to the first five DA times
AllFAT	3-km ensemble with EAKF DA and the FAT run prior to all DA times

displacement is minimal at $t = 45$ min similar to NoDisp likely due to any eastward displacement being offset by, e.g., the microphysics and grid-spacing differences. The Fast experiments are meant to mimic storm displacement errors resulting from biases in the forecast model.

Using the same experiment design as NoDisp (Fig. 3a), we perform another set of experiments using a cycling frequency of 15 min (Cyc15) instead of 5 min based on results from Ying (2019), which showed alignment correction techniques might work better at certain cycling frequencies. In another scenario, 30% of the warm bubbles are randomly removed from NoDisp's initial ensemble members' fields to test the ability of the FAT to handle an unequal number of Truth and ensemble members' storms (Miss; Fig. 3d). By design, ensemble members can have 0–3 warm bubbles remaining. Even so, only two ensemble members have no initial warm bubbles.

The final scenario tests the FAT in a more complex storm setting with initially four Truth storms and three background storms, a modified vertical wind profile (Fig. 2d), a ~1.4 m s^{-1} storm motion bias to the southeast, 30% of warm bubbles randomly removed, and an initial 30-km westward displacement to the warm bubble locations for the northernmost storm (Complex; Fig. 3e). Even though the Truth and ensemble contain primarily supercells, this scenario is more complex than in the previous scenarios due to the increased storm interactions, especially among the northern three Truth storms.

e. Evaluation methods

A mix of subjective and objective evaluation methods are used to assess and compare the different experiments' analyses and forecasts. To assess the filter performance with and
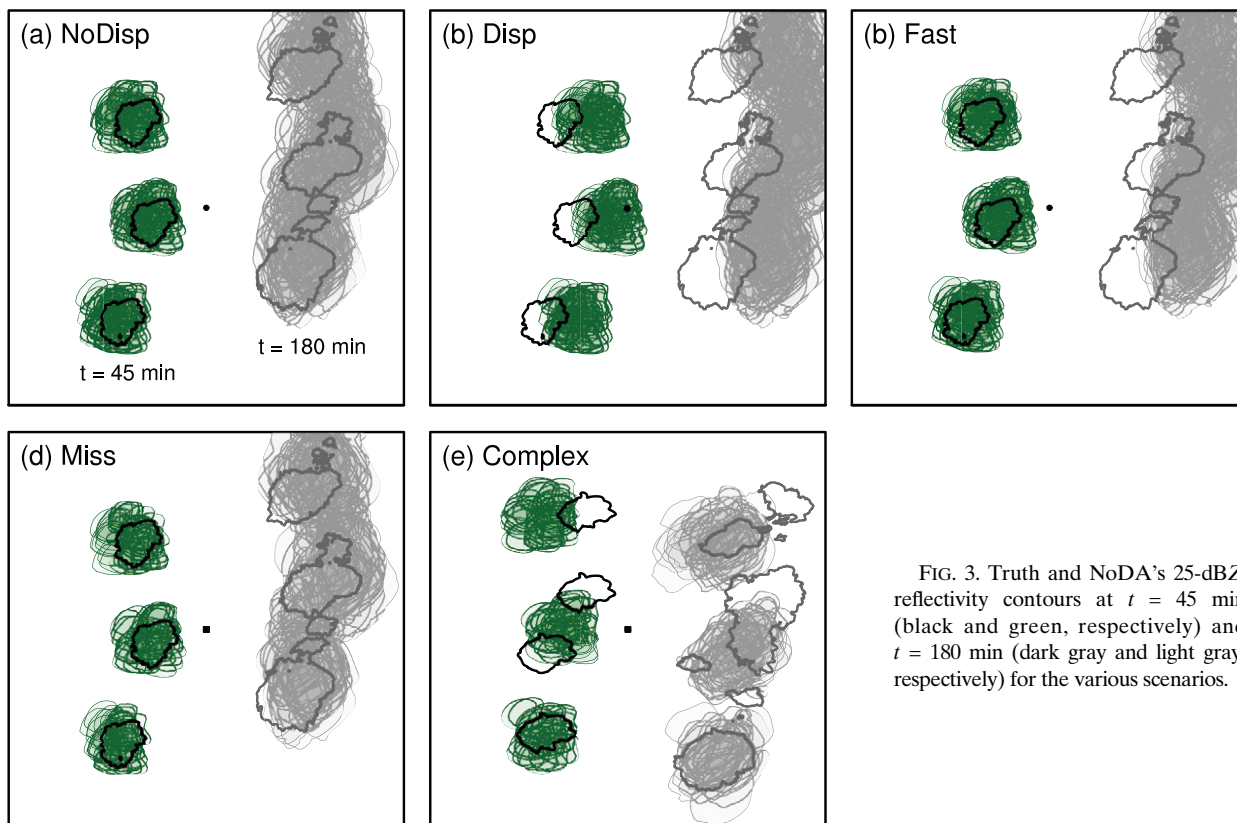

FIG. 3. Truth and NoDA's 25-dBZ reflectivity contours at $t = 45$ min (black and green, respectively) and $t = 180$ min (dark gray and light gray, respectively) for the various scenarios.
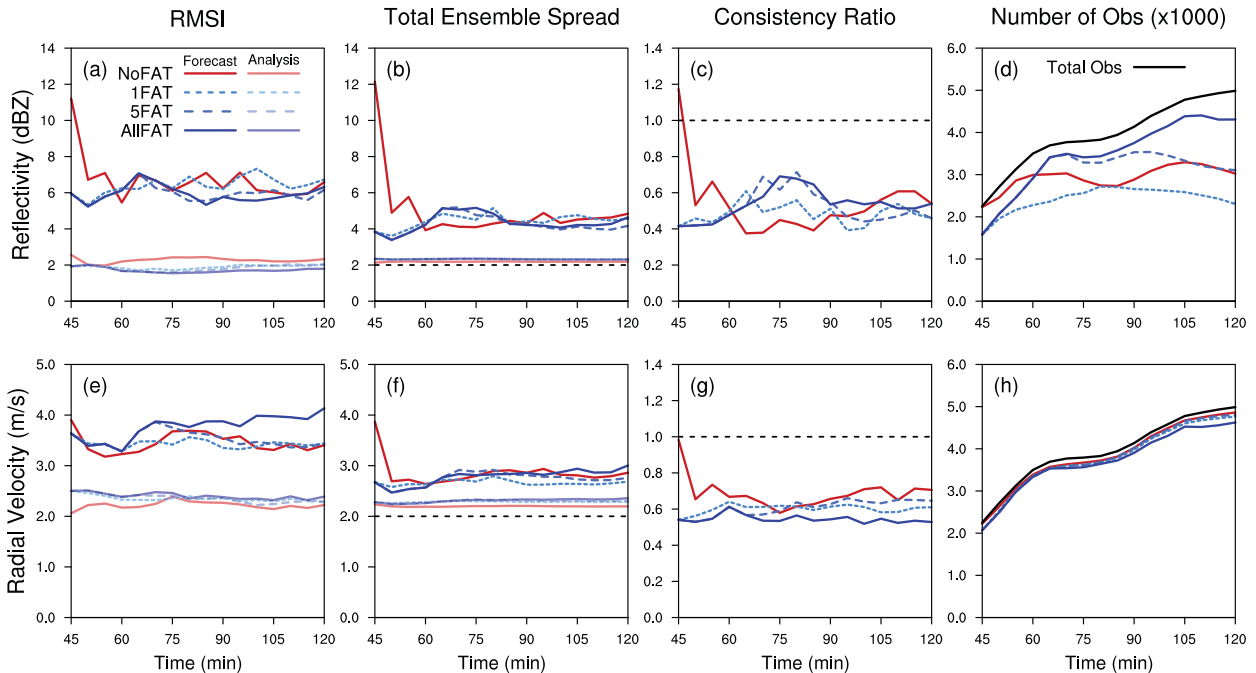
FIG. 4. Observation-space diagnostics and the number of observations assimilated for (a)–(d) reflectivity and (e)–(h) radial velocity are shown. In (b) and (f), the black dashed line represents the observation error. In (c) and (g), the black dashed line indicates an ideal consistency ratio of 1.

without the FAT, observation-space diagnostics, including root-mean-square innovation (RMSI), total ensemble spread (TES), and consistency ratio (Dowell and Wicker 2009; Dowell et al. 2011), and the number of observations assimilated are computed for reflectivity and radial velocity in areas of observed reflectivity > 10 dBZ. The equation for consistency ratio is

$$CR = \left(\frac{TES}{RMSI}\right)^2. \qquad (7)$$

Ideally, RMSI and total ensemble spread, which includes the observation error, are quantitatively similar, which would result in optimal consistency ratios near 1.

Two storm-based attributes—2–5-km updraft helicity (UH; Kain et al. 2008) and column-maximum reflectivity—are used to assess the experiments' forecast performance. First, the neighborhood maximum ensemble probabilities (NMEP; Schwartz and Sobash 2017) for UH greater than 60 m^2 s^{-2} are computed from the forecast initialization times until $t = 180$ min by using a 9-km neighborhood for the maximum filter and then smoothing using a 9-km Gaussian filter. The UH probability swaths are subjectively compared to the location of Truth's UH > 1000 m^2 s^{-2}. For better visual comparison and to put the Truth's UH fields on a similar scale to the 3-km ensembles, a maximum filter is applied to the Truth's UH fields using a 3-km neighborhood. To further assess ensemble forecast spread and accuracy, paintball plots (e.g., Skinner et al. 2018) of the ensemble members' column-maximum reflectivity > 25 dBZ at $t = 180$ min are compared to the 25-dBZ contours of the Truth's

column-maximum reflectivity. Also, the ensembles' average maximum UH values are computed for each forecast to illuminate how the forecasts of UH intensity change with initialization time.

To provide an objective measure of forecast skill, the ensemble fractions skill score (eFSS; Duc et al. 2013) is computed for forecast column-maximum reflectivity greater than 25 dBZ using neighborhood widths of 0, 6, 12, 24, 48, 96, and 192 km as well as the full domain (FD). The equation for the eFSS is

$$eFSS = 1 - \frac{\frac{1}{N \times M}\sum_{n=1}^{N}\sum_{m=1}^{M}(P_{n,m}^{o} - P_{n,m}^{f})^2}{\frac{1}{N \times M}\left[\sum_{n=1}^{N}\sum_{m=1}^{M}(P_{n,m}^{o})^2 + \sum_{n=1}^{N}\sum_{m=1}^{M}(P_{n,m}^{f})^2\right]}, \qquad (8)$$

where N and M are the number of spatial windows and ensemble members, respectively; and P^o and P^f are the observed and forecast fractions, respectively, of grid points exceeding the reflectivity threshold within each spatial-ensemble neighborhood window. The eFSS values for each ensemble forecast are aggregated together for each neighborhood width. An eFSS = 1 indicates a perfect forecast. For a "target" skill, FSS$_{uniform}$ from Roberts and Lean (2008) is computed for each forecast period and is the halfway point between the eFSS for a random forecast, which is the observed base rate, and a perfect forecast (eFSS = 1). eFSS values $<$ FSS$_{uniform}$ indicate the forecast might not provide useful information to the end user (e.g., operational forecaster). Also, while column-maximum reflectivity is used to compute eFSS for this study,
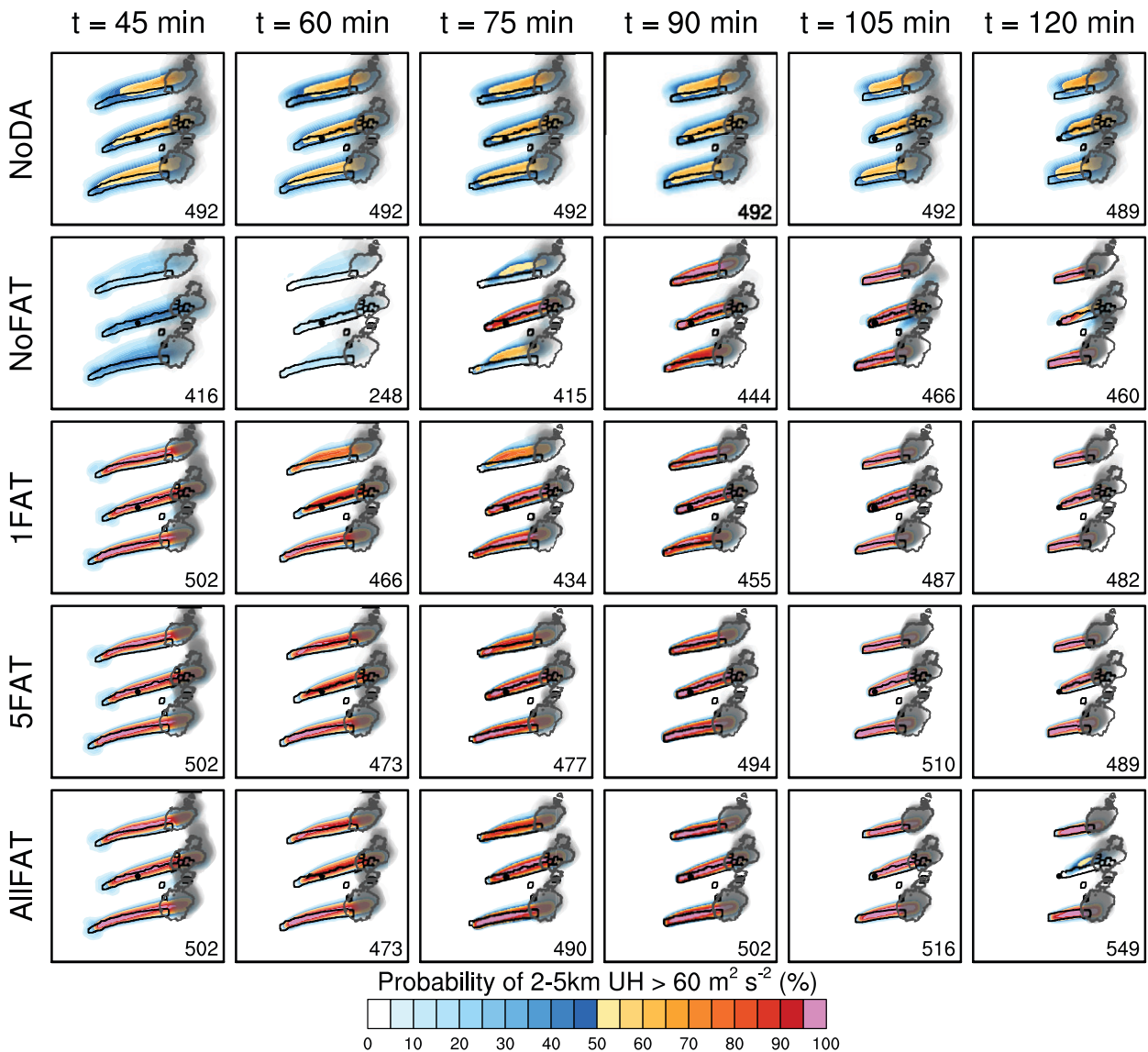
FIG. 5. Probabilities of 2–5-km UH > 60 m^2 s^{-2} (colored shading) and ensemble paintballs of column-maximum reflectivity > 25 dBZ at $t = 180$ min (gray shading) for all NoDisp experiments' forecasts. The Truth's maximum-filtered tracks of UH > 1000 m^2 s^{-2} (black contours) and column-maximum reflectivity > 25 dBZ (dark gray contour) are overlaid. The ensemble average maximum UH value for each forecast is annotated in the bottom right of each subplot.

similar results can be achieved using 2–5-km UH. However, computing eFSS for 2–5-km UH is not as straightforward as for reflectivity since a nonlinear relationship exists for UH between the 250-m and 3-km grids making it more difficult for a direct comparison.

4. Results

a. No initial displacement

This first set of experiments explores the impact of the FAT in a scenario where the ensemble members' individual storm locations are displaced from the Truth storms, but the ensemble mean storms' locations are only minimally displaced. Starting with observation-space diagnostics, all FAT experiments have smaller RMSIs for reflectivity than NoFAT for most analysis times (Fig. 4a). Also, the RMSIs and error growth rates are substantially smaller for the FAT experiments during the first few DA cycles, indicating the FAT helps "spin up" model storms faster. Except from $t = 60$ to 80 min, the three FAT experiments generally have less ensemble spread than the NoFAT experiment (Fig. 4b). Also, the FAT experiments have substantially less spread than NoFAT for the first three DA times due to correcting the initial storm displacements. NoFAT's larger spread for those first few DA cycles results in it having higher consistency ratios for reflectivity
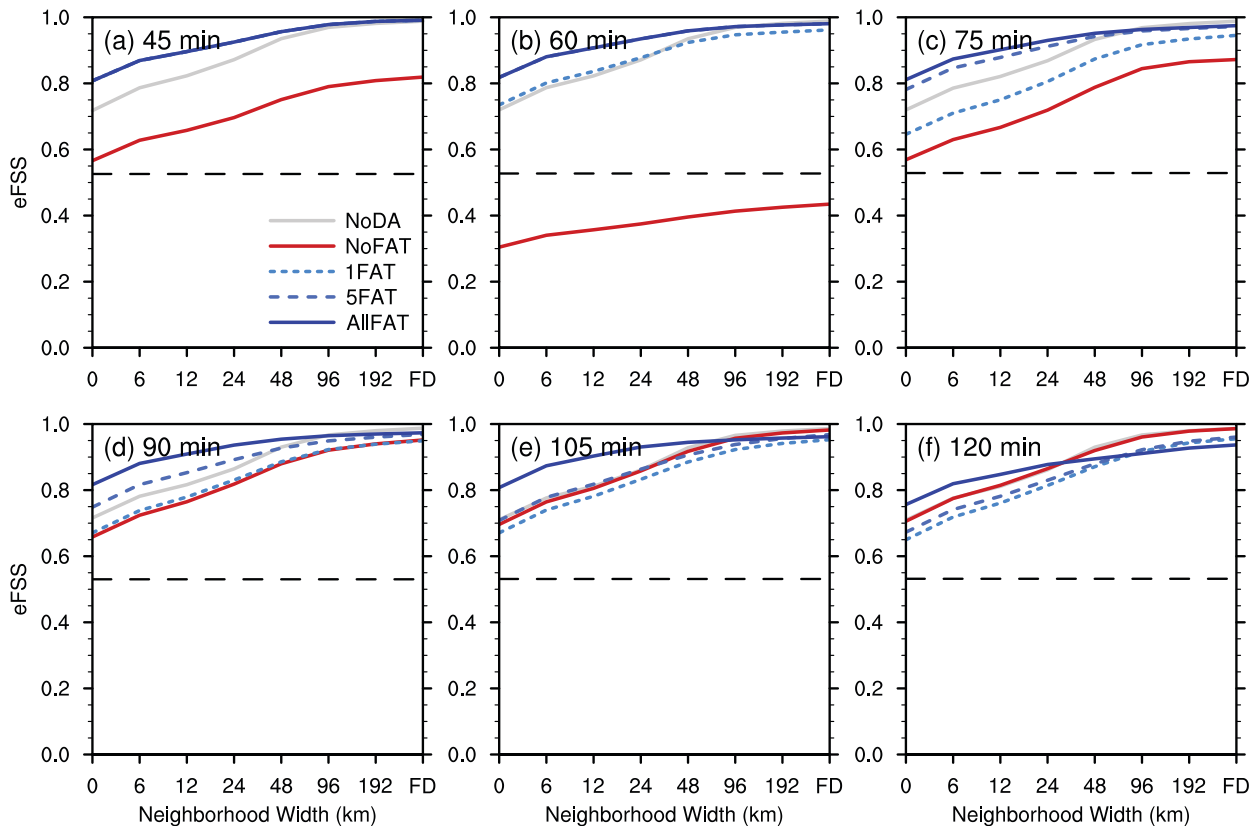
FIG. 6. Average eFSS values for the NoDisp experiments' column-maximum reflectivity for each neighborhood width and forecast period. Black dashed line represents the FSS$_{uniform}$, or target skill.

for those analysis times (Fig. 4c). But, beyond the first few DA cycles, the FAT experiments generally have consistency ratios closer to one. By $t = 120$ min, all experiments have consistency ratios near 0.5, which indicates they are all underdispersive. Using the FAT for all DA times results in the most radar observations being assimilated, while the 1FAT and 5FAT experiments result in the leveling-off of the total number of reflectivity observations being assimilated—similar to the NoFAT experiment—once the FAT ceases to be used (Fig. 4d).

For radial velocity, the AllFAT experiment generally results in larger RMSIs and error growth rates than NoFAT and the other two FAT experiments, especially after $t = 90$ min (Fig. 4e). Similar to reflectivity, the FAT experiments have less spread than NoFAT until $t = 55$ min or so due to correcting the initial storm displacements, but the experiments have similar spread for the remaining analysis times (Fig. 4f). With lower RMSIs and similar spread, NoFAT has higher consistency ratios for most times, while AllFAT has the lowest consistency ratios due to larger RMSIs (Fig. 4g). Also, 1FAT and 5FAT have larger consistency ratios than AllFAT with 5FAT having similar consistency ratios as NoFAT starting at $t = 75$ min. While all experiments assimilate a similar number of radial velocity observations, All-FAT assimilates the fewest (Fig. 4h). In short, the continuous application of the FAT results in smaller errors for reflectivity

and more reflectivity observations being assimilated but also results in larger errors for radial velocity and fewer radial velocity observations being assimilated. This result is consistent with the fact that the FAT corrects the model state using only reflectivity observations and not radial velocity observations. How these observation-space results translate into forecasts of storms will be explored next.

The NoDA ensemble's swaths of probabilities of UH > 60 m^2 s^{-2} reveal the DA experiments are starting from a background forecast with UH probabilities greater than 50% existing along and just north of the Truth's UH tracks for all three storms (first row in Fig. 5). Even with those good priors, however, NoFAT struggles to assimilate the Truth storms into the model until after 30–45 min of cycling (second row in Fig. 5). After the first DA, NoFAT already has substantially smaller UH probabilities and average maximum intensities than NoDA, and its probabilities and intensities bottom out with the forecast initialized at $t = 60$ min. Subsequent forecasts from NoFAT redevelop the storms in the model. These NoFAT results agree with previous radar-DA OSSE studies, such as Xue et al. (2006) and Yussouf and Stensrud (2010), which found that it takes ~10 DA cycles to start producing *accurate* analyses and forecasts of storms. By correcting the ensemble members' individual storm displacements prior to DA, all of the FAT experiments' forecasts produce accurate, high-probability UH swaths with intense UH values for all three storms (last three rows in
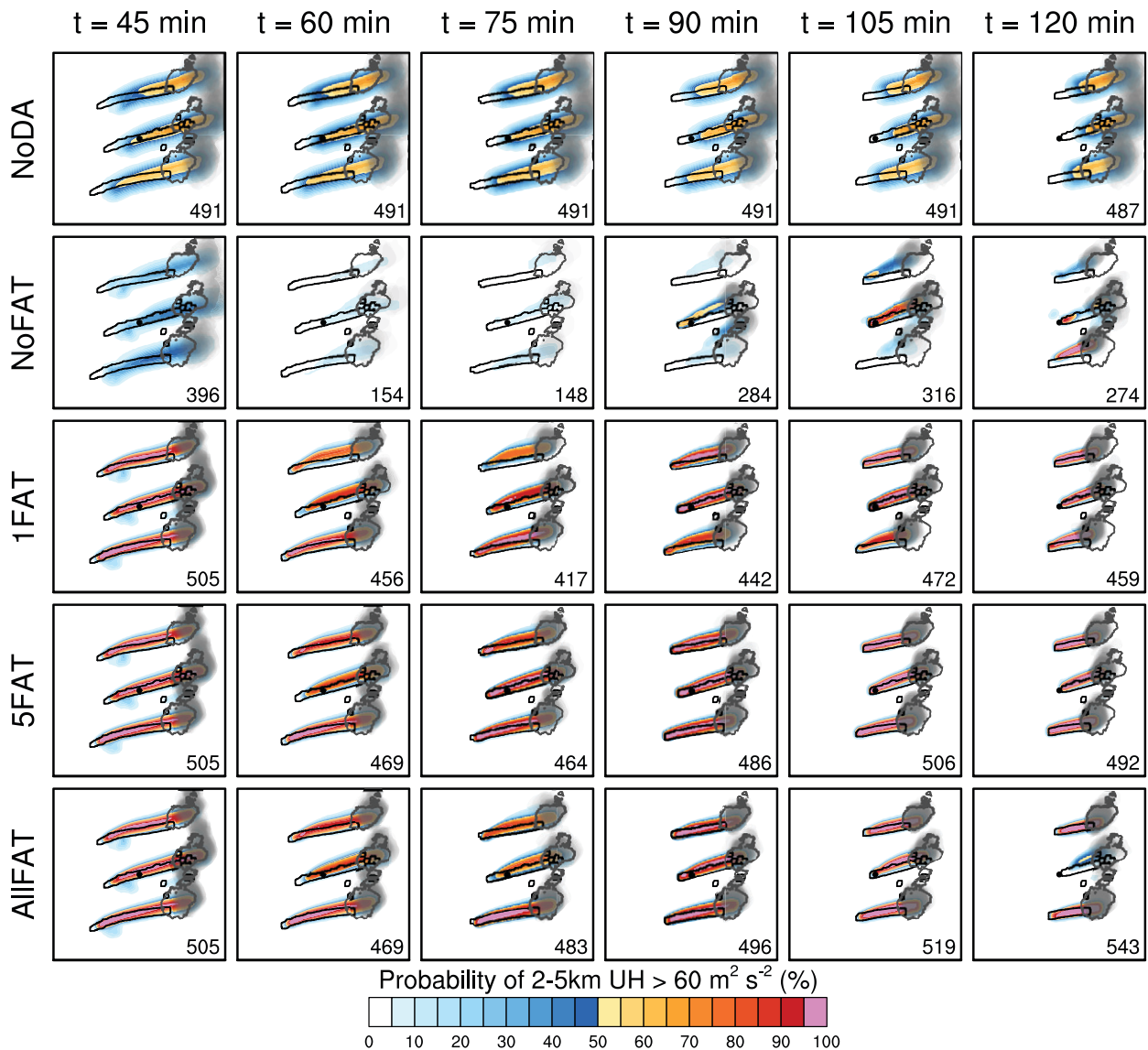
FIG. 7. As in Fig. 5, but for the Disp experiments.

Fig. 5). 1FAT's forecasts initialized at $t = 60$ and 75 min have smaller UH probabilities and intensities than 5FAT and AllFAT (third row of Fig. 5), but improve upon NoFAT for the same times. AllFAT's forecasts perform well until the final initialized forecast, which exhibits lower probabilities along with a deviant motion for the middle storm (last row in Fig. 5). 5FAT's forecasts of UH consistently perform well for all initialization times, including the final initialized forecast (fourth row in Fig. 5).

From an objective perspective on the forecasts, the FAT experiments are also more skillful at forecasting column-maximum reflectivity than the NoFAT experiment for this NoDisp scenario (Fig. 6). Similar to its UH forecast results, NoFAT's eFSS values substantially decrease at all neighborhood widths between the forecast initialized at $t = 45$ min and the forecast initialized at $t = 60$ min (Figs. 6a,b). After $t = 60$ min, NoFAT's skill increases for subsequently initialized

forecasts (Figs. 6c–f). Not surprisingly, all three FAT experiments' forecasts are more skillful than the NoFAT forecasts for most initialization times and scales. Also, the benefit of running the FAT even once is evidenced by the FAT experiments having more skill than NoDA, which is already skillful, at most scales smaller than the full domain (Fig. 6a).

In contrast to those straightforward outcomes, some surprising results are evident. First, only running the FAT once (1FAT) largely prevents the rapid decrease in skill from $t = 45$ min to $t = 60$ min. Second, while AllFAT produces the most skillful forecasts for neighborhood widths up to 48 km, its struggle with the middle storm in the final forecast, as shown in Fig. 5, results in the smallest eFSS values at neighborhood widths greater than 96 km (Fig. 6f). This result indicates that while using the FAT every analysis cycle might lead to more skillful reflectivity forecasts, especially at smaller scales, the repeated use of the
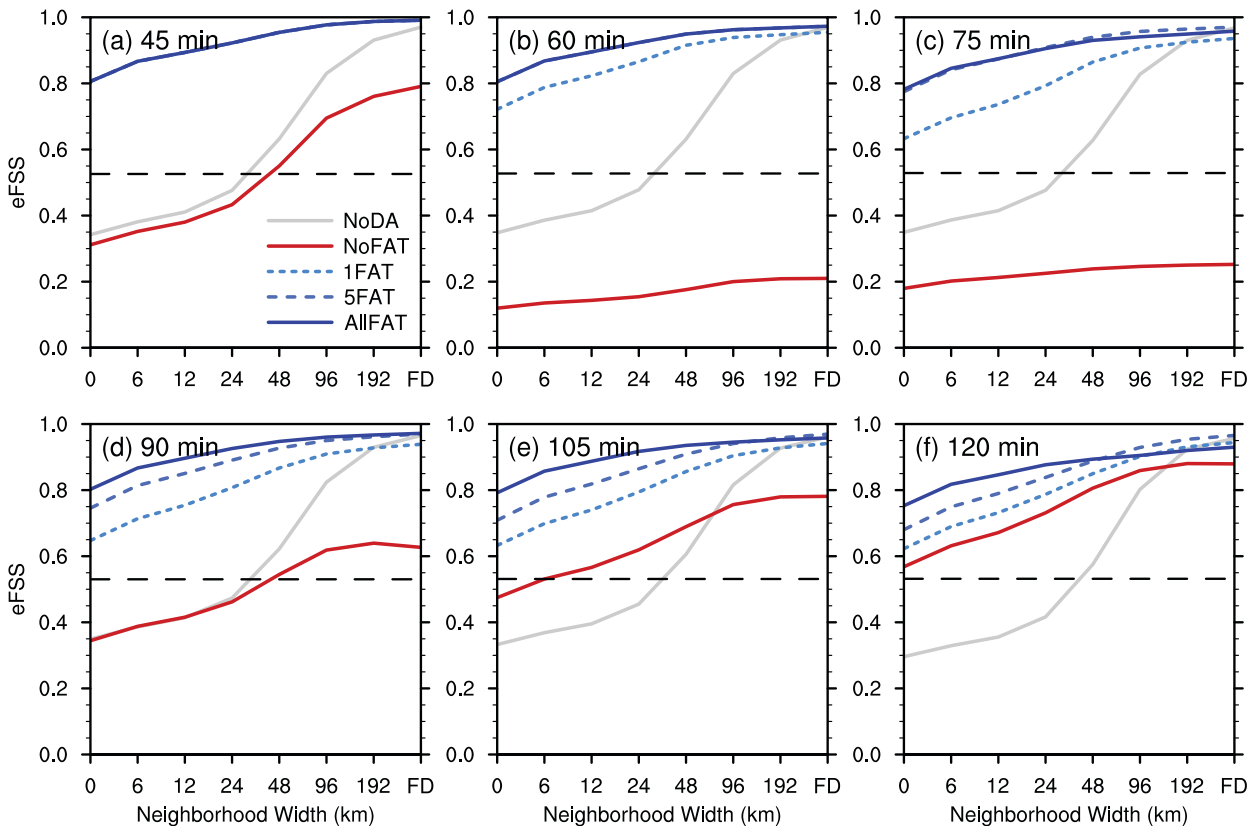
FIG. 8. As in Fig. 6, but for the Disp experiments.

FAT may eventually degrade forecasts of storms. This forecast degradation is at least partly due to the constant use of the FAT leading to minimal ensemble spread. Also, once the FAT stops being used prior to each DA cycle, as in 1FAT and 5FAT, the skill at larger scales is maintained, even though the skill at the smaller scales decreases somewhat. The decrease in skill at the smaller scales for 1FAT and 5FAT is due to small displacements developing after ceasing to run the FAT. Overall, though, the beneficial impact of the FAT is already apparent in this simple multiple-storm scenario—especially at earlier forecast initialization times.

b. Initial displacement

For the Disp set of experiments, a 30-km eastward displacement is added to the initial location of the warm bubbles. The observation-space diagnostics for these experiments are similar to NoDisp's results with the FAT leading to smaller errors for reflectivity, larger errors for radial velocity, and more reflectivity observations being assimilated (not shown). For UH probabilities and intensities, the results are even more striking than for the NoDisp experiments (Fig. 7). NoFAT weakens and dissipates the initial storms during the first 15–30 min of cycling just like for NoDisp, but it struggles to assimilate the storms back into the analyses and forecasts during the remaining cycles (second row in Fig. 7). The initial displacement increases the time needed for NoFAT to "catch up" with the FAT experiments.

Even with the initial displacement added to the storm locations, the three FAT experiments perform just as well as NoDisp's FAT experiments (bottom three rows in Fig. 7), but there are some differences. 1FAT's UH probabilities are lower for some of the storms in the forecasts initialized from $t = 60$ min to $t = 105$ min. AllFAT still has an issue with the middle storm during the final initialized forecast. Also, it has lower probabilities in the forecast initialized at $t = 75$ min. Of the three FAT experiments, 5FAT is the least impacted by the initial displacement error. Also, most of the average maximum UH intensities for all three experiments are up to 5% weaker.

Not surprisingly, the eFSS results agree with our visual assessment of the UH probability swaths and reflectivity paintball contours. NoFAT's eFSS values depict a rapid decrease in skill by $t = 60$ min followed by a slow recovery in skill through the final forecast, when the target skill is once again reached at all neighborhood widths (Fig. 8). Interestingly, NoDA's eFSS results highlight the impact of displacement errors on the eFSS values, which are substantially less for neighborhoods widths less than ~60 km when compared to the FAT experiments and NoDisp's NoDA results. Similar to the NoDisp results, increased usage of the FAT yields consistently more skill at the smaller neighborhood widths. This result is visually apparent with the paintball contours in Fig. 7. For example, 1FAT's reflectivity paintballs for the final forecast are more displaced from the Truth's reflectivity contours than 5FAT's paintballs, and 5FAT's paintballs are more displaced than AllFAT's paintballs. However, for neighborhood

t = 45 min | t = 60 min | t = 75 min | t = 90 min | t = 105 min | t = 120 min
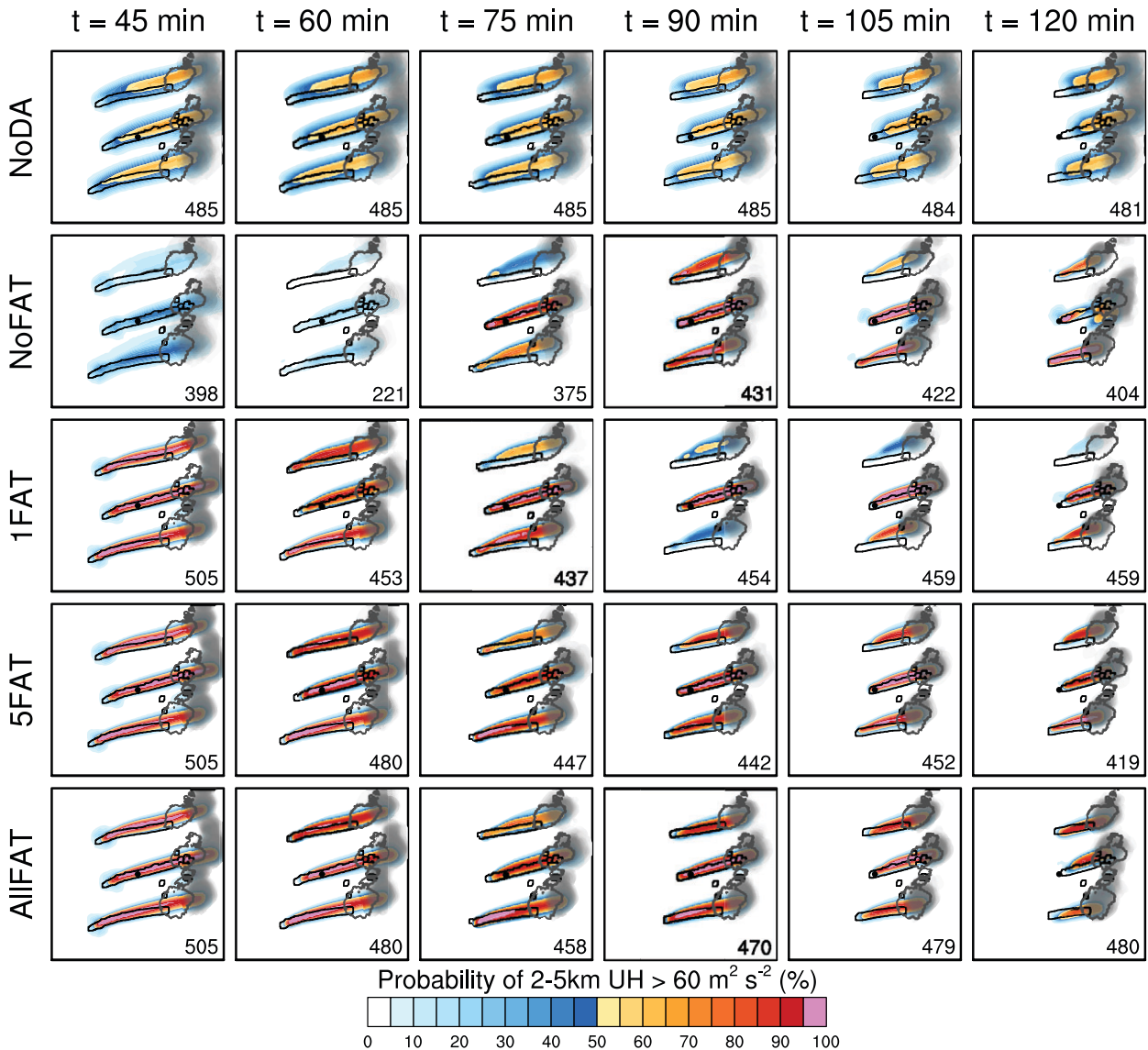

FIG. 9. As in Fig. 5, but for the Fast experiments.

widths greater than ~96 km, 5FAT tends to be the most skillful. The decrease in skill at the largest scales in the last forecast for AllFAT is once again due to its decrease in forecast skill with the middle storm. In summary, these Disp experiments highlight the negative impact of initial storm displacement errors on a forecast system and the FAT's ability to mitigate those negative impacts.

c. Fast storm motion bias

Instead of adding an entire 30-km displacement to the warm bubbles in the initial conditions, the Fast set of experiments slowly adds in the 30-km displacement by adding a storm motion bias to the initial wind profiles. The observation-space diagnostics' results (not shown) are similar to the NoDisp experiments, except for larger errors and more ensemble spread for all experiments. Thus, the consistency ratios are still similar between the two sets of experiments. At $t = 45$ min, the

ensemble mean storm locations are minimally displaced similar to the NoDisp storms (Figs. 3a,c), so besides the longer UH probability swaths, the UH probabilities and intensities are also similar for the first forecast for both sets of experiments (first column in Fig. 9). Again, NoFAT dissipates the original storms in the background forecast and gradually regenerates them over the remaining analyses and forecasts (second row in Fig. 9), yielding decent forecasts for the three storms in the final three forecasts. 1FAT does well for the first three forecasts, but its forecast performance decreases for some of the storms in the last three forecasts (third row in Fig. 9). Conversely, 5FAT and AllFAT both handle the storm motion bias relatively well for *all* forecasts (last two rows in Fig. 9).

After a rapid decrease in skill between $t = 45$ min and $t = 60$ min, NoFAT's skill recovers and is similar to 1FAT's skill at all neighborhood widths for the forecast initialized at $t = 75$ min and similar to 5FAT and AllFAT's skill at all
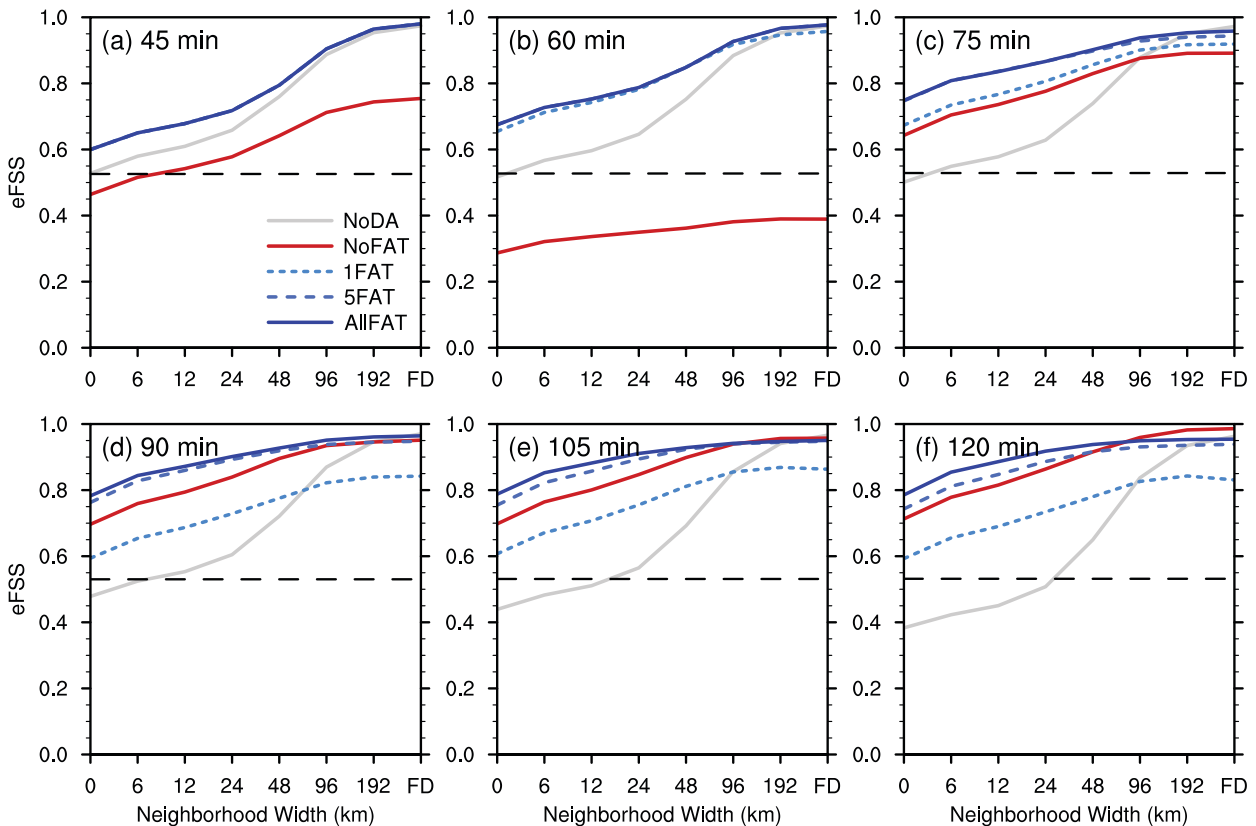
FIG. 10. As in Fig. 6, but for the Fast experiments.

neighborhood widths for the final three forecasts (Fig. 10). 1FAT is the least skillful experiment at all neighborhood widths for the final three forecasts. AllFAT is the most skillful at neighborhood widths less than ~48 km for all forecasts, though 5FAT is nearly as skillful as AllFAT. In the final forecast, NoFAT exhibits the highest skill for scales greater than ~96 km (Fig. 10f), but that higher skill is due to having a lower frequency bias as a result of erroneously having more spurious convection, as evidenced in Fig. 9. Even so, these Fast results show that the use of the FAT throughout the DA period is effective at mitigating the impacts of large storm motion biases on analyses and forecasts of storms.

d. 15-min cycling interval

This next set of experiments explores the impact of cycling less frequently by starting from the NoDisp background fields (Fig. 3a) and using a cycling frequency of 15 min instead of 5 min. Compared to the other NoFAT experiments, NoFAT's initial decrease in forecast performance for Cyc15 is less severe and slower to recover (second row in Fig. 11). For example, NoFAT's UH probabilities and intensities for the forecast initialized at $t = 60$ min are higher for Cyc15 (second row in Fig. 11) than NoDisp (second row in Fig. 5) likely due to not undesirably forcing the model state toward the observations as frequently as for 5-min cycling. Also, based on its lower probabilities and intensities for the forecast initialized at $t = 75$ min,

NoFAT's subsequent recovery is also slower than for NoDisp's NoFAT experiment due to fewer DA times to spin the model up. Even so, the NoFAT experiment performs similarly well as the FAT experiments for the final three forecasts. The three FAT experiments perform well for *all* forecasts, as they did for NoDisp (last three rows in Fig. 11). In fact, AllFAT is able to maintain accurate forecasts of the middle storm through the final forecast.

The eFSS results for Cyc15 generally match the subjective assessment of the UH and reflectivity forecasts and are overall similar to NoDisp's eFSS results (Fig. 12). As for NoDisp, NoFAT's forecast skill decreases from $t = 45$ min to $t = 60$ min, but NoFAT's forecast is more skillful at $t = 60$ min when cycling less frequently (Fig. 12b). Conversely, NoFAT's forecast initialized at $t = 75$ min is less skillful than when cycling more frequently (Fig. 12c). Also, unlike the NoDisp experiments, all three FAT experiments' forecasts are more skillful than the NoFAT's forecasts at all scales for the final three forecasts. In short, the model analyses and forecasts take longer to recover from the storm displacement errors when cycling less frequently. Also, the FAT is able to minimize this dependence on cycling frequency and continue to provide accurate forecasts.

e. Initially missing storms

All of the previous scenarios started out with an equal number of observed and forecast storms, so even though a
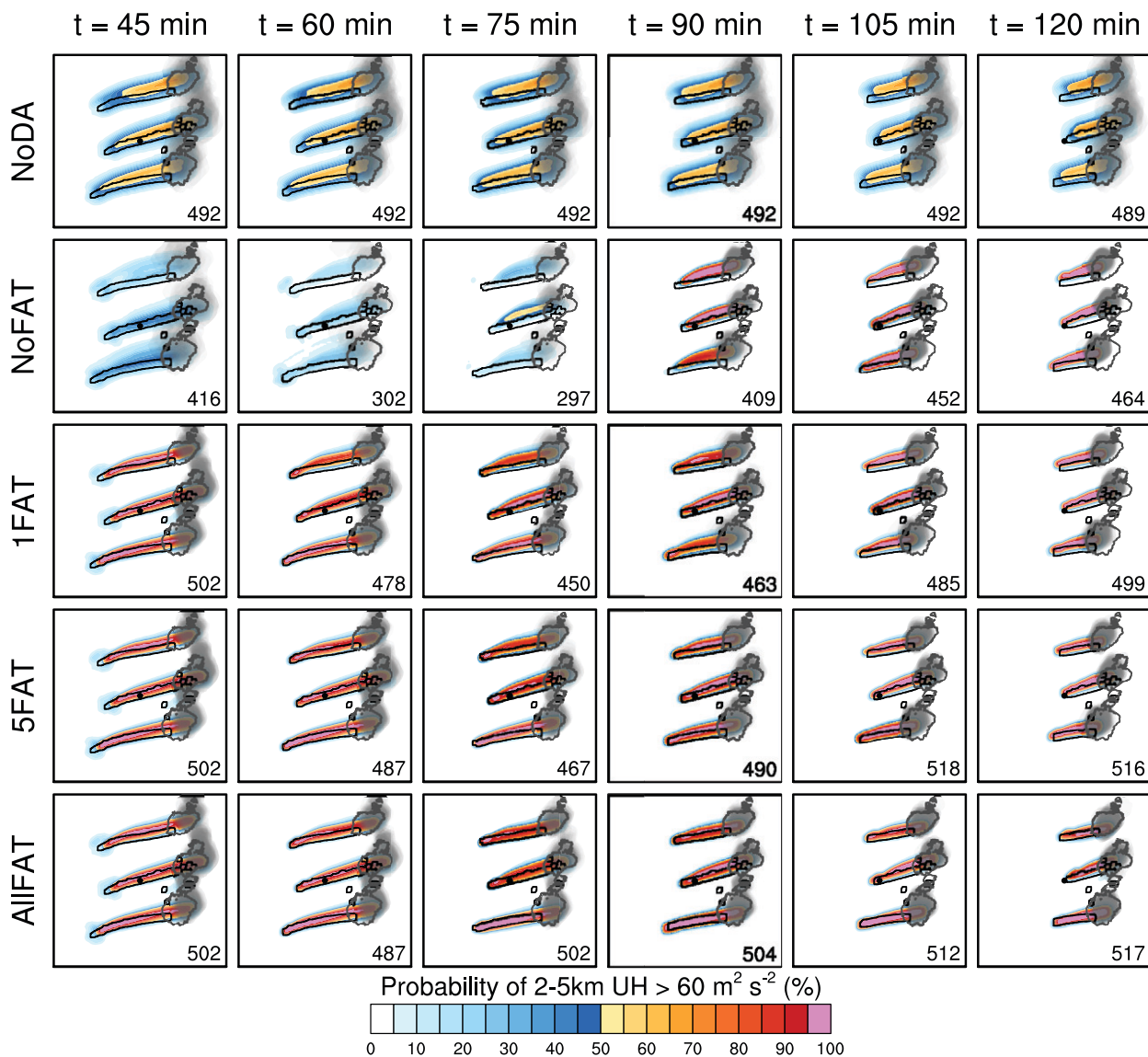
FIG. 11. As in Fig. 5, but for the Cyc15 experiments.

mismatch in the number of storms did occur for some experiments at later DA times, the previous experiments did not test the improved capability of the FAT to handle cases with an *initial* mismatch in the number of storms. The Miss scenario tests that new capability by randomly removing 30% of the initial warm bubbles. As in NoDisp, NoFAT has a rapid decrease in forecast quality within the first 15 min of DA cycling and recovers by $t = 90$ min to produce three decent forecasts (second row in Fig. 13). While 1FAT's UH probabilities and intensities are substantially smaller than in NoDisp for the forecasts initialized at $t = 60$ min and $t = 75$ min (third row in Fig. 13), its UH probabilities and intensities for these two forecasts remain larger than for NoFAT. Also, 1FAT's last three forecasts perform similarly well to the other NoFAT and FAT experiments. Both 5FAT and AllFAT have larger UH probabilities and

intensities than NoFAT and 1FAT through the first three forecasts (bottom two rows in Fig. 13), while 5FAT consistently has the highest probabilities of UH for all three storms in the final four forecasts. Conversely, AllFAT generally has smaller probabilities of UH than the other DA experiments for all three storms.

While all experiments exhibit a decrease in skill between the forecasts initialized at $t = 45$ min and $t = 60$ min, No-FAT experiences the largest decrease in skill (Fig. 14b). Starting with the forecast initialized at $t = 75$ min, NoFAT's eFSS values exceed the target skill at all neighborhood widths again and are similar to the eFSS values of the AllFAT and 5FAT at most neighborhood widths (Figs. 14d–f). 1FAT is generally the least skillful for the last four forecasts, especially the forecasts initialized at $t = 75$ min and $t = 120$ min. Except for the forecast initialized at
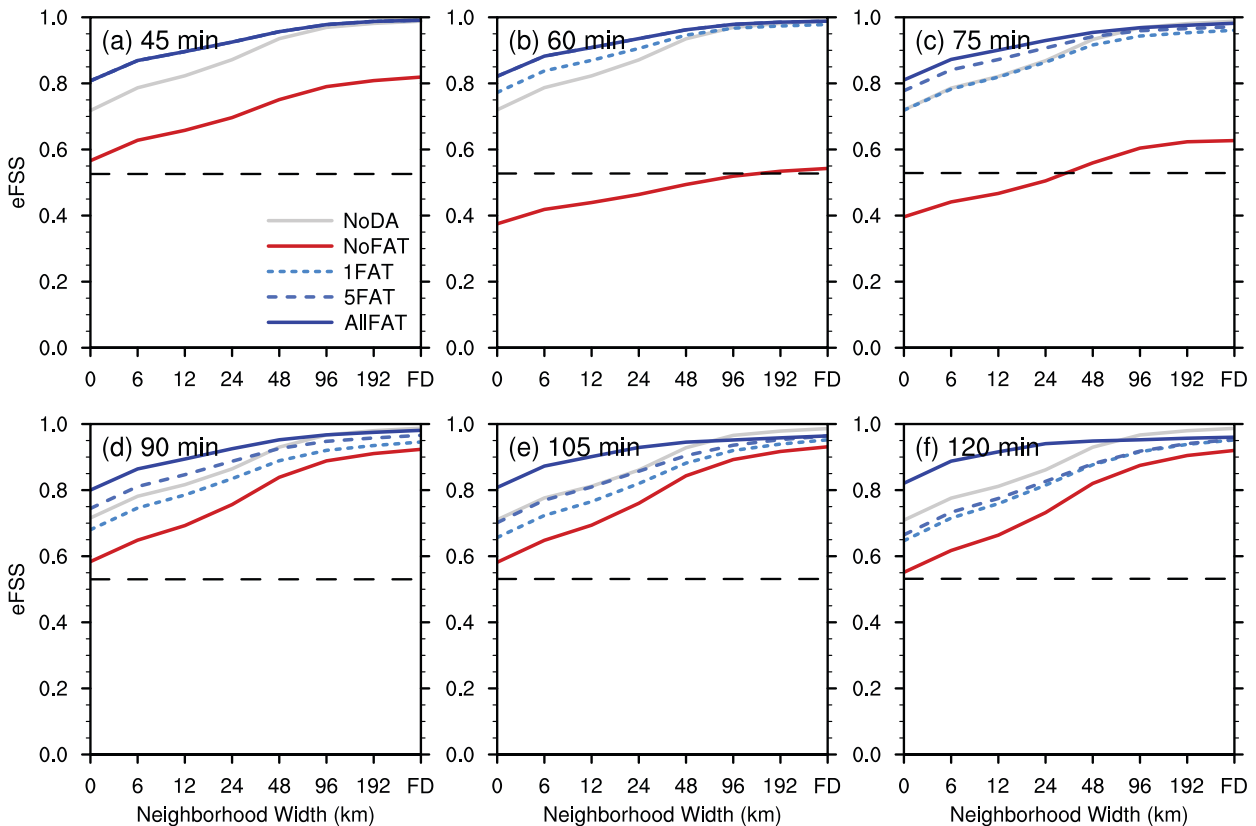
FIG. 12. As in Fig. 6, but for the Cyc15 experiments.

$t = 75$ min, AllFAT is the most skillful at neighborhood widths less than ~24 km. While 5FAT outperforms the other FAT experiments at the largest scales for the final four forecasts, its skill is no better than NoFAT's skill at those scales. Even so, the FAT is still able to help mitigate the decrease in skill during the spinup period and help provide more skill at the smallest scales for all forecasts.

f. Complex storm scenario

This final set of experiments adds realism to the simulations by increasing storm interactions through the addition of another storm and by starting from a decent background for one of the storms (the southern one) but not for the others. As a result, NoFAT produces good UH and reflectivity forecasts for the southern storm, but struggles with the other storms, especially the northern storm, which had the initial westward displacement (second row in Fig. 15). As with the other scenarios, NoFAT exhibits a decrease in forecast performance from $t = 45$ min to $t = 60$ min. Interestingly, that decrease is not evident with the southern storm in this case. NoFAT also produces good forecasts of the two middle storms starting at $t = 90$ min. In fact, NoFAT does well in predicting the demise of the second-most southern storm. Conversely, NoFAT fails to ever redevelop the northern storm.

From a subjective perspective, all three FAT experiments produce similarly better forecasts of UH and reflectivity than NoFAT, based on the higher UH probabilities along the Truth's UH tracks and the reflectivity paintballs filling in the Truth's reflectivity contours at the end of the forecast periods (Fig. 15). It is clear that the FAT experiments do much better in forecasting the northern storm than NoFAT due to correcting the initial storm displacement errors. Interestingly, at $t = 45$ min, the FAT much better predicts one of the middle two storms (the southern one) than the other, even though the initial ensemble-mean warm bubble was located halfway between the two middle Truth warm bubbles. Starting with forecasts initialized at $t = 60$ min, all three FAT experiments predict both of these middle storms well. For the southernmost storm in the domain, the three FAT experiments have lower UH probabilities than NoFAT in the forecasts initialized at $t = 60$ min and $t = 75$ min.

The eFSS values for column-maximum reflectivity agree with the subjective evaluation—all three FAT experiments are more skillful than the NoFAT experiment at forecasting the Complex scenario (Fig. 16). Starting with the forecast initialized at $t = 75$ min, 1FAT and 5FAT are more skillful than AllFAT for all neighborhood widths, indicating again that correcting the storm displacement errors at every DA step may not be optimal in all cases. Even so, all three FAT
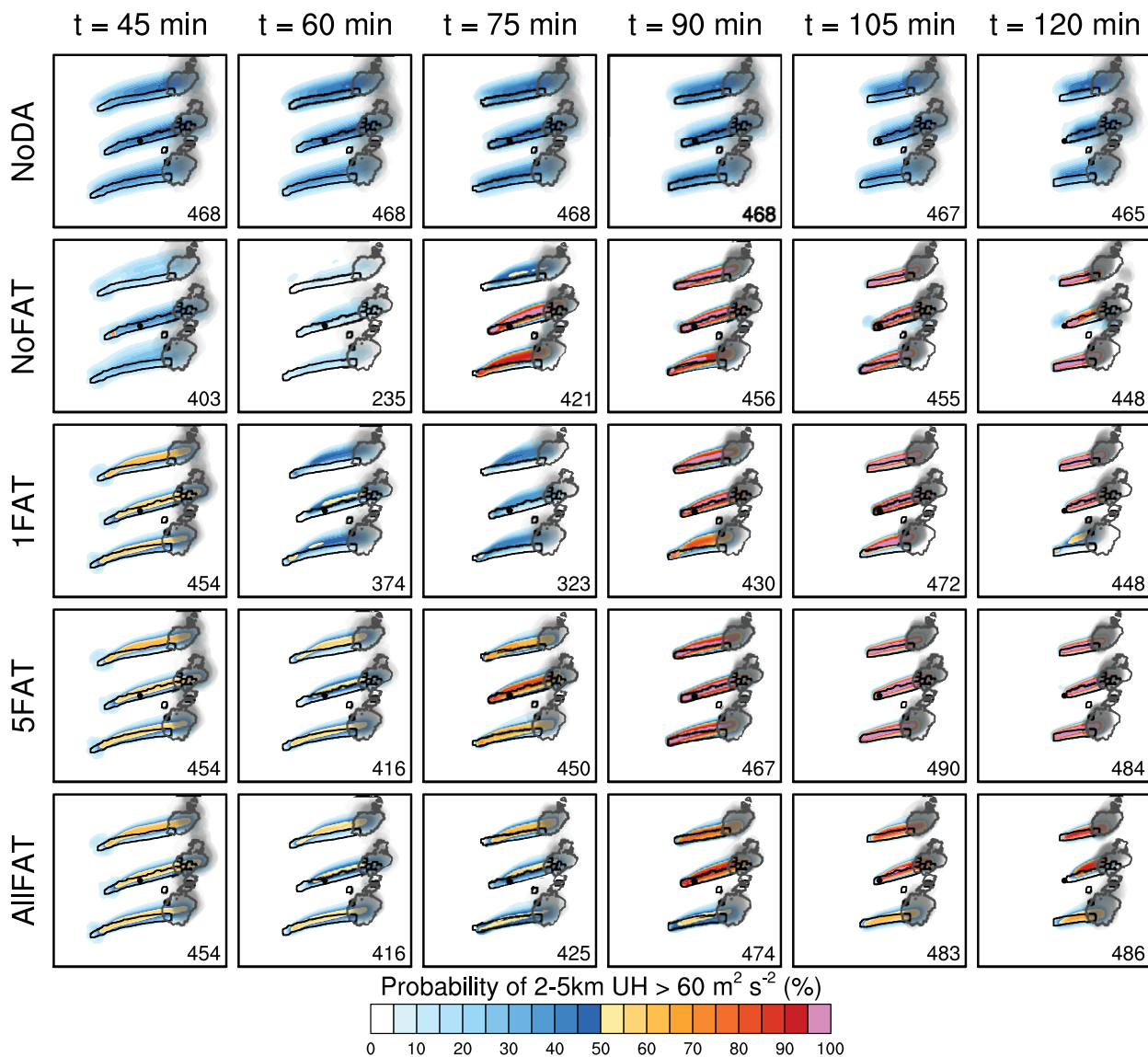
FIG. 13. As in Fig. 5, but for the Miss experiments.

experiments have substantially more skill than NoFAT for most scales and forecasts in this more complex scenario.

5. Summary and discussion

This study expands on the work by S18 by implementing an object-based merging and matching technique into the FAT to handle multiple storms and testing the updated FAT within an idealized OSSE framework using multiple-storm scenarios. As in S18, results from these new experiments reveal the potential benefits to ensemble analyses and forecasts of using the FAT to correct storm displacement errors prior to data assimilation. All three FAT experiments in this study performed subjectively and objectively better than NoFAT in most of the investigated scenarios. Even in the no-initial-displacement

scenario, where the ensemble mean storm displacements were negligible, correcting the individual ensemble members' storm displacement errors with the FAT led to improved UH and reflectivity forecasts. These forecast improvements were largest at early assimilation cycles, where the NoFAT forecasts in all scenarios exhibited a decrease in forecast quality lasting for several DA cycles, in agreement with previous OSSE studies. Except for the Fast case, just one completion of the FAT, as in 1FAT, resulted in markedly improved analyses and forecasts over the NoFAT experiments. Additionally, running the FAT for the first five DA cycles, as in 5FAT, resulted in the best overall performance of all the experiments. At the smaller scales though, the AllFAT experiment showed that continuously running the FAT results in more skillful forecasts of column-maximum reflectivity at smaller scales.
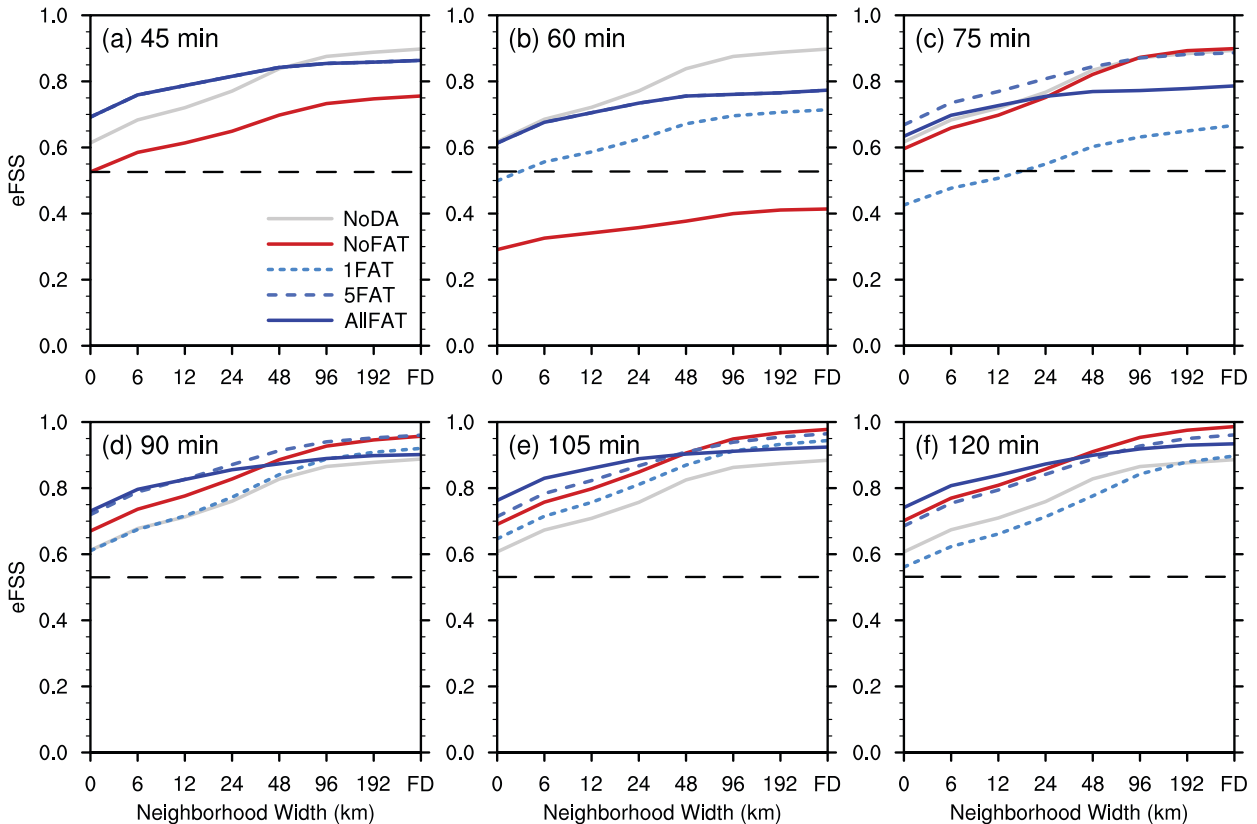
FIG. 14. As in Fig. 6, but for the Miss experiments.

Some more nuanced results also emerged. First, as mentioned above, 1FAT performed well in all scenarios, except for the Fast scenario, which features a large, but reasonable, storm motion bias. The Complex scenario also featured an added storm motion bias, but EAKF DA was able to handle this storm motion bias after running the FAT only once. While both scenarios have storm motion biases, the Complex scenario's storm motion bias is about half the magnitude of the Fast scenario's storm motion bias. This result indicates that in cases of relatively minor storm motion biases, it may be only necessary to run the FAT once, or at least less often, during DA cycling. In addition, some of the experiments showed that continuously running the FAT can cause issues. For example, for the NoDisp and Disp cases, AllFAT has a poorer final forecast for the middle storm than 1FAT and 5FAT. Based on additional experiments where the radar location was shifted ~30 km either to the south or east (not shown), this issue is not due to the location of the radar in the domain.

The crux of this AllFAT problem is likely hinted at by the NoDisp observation-space diagnostics for radial velocity observations, where AllFAT has higher RMSIs than the other experiments for most analysis times. This issue is likely due to the FAT solving for the displacement vectors using only column-maximum reflectivity observations. Thus, the FAT may suboptimally correct the locations of

the model storms in instances where optimizing the alignment with the observed column-maximum reflectivity produces misalignment with the observed radial winds. For example, given the differences in the storm structures and shapes between the Truth simulations and the ensembles, which are mostly due to the differences in grid spacing and microphysics scheme, the storm-relative locations of the rotating updrafts may be different for the Truth and ensemble storms (Fig. 17), resulting in larger innovations and fewer radial velocity observations being assimilated than for the NoFAT experiments. This issue is also likely the culprit to the lower UH probabilities for AllFAT in the Miss scenario. Since fewer radial velocity observations are being assimilated, AllFAT likely struggles to maintain the storms that are being added by the assimilation of reflectivity observations. Without strong, rotating updrafts to support the three storms being assimilated, more of the storms in All-FAT's forecasts weaken and eventually die off.

One possible way to help mitigate these issues for AllFAT is to use radial velocity observations along with the reflectivity observations in the FAT's cost function – something that will be explored in a future study. Another potential solution is to use a finer ensemble grid, such as 1 or 2 km, so that the forecast storms are more similar in structure to the observed storms. Also, this study used the same observation error standard deviation in the FAT's residual cost function as was used
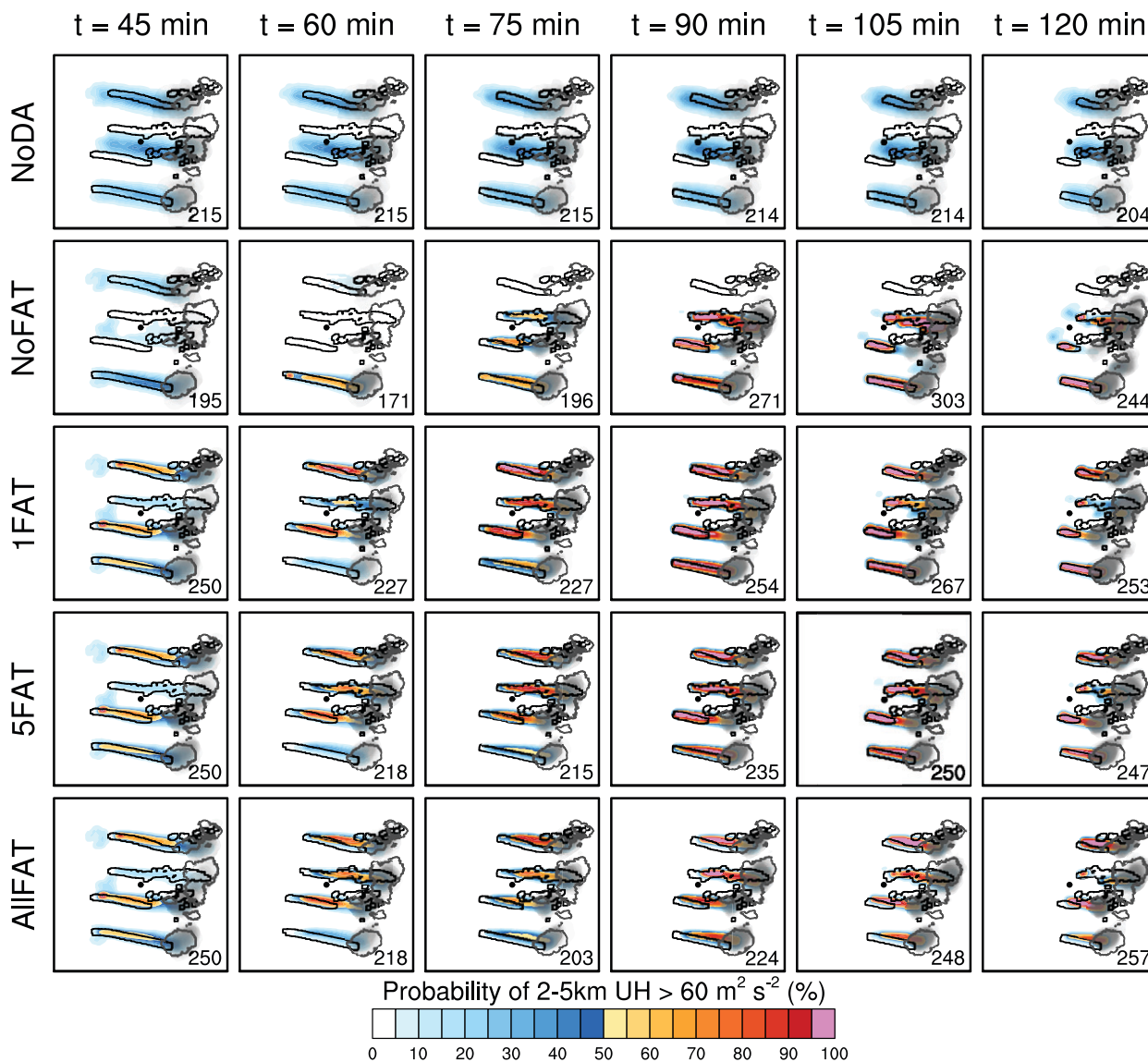
FIG. 15. As in Fig. 5, but for the Complex experiments.

in creating the synthetic reflectivity observations. However, using a small observation error standard deviation (i.e., 2 dB*Z*) in the residual cost function results in a closer fit of the forecast reflectivity to the observed reflectivity. Thus, using a larger observation error standard deviation, such as 5 dB*Z*, could introduce enough uncertainty into the FAT to help prevent this issue of overfitting the reflectivity observations without needing to use the radial velocity observations in the FAT.

The Cyc15 experiments demonstrated that the issues with AllFAT could also potentially be mitigated by cycling less frequently. The Cyc15 experiments showed the FAT may be more beneficial and produce better forecasts at longer (than 5 min) cycling intervals, such as the 15-min interval used by the experimental WoFS. These results support the conclusion in Ying (2019) that there may be an optimal

cycling interval for EnKF DA systems coupled with alignment techniques. Importantly, the fact that AllFAT at 15-min cycling is better than NoFAT and AllFAT at 5-min cycling suggests that the use of the FAT or other alignment techniques could reduce the cycling frequency requirement for storm-scale ensembles.

Additional experiments were run using the FAT every 15 min while cycling with EAKF every 5 min (not shown). Those experiments also show promise in mitigating any issues with continuously running the FAT. These results along with the 1FAT and 5FAT results and the Cyc15 results motivate the need for an *adaptive* FAT approach. In a way, the FAT presented here is already adaptive with the minimum displacement vector and decreasing-RMSE requirements, which can "adaptively" determine when not to adjust the model state variables with the displacement
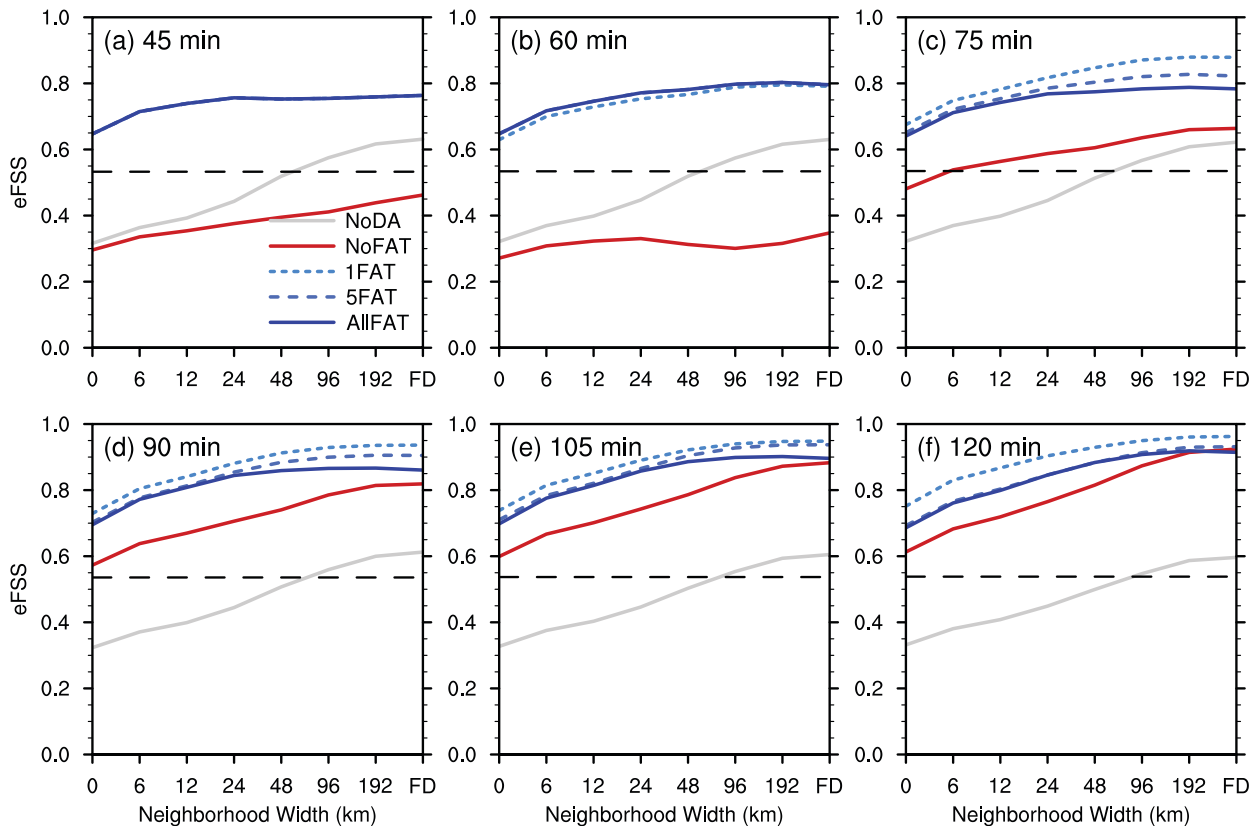
FIG. 16. As in Fig. 6, but for the Complex experiments.

vectors. However, those requirements are on a per-ensemble-member-basis, so some members' background forecasts will be adjusted while other members' forecasts will not be adjusted at the same time. Thus, a future research path to explore is to adaptively determine when the FAT should be run for the entire ensemble. For example, ensemble mean innovations could potentially be used to determine when the FAT should run; the FAT probably should not run when the mean innovations are large for radial velocity but small for reflectivity.

So far, the FAT has only been tested in an idealized framework using OSSEs with homogeneous environments and isolated supercell storms. In future studies, the FAT will be further tuned and tested using real-data cases with a DA and forecast system similar to the current experimental WoFS (Jones et al. 2020). Testing the FAT with multiple real-data cases and convective modes will give a better sense of the benefits of correcting storm displacement errors in real-world ensembles, whose forecast errors may not be well approximated by OSSEs. These real-data tests could also reveal potential issues with correcting storm displacement errors in nonhomogeneous environments. For example, the FAT could potentially introduce imbalances in the model state wind fields by introducing spurious areas of convergence and divergence. Also, while the FAT as designed here should handle non-supercell-shaped storms well, more thorough tests with

more complex storm shapes and modes have yet to be completed and could require additional modifications to the FAT methodology. Before running real-data experiments, the FAT will need to be further modified to account for terrain effects, as implemented by Nehrkorn et al. (2014). Additionally, we are interested in comparing the FAT-EnKF system with localized particle filter (LPF) and hybrid LPF-EnKF DA systems (Grooms and Robinson 2021; Poterjoy 2022) since the FAT and PF methods both address non-Gaussianity in different ways.
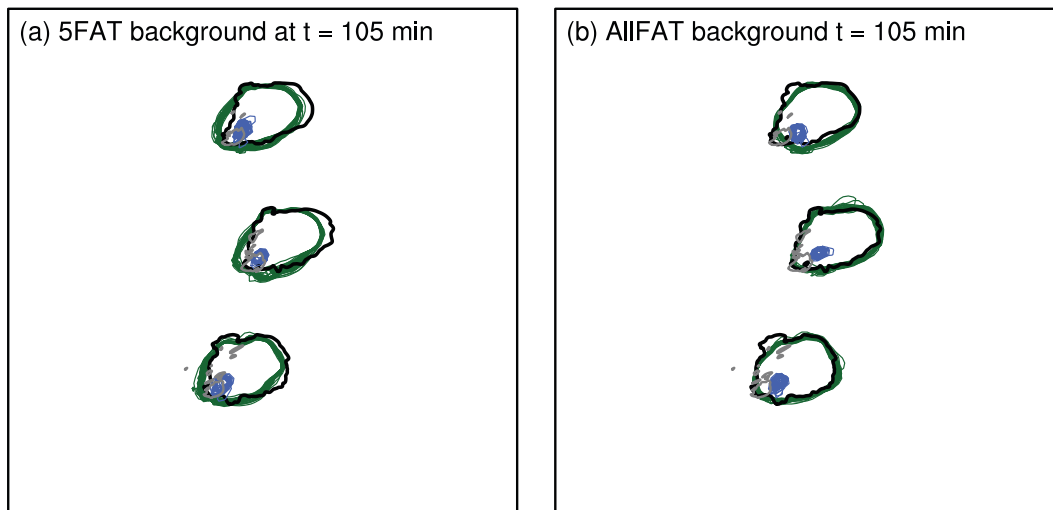
Fig. 17. Ensemble background of 25-dBZ column-maximum reflectivity contours (green lines) and 20 m s^{-1} column-maximum vertical-wind contours (blue lines) for (a) 5FAT and (b) AllFAT at $t = 105$ min. The Truth's column-maximum reflectivity contour (black line) and 20 m s^{-1} column-maximum vertical-wind contour (gray line) at $t = 105$ min are also shown.

computing assistance. We also thank the two anonymous reviewers for their comments and suggestions, which greatly improved the paper.

Data availability statement. The model output and code used in this study are not currently available in a publicly accessible repository. The model output and code used to generate the results herein are available from the authors 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. Arellano, 2009: The Data Assimilation Research Testbed: A community facility. *Bull. Amer. Meteor. Soc.*, **90**, 1283–1296, https://doi.org/10.1175/2009BAMS2618.1.

Brewster, K., 2003: Phase-correcting data assimilation and application to storm-scale numerical weather prediction. Part I: Method description and simulation testing. *Mon. Wea. Rev.*, **131**, 480–492, https://doi.org/10.1175/1520-0493(2003)131<0480:PCDAAA>2.0.CO;2.

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 C. Snyder, 2011: Ensemble Kalman filter assimilation of radar observations of the 8 May 2003 Oklahoma City supercell: Influences of reflectivity observations on storm-scale analyses. *Mon. Wea. Rev.*, **139**, 272–294, https://doi.org/10.1175/2010MWR3438.1.

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.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, **99**, 10 143–10 162, https://doi.org/10.1029/94JC00572.

Flora, M. L., P. S. Skinner, C. K. Potvin, A. E. Reinhart, T. A. Jones, N. Yussouf, and K. H. Knopfmeier, 2019: Object-based verification of short-term, storm-scale probabilistic mesocyclone guidance from an experimental Warn-on-Forecast System. *Wea. Forecasting*, **34**, 1721–1739, https://doi.org/10.1175/WAF-D-19-0094.1.

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

Grooms, I., and G. Robinson, 2021: A hybrid particle-ensemble Kalman filter for problems with medium nonlinearity. *PLOS ONE*, **16**, e0248266, https://doi.org/10.1371/journal.pone.0248266.

Houtekamer, P. L., and F. Zhang, 2016: Review of the ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, **144**, 4489–4532, https://doi.org/10.1175/MWR-D-15-0440.1.

Jones, T. A., P. Skinner, N. Yussouf, K. Knopfmeier, A. Reinhart, and D. Dowell, 2019: Forecasting high-impact weather in landfalling tropical cyclones using a Warn-on-Forecast System. *Bull. Amer. Meteor. Soc.*, **100**, 1405–1417, https://doi.org/10.1175/BAMS-D-18-0203.1.

——, and Coauthors, 2020: Assimilation of *GOES-16* radiances and retrievals into the Warn-on-Forecast System. *Mon. Wea. Rev.*, **148**, 1829–1859, https://doi.org/10.1175/MWR-D-19-0379.1.

Kain, J. S., and Coauthors, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting*, **23**, 931–952, https://doi.org/10.1175/WAF2007106.1.

Kerr, C. A., and X. Wang, 2020: Ensemble-based targeted observation method applied to radar radial velocity observations on idealized supercell low-level rotation forecasts: A proof of concept. *Mon. Wea. Rev.*, **148**, 877–890, https://doi.org/10.1175/MWR-D-19-0197.1.

Mansell, E. R., C. L. Ziegler, and E. C. Bruning, 2010: Simulated electrification of a small thunderstorm with two-moment bulk microphysics. *J. Atmos. Sci.*, **67**, 171–194, https://doi.org/10.1175/2009JAS2965.1.

——, D. T. Dawson II, and J. M. Straka, 2020: Bin-emulating hail melting in three-moment bulk microphysics. *J. Atmos. Sci.*, **77**, 3361–3385, https://doi.org/10.1175/JAS-D-19-0268.1.

Nehrkorn, T., B. Woods, T. Auligné, and R. N. Hoffman, 2014: Application of feature calibration and alignment to high-resolution analysis: Examples using observations sensitive to cloud and water vapor. *Mon. Wea. Rev.*, **142**, 686–702, https://doi.org/10.1175/MWR-D-13-00164.1.

——, ——, R. N. Hoffman, and T. Auligné, 2015: Correcting for position errors in variational data assimilation. *Mon. Wea. Rev.*, **143**, 1368–1381, https://doi.org/10.1175/MWR-D-14-00127.1.

Polak, E., and G. Ribière, 1969: Note sur la convergence de méthodes de directions conjuguées. *ESAIM: Math. Modell. Numer. Anal.*, **3**, 35–43, http://eudml.org/doc/193115.

Poterjoy, J., 2022: Regularization and tempering for a moment-matching localized particle filter. *Quart. J. Roy. Meteor. Soc.*, https://doi.org/10.1002/qj.4328, in press.

Potvin, C. K., and M. L. Flora, 2015: Sensitivity of idealized supercell simulations to horizontal grid spacing: Implications for Warn-on-Forecast. *Mon. Wea. Rev.*, **143**, 2998–3024, https://doi.org/10.1175/MWR-D-14-00416.1.

——, and Coauthors, 2020: Assessing systematic impacts of PBL schemes on storm evolution in the NOAA Warn-on-Forecast System. *Mon. Wea. Rev.*, **148**, 2567–2590, https://doi.org/10.1175/MWR-D-19-0389.1.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, https://doi.org/10.1175/2007MWR2123.1.

Schwartz, C. S., and R. A. Sobash, 2017: Generating probabilistic forecasts from convection-allowing ensembles using neighborhood approaches: A review and recommendations. *Mon. Wea. Rev.*, **145**, 3397–3418, https://doi.org/10.1175/MWR-D-16-0400.1.

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.

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 D. J. Stensrud, 2013: The impact of covariance localization for radar data on EnKF analyses of a developing MCS: Observing system simulation experiments. *Mon. Wea. Rev.*, **141**, 3691–3709, https://doi.org/10.1175/MWR-D-12-00203.1.

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.

——, and Coauthors, 2013: Progress and challenges with Warn-on-Forecast. *Atmos. Res.*, **123**, 2–16, https://doi.org/10.1016/j.atmosres.2012.04.004.

Stratman, D. R., and K. A. Brewster, 2017: Sensitivities of 1-km forecasts of 24 May 2011 tornadic supercells to microphysics parameterizations. *Mon. Wea. Rev.*, **145**, 2697–2721, https://doi.org/10.1175/MWR-D-16-0282.1.

——, C. K. Potvin, and L. J. Wicker, 2018: Correcting storm displacement errors in ensemble using the feature alignment technique (FAT). *Mon. Wea. Rev.*, **146**, 2125–2145, https://doi.org/10.1175/MWR-D-17-0357.1.

Wheatley, D. M., K. H. Knopfmeier, T. A. Jones, and G. J. Creager, 2015: Storm-scale data assimilation and ensemble forecasting with the NSSL experimental Warn-on-Forecast System. Part I: Radar data experiments. *Wea. Forecasting*, **30**, 1795–1817, https://doi.org/10.1175/WAF-D-15-0043.1.

Xue, M., M. Tong, and K. K. Droegemeier, 2006: An OSSE framework based on the ensemble square-root Kalman filter for evaluating impact of data from radar networks on thunderstorm analysis and forecast. *J. Atmos. Oceanic Technol.*, **23**, 46–66, https://doi.org/10.1175/JTECH1835.1.

Ying, Y., 2019: A multiscale alignment method for ensemble filtering with displacement errors. *Mon. Wea. Rev.*, **147**, 4553–4565, https://doi.org/10.1175/MWR-D-19-0170.1.

Yussouf, N., and D. J. Stensrud, 2010: Impact of phased-array radar observations over a short assimilation period: Observing system simulation experiments using an ensemble Kalman filter. *Mon. Wea. Rev.*, **138**, 517–538, https://doi.org/10.1175/2009MWR2925.1.

——, T. A. Jones, and P. S. Skinner, 2020: Probabilistic high-impact rainfall forecasts from landfalling tropical cyclones using Warn-on-Forecast System. *Quart. J. Roy. Meteor. Soc.*, **146**, 2050–2065, https://doi.org/10.1002/qj.3779.