Severe Weather Verification of an FV3-LAM Regional Ensemble during the 2022 NOAA Hazardous Weather Testbed Spring Forecasting Experiment

MARCUS JOHNSON,[a] NATHAN SNOOK,[a] JUN PARK,[a] MING XUE,[a,b] KEITH A. BREWSTER,[a,b] TIMOTHY SUPINIE,[c] AND XIAO-MING HU[a]

[a] *Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, Oklahoma*
[b] *School of Meteorology, University of Oklahoma, Norman, Oklahoma*
[c] *NOAA/Storm Prediction Center, Norman, Oklahoma*

(Manuscript received 23 February 2024, in final form 15 October 2024, accepted 5 December 2024)

ABSTRACT: As part of the 2022 NOAA Hazardous Weather Testbed Spring Forecasting Experiment, the Center for Analysis and Prediction of Storms produced Finite-Volume Cubed-Sphere–Based Limited Area Model (FV3-LAM) real-time ensemble forecasts to study its use in convection-allowing ensemble forecasts for the severe weather forecasting problem and to inform the optimization of the upcoming operational Rapid Refresh Forecast System. We evaluate deterministic and ensemble forecasts in terms of surrogate severe weather reports (SSRs) and surrogate severe probability forecasts (SSPFs) created from simulated 0–3- and 2–5-km updraft helicity (UH) and 10-m wind speed. Forecasts are verified against observed storm reports (OSRs) and observed severe probabilistic fields (OSPFs) derived from tornado, hail, wind, and all types of local storm reports, and 0600 UTC day 1 convective outlooks issued by the Storm Prediction Center (SPC) for three cases. UH ensemble SSPFs have better reliability and discrimination when verified with OSRs from all storm reports. Spatial smoothing generally increases reliability, while smaller smoothing lengths optimize discrimination ability. Case studies demonstrate that UH SSPFs are consistent with SPC day 1 convective outlooks, indicating that these forecasts are qualitatively similar to operational guidance. The ensemble mean of SSRs is generally more skillful than individual members when based on UH but not 10-m wind. In fact, SSRs and SSPFs based on 10-m wind display little skill in predicting severe wind events; we therefore conclude that UH seems to better inform severe hazard risk, even compared to prognosed surface wind speed. Model resolution dictates its ability to prognose rotating updrafts and severe wind.

KEYWORDS: Ensembles; Forecast verification/skill; Numerical weather prediction/forecasting

1. Introduction

Numerical weather prediction of convective hazards has steadily improved in recent years with increasing computing power, allowing for finer horizontal resolution, more sophisticated model physics, improved data assimilation, and an increasing forecast ensemble size. As part of the Unified Forecast System vision, the Rapid Refresh Forecast System (RRFS) is expected to replace several regional operational models of the National Weather Service. It will feature the Finite-Volume Cubed-Sphere (FV3) model (e.g., Lin 2004) as its dynamic core. As part of the 2022 National Oceanic and Atmospheric Administration (NOAA) Hazardous Weather Testbed (HWT) Spring Forecasting Experiment (SFE), the Center for Analysis and Prediction of Storms (CAPS) ran convection-allowing ensemble forecasts with 3-km grid spacing utilizing multiple-physics combinations, with the goal of helping determine the optimal configuration of an operational multiphysics RRFS system.

The FV3 dynamic core's operational implementation at convection-allowing resolution is limited to two high-resolution window FV3 members in the High-Resolution Ensemble Forecast (Roberts et al. 2019), which replaced the high-resolution window Nonhydrostatic Multiscale Model on the B-Grid (Janjic 2005) members in May 2021. Its performance as a convection-allowing model has been examined in the past few years through NOAA HWT and hydrometeorology testbed experiments. Zhang et al. (2019) first demonstrated comparable precipitation forecast skill between FV3 and the Weather Research and Forecasting Model during the 2018 NOAA HWT SFE. Snook et al. (2020) noted the greater performance of localized probability-matched ensemble means compared to probability-matched ensemble means in precipitation forecasts using an FV3 ensemble. Since the FV3-Based Limited Area Model (FV3-LAM) has been evolving, including the addition of and updating to physics parameterizations, continued testing is needed. Supinie et al. (2022) noted regional precipitation biases over the contiguous United States (CONUS) attributable to microphysics schemes for FV3-LAM-based ensemble forecasts for the winter season of 2020–21 and recommended different land surface models (LSMs) depending on which fields forecasters want to prioritize for optimization. The operational High-Resolution Rapid Refresh (Benjamin et al. 2016) still generally outperforms FV3 (Gallo et al. 2021), likely owing to its long-time operational tuning and more compatible initial conditions (ICs). Although FV3-LAM-based forecasts have exhibited useful skill at convection-allowing grid spacings, there remains substantial room for improvement and optimization, especially when different physics combinations are to be used in the ensemble.

Johnson's current affiliations: Cooperative Institute for Severe and High-Impact Weather Research and Operations, University of Oklahoma, Norman, Oklahoma and NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma

Corresponding author: Marcus Johnson, marcus.johnson@ou.edu

DOI: 10.1175/WAF-D-24-0034.1

© 2025 American Meteorological Society. This published article is licensed under the terms of the default AMS reuse license. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

Brought to you by NOAA Library | Unauthenticated | Downloaded 05/29/25 05:25 PM UTC

High-resolution forecast evaluation for convection-allowing models often invokes techniques to mitigate the "double penalty" issue in which small spatial and temporal displacements can simultaneously decrease hits and increase false alarms (e.g., Ebert 2008; Gilleland et al. 2009). These techniques include the use of a neighborhood, scale separation, object-based methods, and field deformation (e.g., Gilleland et al. 2009; Schwartz 2017). Forecasts of severe weather and its associated hazards can be verified against Multi-Radar Multi-Sensor products over the United States, such as azimuthal shear and maximum estimated size of hail, or against regions of intense reflectivity (e.g., Flora et al. 2019; Johnson et al. 2023). One neighborhood approach for severe weather hazards employs "surrogate severe weather reports" (SSRs; e.g., Sobash et al. 2011). Analogous to "practically perfect" forecasts (e.g., Brooks et al. 1998; Hitchens et al. 2013), SSRs map predicted model events [i.e., updraft helicity (UH) exceeding a prescribed threshold] that can be considered "surrogates" of severe weather occurrences to a coarser grid, to which a Gaussian smoother can be applied to create a surrogate severe probability forecast (SSPF). SSPFs have demonstrated skill for both deterministic and probabilistic forecasts of convective hazards for warm- and cool-season events (e.g., Sobash et al. 2016, 2019).

This study presents evaluations of the FV3-LAM ensemble forecasts produced by CAPS during the 2022 HWT SFE, focusing specifically on skill in predicting severe convective hazards. The forecasts to be verified in this paper are a 10-member subset with perturbed ICs that is part of a larger 21-member ensemble over the CONUS. SSRs are created from simulated UH in the 0–3-km layer, UH in the 2–5-km layer, and 10-m wind speed ($|\mathbf{U}|$ but stylized as U for simplicity) and are verified against observed tornado, hail, and wind local storm reports. Similar to prior studies, probabilistic forecasts are created by applying a Gaussian filter to SSRs to create SSPFs and are verified against similarly upscaled observed storm reports (OSRs) or their smoothed observed severe probabilistic fields (OSPFs). Such a study can clarify which hazards the ensemble can forecast with skill and what deficiencies remain to be improved.

The remainder of this paper is organized as follows: Section 2 details the methods used including the forecast configuration and Model Evaluation Tools employed for verification; section 3 evaluates the performance of the individual members and ensemble forecasts; section 4 provides an operational comparison in terms of Storm Prediction Center (SPC) day 1 convective outlooks; and section 5 summarizes and discusses the results.

2. Model configuration and verification

a. *FV3-LAM ensemble configuration*

The 21-member FV3-LAM ensemble run by CAPS during the 2022 HWT SFE contains one reference member and three subensembles designed to examine the performance of different ensemble perturbation strategies. One five-member subensemble is multiphysics with identical ICs and lateral boundary conditions (LBCs), the 10-member subensemble is multiphysics with IC perturbations, and the other five-member subensemble is multiphysics with stochastic physics and IC perturbations. Forecasts are performed on a domain centered over the CONUS (Fig. 1) using ~3-km horizontal spacing with 65 vertical levels. Forecasts are initialized at 0000 UTC each weekday from 2 May to 3 June 2022 and run for 84 h. The FV3-LAM forecasts for the 12–36-h time period examined in this study span a total of 16 days between 5 May and 3 June. The FV3-LAM dynamic core employed for these forecasts was obtained from the Global Systems Laboratory (GSL) repository (https://github.com/NOAA-GSL/ufs-weather-model) on 30 March 2022.

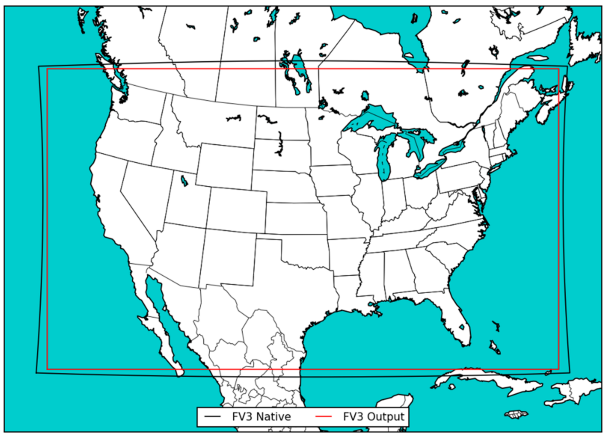

FIG. 1. Computational domains of the FV3 native model (black) and FV3 ouput (red) over the CONUS for the 2022 NOAA HWT SFE.

In this study, we consider one of the three subensembles run by CAPS during the 2022 HWT SFE: a multiphysics, perturbed IC subensemble containing 10 members. We also compare individual member performance with that of the CAPS FV3-LAM ensemble's reference member (M0B0L0_PG) and the control member of the multiphysics ensemble (M0B0L0_P; Table 1). These members are designated as such because the reference member M0B0L0_PG contains Global Forecast System (GFS) ICs and LBCs, to which IC strategies can be compared. The control member of the multiphysics ensemble M0B0L0_P uses experimental RRFS ensemble-variational data assimilation analyses as ICs and GFS forecasts as LBCs. All members with a "_PI" suffix use experimental RRFS ensemble Kalman filter perturbed ICs and Global Ensemble Forecast System (GEFS) forecast LBCs.

The naming convention for the members contains information about their configuration as follows: M0 and M1 identify the microphysics schemes, the partially two-moment Thompson aerosol-aware scheme (Thompson and Eidhammer 2014) or the fully two-moment National Severe Storms Laboratory (NSSL; Mansell et al. 2010) scheme, respectively. The B in the name identifies the planetary boundary layer (PBL) scheme used, with B0 for the Mellor–Yamada–Nakanishi–Niino (MYNN; Nakanishi and Niino 2006), B1 for the Shin–Hong (Shin and Hong 2015), and B2 for the turbulent kinetic energy–based moist eddy-diffusivity mass-flux (TKE-EDMF; Han and Bretherton 2019) scheme. MYNN surface physics

Brought to you by NOAA Library | Unauthenticated | Downloaded 05/29/25 05:25 PM UTC

TABLE 1. Physics and IC/LBC configurations of one reference member (M0B0L0_PG), one control member from the multiphysics subensemble (M0B0L0_P), and the 10-member multiphysics with IC/LBC perturbation subensemble (M#B#L#_PI) during the 2022 NOAA HWT SFE. Experiment names denote the microphysics (M), PBL scheme (B), and LSM (L). Physics options and ICs/LBCs are denoted in the text.

Experiment	Microphysics	PBL	Surface	LSM	IC/LBC
M0B0L0_PG	Thompson	MYNN	MYNN	Noah	GFS/GFS
M0B0L0_P	Thompson	MYNN	MYNN	Noah	RRFS CNTL/GFS
M0B0L0_PI	Thompson	MYNN	MYNN	Noah	RRFS01/GEFS m1
M0B1L0_PI	Thompson	Shin–Hong	GFS	Noah	RRFS02/GEFS m2
M0B2L1_PI	Thompson	TKE-EDMF	GFS	Noah-MP	RRFS03/GEFS m3
M0B0L1_PI	Thompson	MYNN	GFS	Noah-MP	RRFS04/GEFS m4
M0B2L2_PI	Thompson	TKE-EDMF	GFS	RUC	RRFS05/GEFS m5
M1B0L0_PI	NSSL	MYNN	MYNN	Noah	RRFS06/GEFS m6
M1B1L0_PI	NSSL	Shin–Hong	GFS	Noah	RRFS07/GEFS m7
M1B2L1_PI	NSSL	TKE-EDMF	GFS	Noah-MP	RRFS08/GEFS m8
M1B0L1_PI	NSSL	MYNN	GFS	Noah-MP	RRFS09/GEFS m9
M1B2L2_PI	NSSL	TKE-EDMF	GFS	RUC	RRFS10/GEFS m10

(Olson et al. 2021) are coupled with the MYNN PBL (but not vice versa), while GFS surface physics (Long 1986) are also employed. The L in the name refers to the LSM with L0 denoting the Noah (Chen and Dudhia 2001), L1 being the Noah-MP (Niu et al. 2011), and L2 denoting the Rapid Update Cycle (RUC; Smirnova et al. 2016) model. All members use the Rapid Radiative Transfer Model for general circulation models (e.g., Clough et al. 2005) to parameterize radiative heat fluxes. Overall, two microphysics and three PBL schemes, and three LSMs are used. These are candidate schemes that have the potential to be adopted by the operational RRFS.

b. *Model Evaluation Tools and verification datasets*

In this study, we evaluate the individual members and ensemble consensus products from the CAPS FV3-LAM ensemble in terms of their ability to accurately predict severe weather occurrences using three simulated severe weather proxy variables: 0–3-km UH, 2–5-km UH, and 10-m wind speed (the magnitude of the 2D wind vector $|\mathbf{U}|$ but stylized as U for simplicity). A threshold is applied to each of these fields during verification to produce SSRs and SSPFs via a process described later in this section. The method to generate SSRs and SSPFs is similar to those employed by Sobash et al. (2016, 2019) and uses Model Evaluation Tools v11.0 (Brown et al. 2021) to format forecast/observation fields and perform verification. Additionally, a binary filter is applied to 10-m U using an hourly maximum 1-km height reflectivity threshold of 45 dBZ and a square neighborhood with a 40-km length. These values are chosen to isolate 10-m U to thunderstorm and convective events (on which the SPC wind reports are conditioned) while producing forecasts with skill relative to the fractions skill score (FSS; Roberts and Lean 2008) (Fig. 2). Here, skill represents the FSS if the SSPF at every grid point was equal to the observed fractional coverage f_0, given by the formula $0.5 + (f_0/2)$. At each grid point, if 1-km height reflectivity exceeds this threshold anywhere inside the neighborhood, then the 10-m U value is retained; otherwise, 10-m U is set to 0.

CAPS FV3-LAM hourly maximum forecasts of the severe weather proxies are combined into a single dataset that contains the maximum value at each grid point over the 12–36-h forecast period using Model Evaluation Tool's "pcp_combine." This temporal range is chosen as it corresponds with the 1200–1200 UTC timing used by SPC local storm reports and the day 1 convective outlook issued by 0600 UTC. Next, the 3-km gridded output is upscaled to the National Centers for Environmental Prediction Grid 211, which uses a horizontal grid spacing of approximately 80 km. Grid 211's horizontal resolution is consistent with SPC forecasts, which define probabilities for occurrences within 25 mi (~40 km) of a point, and it has been done in prior severe weather forecast validation studies (e.g., Clark et al. 2018; Loken et al. 2020). If the 3-km forecast fields exceed a predetermined threshold, then the closest Grid 211 grid point (determined by latitude/longitude) is set to 1. The selection of threshold values is explored in

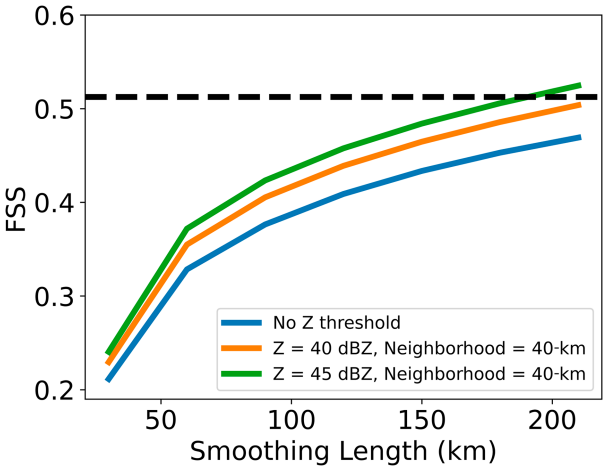

FIG. 2. FSS of 10-m U SSPFs with no filtering (blue), filtering with a reflectivity threshold of 40 dBZ and a 40-km neighborhood (orange), and filtering with a reflectivity threshold of 45 dBZ and a 40-km neighborhood (green) verified with Wind OSPFs. The skill line is shown as a black dashed line.

Brought to you by NOAA Library | Unauthenticated | Downloaded 05/29/25 05:25 PM UTC

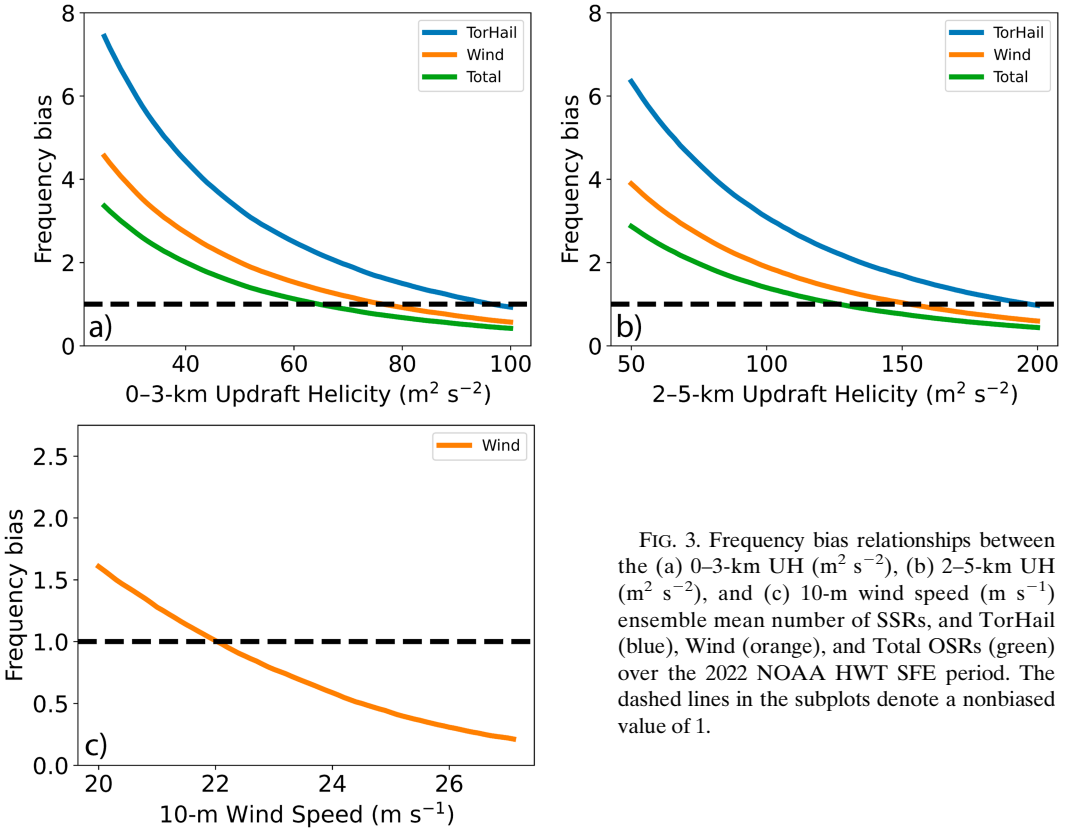

FIG. 3. Frequency bias relationships between the (a) 0–3-km UH (m^2 s^{-2}), (b) 2–5-km UH (m^2 s^{-2}), and (c) 10-m wind speed (m s^{-1}) ensemble mean number of SSRs, and TorHail (blue), Wind (orange), and Total OSRs (green) over the 2022 NOAA HWT SFE period. The dashed lines in the subplots denote a nonbiased value of 1.

section 3. The resulting binary grid of ones and zeroes represents the SSRs (1 = event, 0 = nonevent). Finally, a Gaussian smoother of varying standard deviations (smoothing lengths) is used to transform the SSRs into SSPFs. For the ensemble SSPFs, the ensemble mean of SSRs is calculated using Model Evaluation Tool's "GenEnsProd," after which the Gaussian smoother is applied.

SSRs and SSPFs are verified against OSRs and OSPFs. Similar to SSRs, OSRs are constructed by setting the nearest grid point on Grid 211 to 1 for each SPC-filtered local storm report. Separate OSRs are made for reports containing tornadoes or hail ("TorHail"), wind reports ("Wind"), and a combination of all three report types ("Total"). These reports are processed by the SPC to eliminate duplicate occurrences using spatial and temporal filters. Then, a Gaussian smoother can be applied to transform the OSRs into OSPFs. Local storm reports can be biased, particularly in sparsely populated areas where severe weather may go unreported (e.g., Anderson et al. 2007). However, local storm reports still represent the best available dataset for observed severe weather. Further, the storm reports used for verification during the 2022 NOAA HWT SFE period qualitatively exhibit at most weak correlation with the location of population centers (not shown).

Deterministic and probabilistic verifications are performed using Model Evaluation Tool's "grid_stat" for forecasts and observations on Grid 211. Because the 3-km FV3-LAM model output is already coarsened to the ~80-km Grid 211, no neighborhood is applied during verification. Verification is limited to the land area of the CONUS, as local storm reports are not typically available over the ocean. Finally, Model Evaluation Tool's "aggregate" and "aggregate_stat" are applied to compute contingency tables and verification metrics for the entire 2022 NOAA HWT SFE period for which forecasts are available. FSS is calculated external to Model Evaluation Tools directly from SSPFs and OSPFs.

3. Point-based verification

Given that SSRs are entirely dependent on the thresholds chosen, we first examine biases in reference to TorHail, Wind, and Total OSRs across a range of model variable thresholds (Fig. 3) over the 2022 HWT SFE period. The biases analyzed in this paper are largely frequency biases, which can range from 0 to infinity, with a value of 1 representing a nonbiased forecast. The 0–3- and 2–5-km UH SSRs are very similar: Frequency biases across the three types of OSRs are large at smaller thresholds, gradually decreasing to below 1.0 for large UH values. Frequency bias is highest for TorHail, followed by Wind, and finally Total OSRs, due to the frequency of these OSRs (Table 2). The frequency biases are expected as severe wind events are typically more prevalent than severe hail and tornado events. As such, the optimal threshold for verifying SSRs, which we define as the threshold which results in a frequency bias of 1.0, against TorHail OSRs is much larger than for Wind and Total OSRs. While UH is related to a storm's vertical motion and rotation (which can indicate a storm with

Brought to you by NOAA Library | Unauthenticated | Downloaded 05/29/25 05:25 PM UTC

TABLE 2. Thresholds resulting in a frequency bias closest to 1 between the UH and 10-m *U* ensemble mean number of SSRs and TorHail, Wind, and Total OSRs over the 2022 NOAA HWT SFE period. The number of SSRs is listed next to the model thresholds, and the number of OSRs is listed next to the storm reports.

	TorHail reports (347)	Wind reports (566)	Total reports (768)
0–3-km UH	≥97 m^2 s^{-2} (344)	≥76 m^2 s^{-2} (573)	≥65 m^2 s^{-2} (759)
2–5-km UH	≥198 m^2 s^{-2} (343)	≥152 m^2 s^{-2} (571)	≥126 m^2 s^{-2} (773)
10-m wind speed	—	≥22.0 m s^{-1} (570)	—

tornado, hail, and wind hazard potential), 10-m wind speed on its own contains no information on storm rotation or hail potential and is therefore only verified against Wind OSRs. Frequency bias in 10-m *U* SSRs moderately decreases over a small threshold range, from ~1.5 to 0.25 over a range of 7 m s^{-1}. Given that this large decrease does not occur when UH is verified against Wind OSRs, this might reflect limits in the numerical weather prediction model's ability to predict strong wind events at 3-km grid spacing, at which severe wind events (e.g., downbursts) might not be adequately resolved. It should be noted that a rotating updraft (i.e., high UH) is not necessary for a storm to produce severe winds, but in general, storms with rotating updrafts do have greater potential to produce severe winds (e.g., Gallus et al. 2008). The optimal thresholds, defined using frequency bias as discussed above, are listed in Table 2, and they are used for subsequent verification statistics presented in this study.

Ideally, the number of SSRs and OSRs would be equal everywhere (values near zero in Fig. 4). However, bias between SSRs and OSRs varies geographically. Both 0–3- and 2–5-km UH exhibit similar behaviors in regard to over/underprediction relative to OSRs: TorHail OSR correlations are weak overall, although there is a general trend for some underprediction in the upper Midwest and mid-Atlantic. Verification with Wind OSRs exhibits two strong correlations: underprediction of OSRs in the mid-Atlantic region and overprediction spanning the lower Great Plains to the upper Midwest. We note here that, in addition to the population density bias,

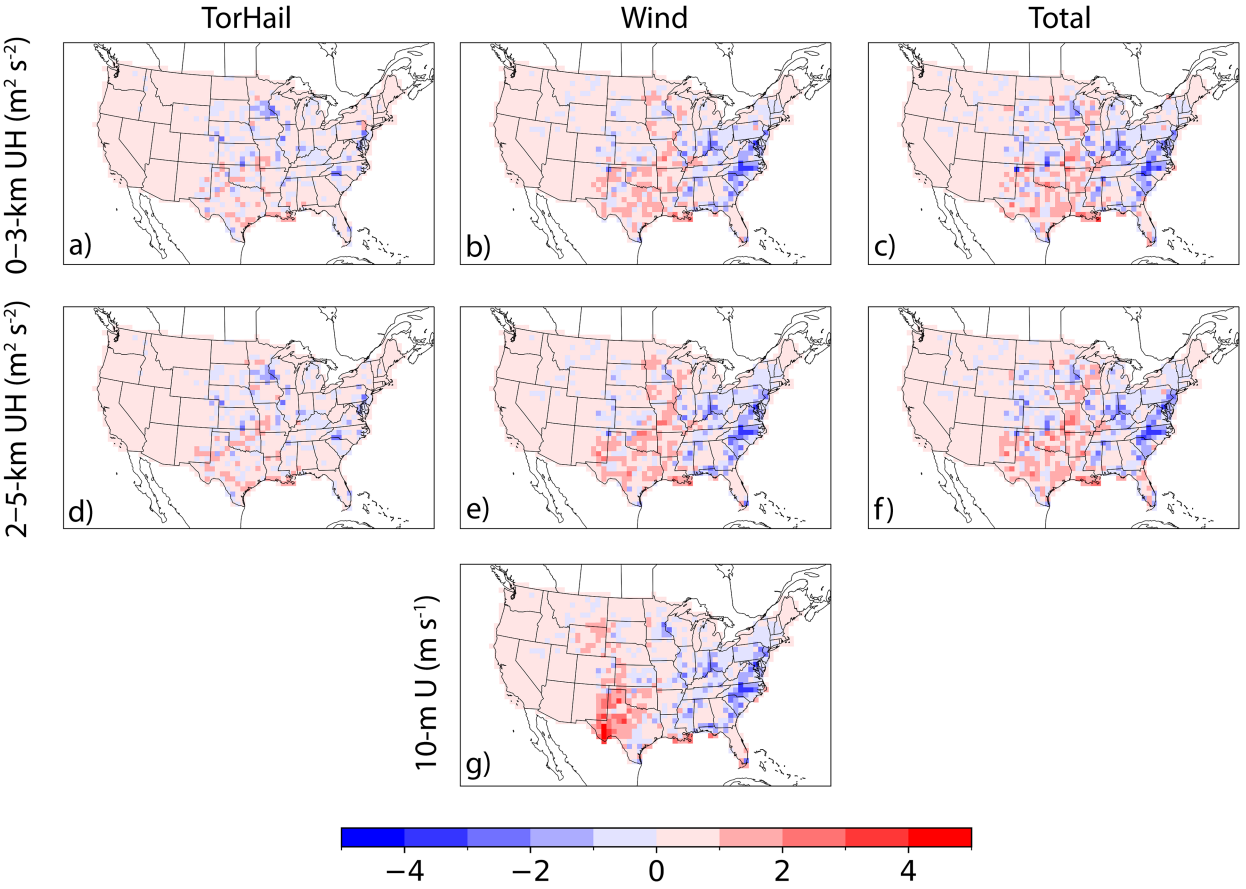

FIG. 4. The difference between the (a)–(c) 0–3-km UH (m^2 s^{-2}), (d)–(f) 2–5-km UH (m^2 s^{-2}), and (g) 10-m *U* (m s^{-1}) ensemble mean of SSRs and TorHail, Wind, and Total OSRs.

Brought to you by NOAA Library | Unauthenticated | Downloaded 05/29/25 05:25 PM UTC

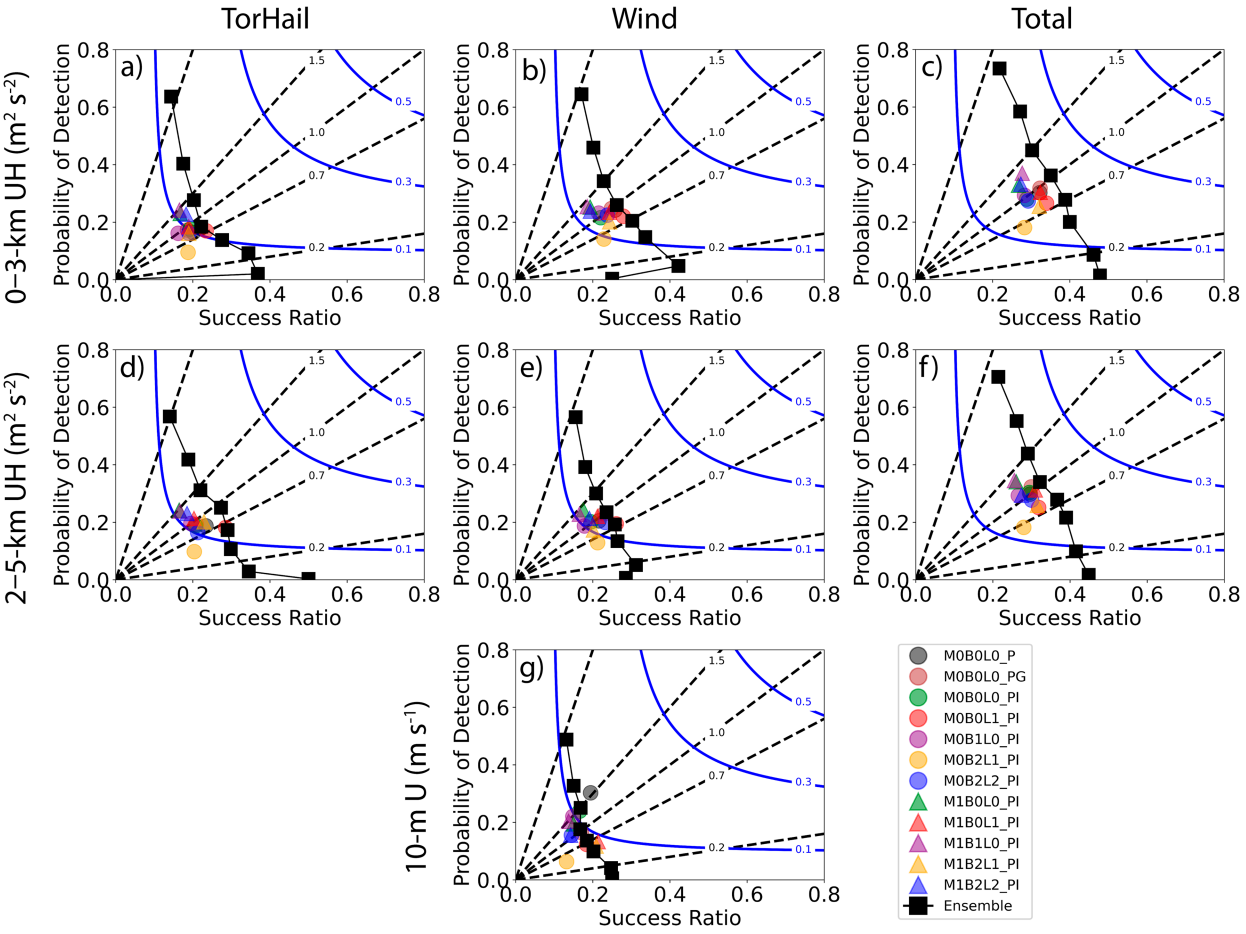

FIG. 5. Performance diagrams with (a)–(c) 0–3- and (d)–(f) 2–5-km UH ($m^2\ s^{-2}$) SSRs compared with TorHail, Wind, and Total OSRs and (g) 10-m U ($m\ s^{-1}$) SSRs compared to Wind OSRs. The black dashed lines represent lines of constant frequency bias, while the solid blue lines are lines of constant critical success index. Individual members are denoted by the color and marker shape in the legend. The ensemble mean number of SSRs is included in the subplots as a black line, with each square marker representing a probabilistic threshold $(0.1, 0.2, \ldots, 1)$ to define a forecast event.

the characteristics of local wind damage markers can also bias wind reports. For example, the presence and type of trees in the region, power lines, and other structures can affect the extent to which the damage is recorded. Total OSRs exhibit similar spatial patterns as for Wind OSRs, reflecting the larger number of wind reports compared to tornado and hail reports. Sobash et al. (2019) noticed a similar spatial distribution between UH SSRs and tornadic OSRs, attributing this bias to regional differences in tornadic environments. Geographic biases for 10-m U are similar to wind verification with UH SSRs, although the overprediction is offset further west.

a. *Deterministic forecasts*

Although ensemble forecasting can provide advantages over deterministic forecasts (e.g., developing an envelope of spread that contains more forecast possibilities and calculating ensemble consensus products that often outperform the best individual member), we first examine the performance (Fig. 5) of individual member (Table 1) forecasts to identify frequency biases and compare the skill among differently configured members. The configuration of the 10-member ensemble allows for comparisons focusing on the impact of microphysics (M0 vs M1 members) and PBL (e.g., M0B0L0_PI vs M0B1L0_PI) schemes, and LSM (e.g., M0B0L0_PI vs M0B0L1_PI) assuming the initial and boundary conditions are of similar quality across the members. As previously mentioned, the members are also compared to the M0B0L0_P and M0B0L0_PG deterministic forecasts that represent another subensemble's control member and the reference member, respectively. Finally, the ensemble mean of SSRs is compared to OSRs at several probabilistic thresholds ($p = 0.1, 0.2, \ldots, 1$) to illustrate the skill of the entire 10-member ensemble.

Performance diagrams (Roebber 2009) are constructed by comparing each member's SSRs with relevant OSRs in terms of probability of detection and success ratio; performance is overall similar among the microphysics, PBL, and LSM schemes (Fig. 5). When verifying UH SSRs, the member with the greatest underprediction frequency bias and lowest skill is

Brought to you by NOAA Library | Unauthenticated | Downloaded 05/29/25 05:25 PM UTC

the M0B2L1_PI member, which uses Thompson microphysics and TKE-EDMF PBL schemes, and the Noah-MP LSM. Typically, members using NSSL microphysics ("M1" members) produce relatively large frequency bias and higher skill than those with Thompson microphysics ("M0" members). Members with the MYNN PBL scheme have relatively large frequency bias and higher skill than corresponding members using the TKE-EDMF PBL scheme when other physics are identical. Likewise, RUC LSM members exhibit relatively larger frequency bias and higher skill than Noah-MP LSM members. The reference and control members are among the least-biased (i.e., frequency bias near 1) and most-skillful members. This is expected for the control member of the multiphysics subensemble, whose ICs are supposed to be more like the ensemble Kalman filter mean. It is less clear about the reference member that uses GFS analyses as ICs. Both members use GFS forecasts as the LBCs, which have higher spatial resolution than the GEFS forecasts used in the other members. The ensemble mean of SSRs (plotted for each probability from 0.1, 0.2, … , 1) is more skillful than even the two reference/control members across a range of probability thresholds, demonstrating the collective value of an ensemble. Therefore, for UH SSRs, it appears that the combination of the reference/control physics suite with the GFS LBCs can provide higher skills, but the ensemble mean is somewhat superior despite the lower skill with individual multiphysics members. The goal of having an effective ensemble of forecasts is achieved.

There is little difference in ensemble member performance when verifying 10-m U SSRs against Wind OSRs (Fig. 5g). Microphysics differences are not consistent for 10-m U SSRs. It appears that the MYNN PBL members are slightly less biased (i.e., frequency bias closer to 1) and more skillful than the TKE-EDMF PBL members. The two reference/control members are relatively skillful compared to other ensemble members, and the M0B0L0_P member (along with some of the members in the 10-member ensemble) is more skillful than the ensemble at different probability thresholds. We note that the M0B0L0_P member does contain a large frequency bias, so its larger critical success index could result from overestimating 10-m U. Overall, 10-m U is a relatively poor predictor of hazardous wind events, indicating that the ensemble struggles to directly prognose such events. Instead, it exhibits greater skill when relying on diagnostic metrics such as UH to prognose storms that can produce strong surface winds.

b. *Probabilistic forecasts*

As previously mentioned, probabilistic forecasts are constructed from the 10-member ensemble by calculating the ensemble mean of SSRs. The SSRs used to calculate the ensemble mean are constructed from the thresholds in Table 2. Then, a Gaussian smoother is applied to the ensemble mean to calculate ensemble SSPFs. We begin by examining probabilistic forecast reliability and sharpness (Fig. 6). Reliability is a measure of how forecast probabilities match the corresponding observations, and sharpness is a measure of the distribution of probabilistic forecasts (Murphy 1993). Sharp forecasts have most forecast probabilities near 0 or 1, which demonstrates the ensemble's forecasting consistency (i.e., all forecast occurrences, or lack thereof) across its members. Generally, 0–3-km UH ensemble SSPFs are most skillful when verified against Total OSRs, followed by Wind and TorHail OSRs (Figs. 6a–c). The 0–3-km UH is typically overforecast when compared to OSRs, so reliability typically increases as smoothing radius increases, as these overpredictions are smoothed to smaller probabilities. However, small sample size (i.e., OSRs from tornado and hailstorm reports) can severely decrease reliability. Given the overall rarity of these severe events, it is difficult to determine the sharpness of the forecasts as most probabilities are near zero. Larger smoothing radii remove higher-probability forecasts, reducing their sharpness. Reliability and sharpness exhibit similar trends for 2–5-km UH ensemble SSPFs (Figs. 6d–f). Ensemble SSPFs verified against Total OSRs provide the best reliability relative to Wind and TorHail OSRs. Forecast skill, in terms of reliability, generally increases as smoothing radius increases, and sharpness is difficult to determine given the rareness of the severe events. The 10-m U ensemble SSPFs indicate substantial overforecasting of wind events (Fig. 6g). Ensemble SSPFs with varying smoothing radii are not as stratified relative to those using UH, as the observed relative frequency is typically small across all smoothing radii. Finally, sharpness is similar to that of UH ensemble SSPFs: Probabilities are smoothed to smaller values as smoothing radius increases, but most probabilities are near the small climatological mean. A small climatological mean can make it more difficult to ascertain reliability (especially its sensitivity to smoothing), sharpness, and other forecast metrics that use climatology as a reference for skill. We attempt to mitigate these issues by, for example, selecting model thresholds that result in frequency biases near 1 and supplementing metrics with those that do not consider climatology (e.g., relative operating characteristic curves).

Relative operating characteristic (Mason 1982) curves measure an ensemble's ability to discriminate between observed events and nonevents by comparing false alarm rate and probability of detection. Forecast skill is often expressed using the area under the relative operating characteristic curve (AUC); a perfect forecast has an AUC of 1.0, while a forecast with an AUC of 0.5 is no better than random chance at distinguishing events from nonevents (i.e., false alarm rate = probability of detection). In terms of AUC, 0–3-km UH ensemble SSPFs verified against Wind OSRs are slightly more skillful than those verified against TorHail OSRs at larger smoothing radii, but forecasts show moderate skill (AUC of 0.6–0.8) when verified with both OSRs (Figs. 7a,b). AUC for 0–3-km UH SSPFs is largest when verified against Total OSRs (Fig. 7c) and with a smoothing radius of 90 km across all OSRs. The 2–5-km UH ensemble SSPFs exhibit more skill when verified with TorHail OSRs than with Wind OSRs, implying that low-level UH is a better diagnostic for severe wind events than the midlevel mesocyclone/rotating updraft (Figs. 7d,e). Still, forecasts are moderately skillful with both verifications. Like 0–3-km UH ensemble SSPFs, 2–5-km UH ensemble SSPFs are more skillful for predicting Total OSRs than for TorHail or Wind

Brought to you by NOAA Library | Unauthenticated | Downloaded 05/29/25 05:25 PM UTC

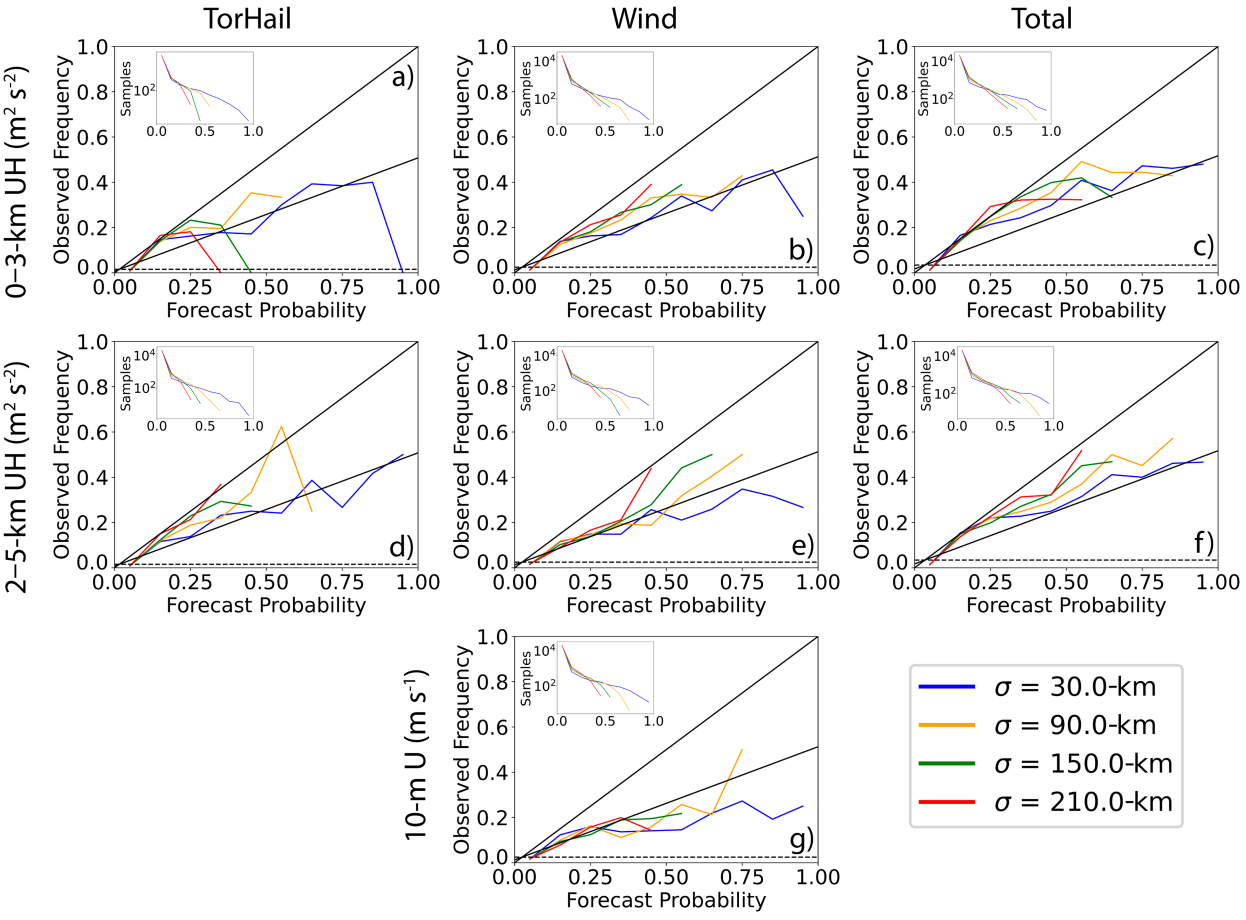

FIG. 6. Reliability and frequency diagrams using (a)–(c) 0–3-km UH, (d)–(f) 2–5-km UH, and (g) 10-m U ensemble SSPFs. Ensemble SSPFs are verified with TorHail, Wind, and Total OSRs, and increasing smoothing radii are denoted by line color. The horizontal dashed line in each reliability diagram represents the climatological mean of forecasts.

OSRs considered separately (Fig. 7f). Further, a 90-km smoothing radius generates the largest AUC when verifying 2–5-km UH, either outright or combined with another radius. The 10-m U ensemble SSPFs are less skillful than 0–3- or 2–5-km UH ensemble SSPFs when verified with Wind OSRs in terms of AUC (Fig. 7g). This suggests that even though 10-m U is a model field more directly related to surface wind hazards, deficiencies in model predictions of 10-m U result in poorer ability to discriminate between events and nonevents than predicted low-/midlevel UH.

The Brier score and Brier skill score are useful and commonly used metrics for probabilistic forecast verification. The Brier score can be decomposed into three terms: reliability, resolution, and uncertainty (Murphy 1973). The Brier skill score normalizes the Brier score using a reference forecast (e.g., climatology; Wilks 1995). Brier score components and Brier skill scores are presented in Table 3 for UH SSPFs verified against Total OSRs and for 10-m U SSPFs verified against Wind OSRs. Scores are shown for varying smoothing radii, with the optimal radius (in terms of Brier score metrics) in bold. For ensemble SSPFs, the reliability component of the Brier score is minimized (i.e., higher reliability) at large smoothing radii (180 or 210 km), in agreement with the prior reliability diagrams (Fig. 6). Even when the smoothing radius is extended to 300 km, the reliability component for 2–5-km UH and 10-m U SSPFs do not reach a minimum (not shown). The resolution component of the Brier score (which measures the ability of a forecast to resolve different event frequencies) is maximized at $\sigma = 60$ km for all SSPFs. AUC is maximized at a similar scale ($\sigma = 90$ km; see Fig. 7). Brier skill score is maximized for σ of around 90–120 km for UH ensemble SSPFs and 240 km (not shown) for 10-m U SSPFs. The 10-m U ensemble SSPF is such a poor indicator of Wind OSRs that its Brier skill score is negative for all smoothing radii examined. Overall, no one smoothing radius is optimal across the board; rather, these results indicate that different desired forecast qualities are optimized using different spatial lengths for smoothing.

The FSS (Roberts and Lean 2008) is a measure of how well the spatial distribution of forecast events matches that of observed events. FSS is calculated here by verifying 0–3- and 2–5-km UH and 10-m U SSPFs against OSPFs at different smoothing lengths. Forecasts are considered skillful if FSS exceeds $0.5 + (f_0/2)$, where f_0 is the observed fractional coverage.

Brought to you by NOAA Library | Unauthenticated | Downloaded 05/29/25 05:25 PM UTC

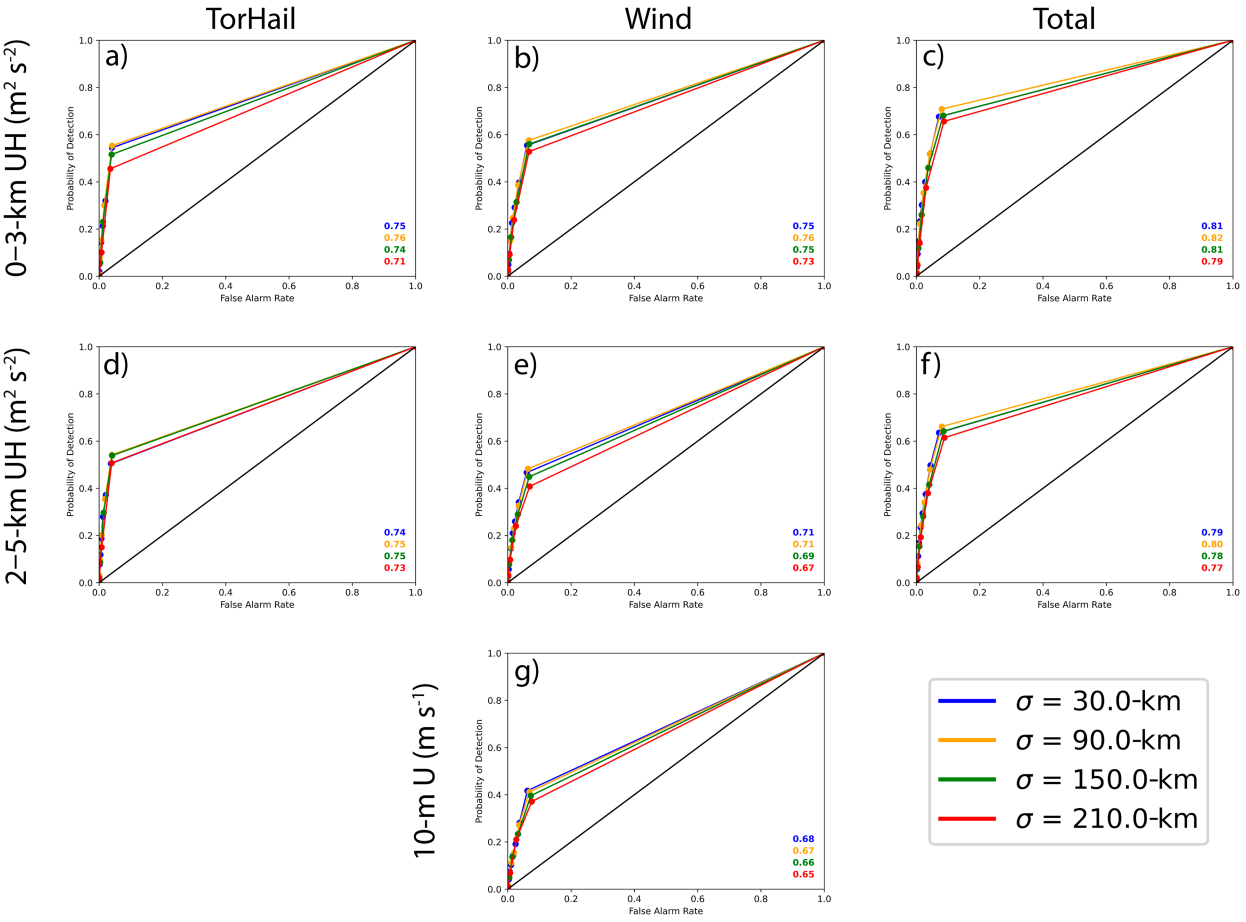

FIG. 7. Relative operating characteristic curves using (a)–(c) 0–3-km UH, (d)–(f) 2–5-km UH, and (g) 10-m *U* ensemble SSPFs. The ensemble SSPFs are verified with TorHail, Wind, and Total OSRs. Smoothing radii are denoted by line color, while each subplot also contains the AUC.

This equation is used to compute the skill lines in Fig. 8. As FSS generally increases monotonically with increasing smoothing length, the smallest length scale at which the skill threshold is exceeded can be considered the minimum skillful scale for that forecast in terms of FSS. The 0–3-km UH SSPFs indicate skillful forecasts compared to TorHail, Wind, and Total OSPFs at small smoothing lengths (e.g., ~40–60 km; Fig. 8a). FSS is highest when verified against Total OSPFs. Similarly, 2–5-km UH SSPFs produce skillful forecasts, although Wind OSPFs have a larger minimum skillful scale (Fig. 8b). In fact, UH SSPFs verified with Wind OSPFs produce the smallest FSS relative to verification with TorHail and Total OSPFs. At larger smoothing radii, FSS using TorHail OSPFs exceeds that of Total OSPFs for 2–5-km

TABLE 3. Brier score components (BS_{rel}: reliability; BS_{res}: resolution) and Brier skill scores relative to spatial smoothing lengths for 0–3-km UH, 2–5-km UH, and 10-m *U* ensemble SSPFs. UH SSPFs are verified with Total OSRs, while 10-m *U* SSPFs are verified with Wind OSRs. The uncertainty component of the Brier score is noted after each model variable, and optimal scores are in bold.

	0–3-km UH (0.032 91)			2–5-km UH (0.032 91)			10-m *U* (0.024 48)		
σ (km)	BS_{rel}	BS_{res}	BSS	BS_{rel}	BS_{res}	BSS	BS_{rel}	BS_{res}	BSS
30	0.003 04	0.005 61	0.077 80	0.003 70	0.004 93	0.037 42	0.005 15	0.001 23	−0.160 04
60	0.002 20	**0.005 85**	0.110 91	0.002 65	**0.005 13**	0.075 43	0.003 81	**0.001 28**	−0.103 50
90	0.001 82	0.005 70	**0.117 62**	0.002 13	0.004 92	0.085 03	0.003 18	0.001 16	−0.082 48
120	0.001 64	0.005 35	0.112 58	0.001 81	0.004 63	**0.085 54**	0.002 62	0.001 07	−0.063 38
150	0.001 49	0.004 96	0.105 49	0.001 58	0.004 22	0.080 36	0.002 23	0.000 96	−0.051 74
180	**0.001 40**	0.004 66	0.099 13	0.001 48	0.004 02	0.077 32	0.001 98	0.000 93	−0.042 86
210	0.001 40	0.004 30	0.088 13	**0.001 24**	0.003 77	0.076 94	**0.001 74**	0.000 85	**−0.036 34**

Brought to you by NOAA Library | Unauthenticated | Downloaded 05/29/25 05:25 PM UTC

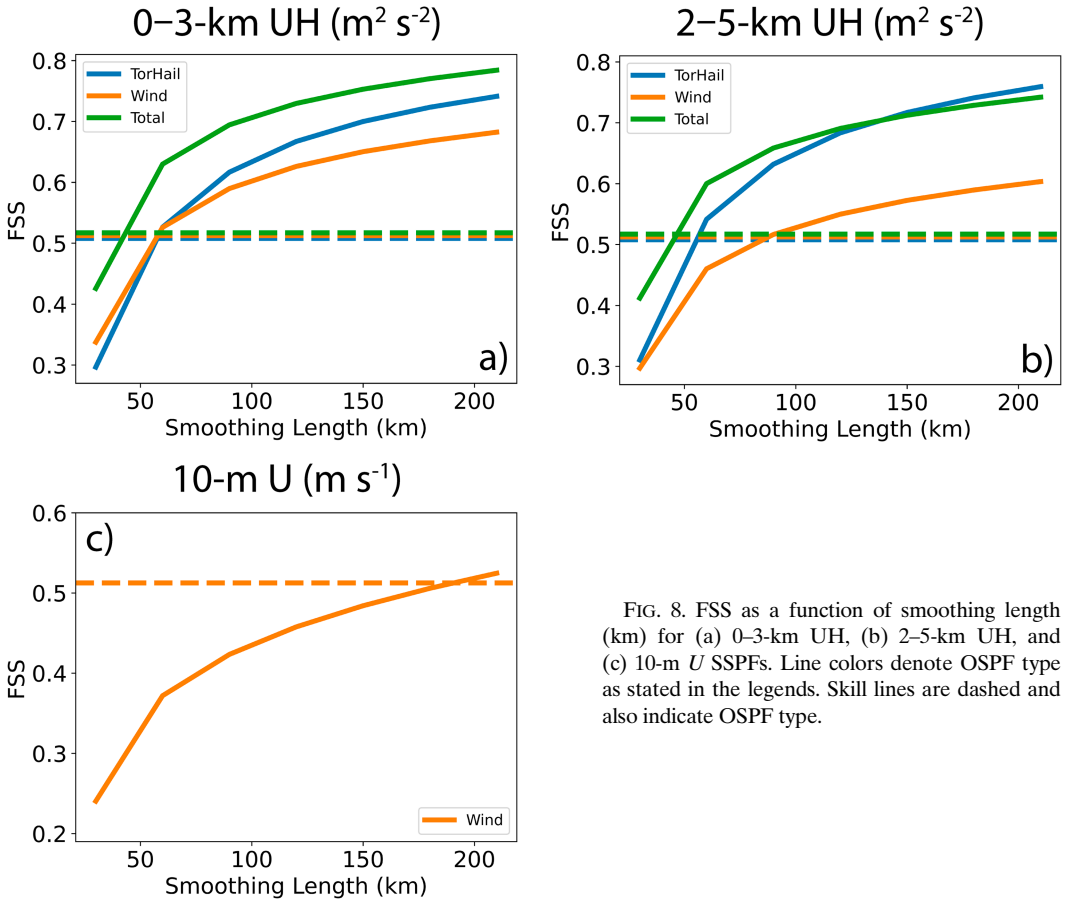

FIG. 8. FSS as a function of smoothing length (km) for (a) 0–3-km UH, (b) 2–5-km UH, and (c) 10-m U SSPFs. Line colors denote OSPF type as stated in the legends. Skill lines are dashed and also indicate OSPF type.

UH SSPFs. While less skillful than UH SSPFs (consistent with previous metrics), 10-m U SSPFs do produce skillful forecasts in terms of FSS for smoothing lengths near 200 km when verified with Wind OSPFs (Fig. 8c). Still, the spatial verification of severe wind might be problematic given the potential for events without deep convection (e.g., low UH) or the difficulty in directly resolving high-speed wind events (e.g., wind speed prognosis).

4. Comparisons with SPC day 1 convective outlooks

The 10-member CAPS FV3-LAM ensemble forecasts are compared with SPC convective outlooks to determine their consistency with operational guidance. As in section 3, ensemble forecasts of 0–3-km UH, 2–5-km UH, and 10-m U are utilized to create SSRs for each member, from which the ensemble mean can be calculated. A Gaussian smoother with $\sigma = 30$ km is applied to the SSRs to create SSPFs. This small smoothing length is chosen to allow for limited spatial errors while retaining smaller-scale structures. SPC day 1 0600 UTC convective outlooks (valid for the same 1200–1200 UTC period covered by the forecasts; e.g., Edwards et al. 2015) are overlaid on the ensemble SSPFs (Fig. 9). The FV3-LAM model output was unavailable to forecasters at 0600 UTC, so the SPC outlooks at this time were independent forecasts. Three cases during the 2022 NOAA HWT SFE period are selected to illustrate different forecast behaviors: 5, 13, and 23 May 2022.

On 5 May 2022, the SPC issued an enhanced risk threat over the ArkLaTex (Arkansas, Louisiana, Texas) region due to a 30%–45% chance of severe wind. Deep convection was anticipated in advance of an approaching cold front (Fig. 10a), supported by upper-level divergence downstream of an upper-level trough, moderate (~2000 J kg^{-1}) mixed-layer convective available potential energy, and moderate effective (40–50 kt; 1 kt $\approx$ 0.51 m s^{-1}) shear. On 13 May, the SPC issued a slight risk area extending from the upper peninsula of Michigan to Oklahoma due to an occluding cold front (Fig. 10b) and areas of moderate (~2000 J kg^{-1}) convective available potential energy forecasted in this region, but little shear. 23 May had two areas of interest; the SPC issued a slight risk over southwest Texas and a marginal risk in the Southeast United States. Over southwest Texas, an approaching shortwave trough and forecasted dryline (not shown) were anticipated to contribute to convection, although little shear precluded more severe storms despite a predicted low-level jet. An approaching shortwave trough also supported convection in the Southeast United States downstream of a surface low (Fig. 10c) with strong shear, but weak midlevel lapse rates limited stronger deep convection.

Brought to you by NOAA Library | Unauthenticated | Downloaded 05/29/25 05:25 PM UTC

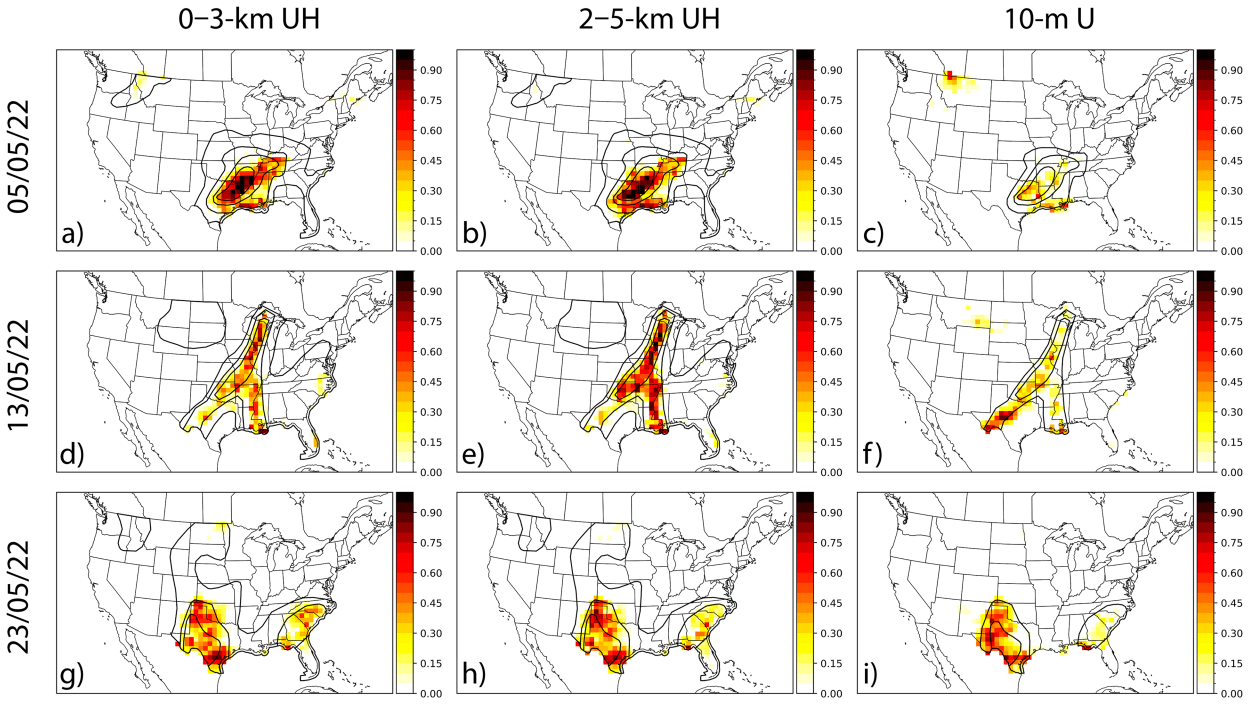

FIG. 9. The 0–3-km UH, 2–5-km UH, and 10-m *U* ensemble SSPFs for the (a)–(c) 5 May, (d)–(f) 13 May, and (g)–(i) 23 May 2022 cases. (left),(center) SPC day 1 categorical outlooks and (right) wind convective outlooks valid from 1200 to 1200 UTC are overlaid on the UH and wind SSPFs, respectively, as black contours. Convective outlook contours range from thunderstorm, marginal, slight, to enhanced risks, while wind outlook contours range from 5%, 15%, to 30% probabilities.

UH ensemble SSPFs demonstrate high confidence in severe weather on 5 May in the ArkLaTex region, with maximum SSPFs exceeding 0.9 (Figs. 9a,b). Further, the nonzero SSPFs in this area compare favorably to the marginal, slight, and enhanced risks issued in the SPC convective outlooks. Subjectively, the UH ensemble SSPFs match up well with observed storm reports, with tornado, hail, and wind reports spanning ArkLaTex into middle Tennessee (Figs. 9a,b and 11a). There are missed wind reports in southern Alabama and the Florida Panhandle and missed hail reports in north-central Oklahoma. The SPC outlooks in these regions range from no risk to marginal risk, indicating similarly low operational forecaster confidence in these areas. The 10-m *U* ensemble SSPFs (Fig. 9c) generally demonstrate low confidence in severe wind events but are typically contained within the SPC wind outlook and cover many of the observed wind reports in the ArkLaTex region (Fig. 11b).

UH ensemble SSPFs are less confident of severe weather on 13 May compared to those of 5 May, although 2–5-km UH SSPFs are slightly larger than those from 0 to 3-km UH (Figs. 9d,e). This generally agrees with the SPC's assessment; SPC only issued a day 1 slight risk on 13 May compared to a day 1 enhanced risk on 5 May. Storms on 13 May produced a large number of hail reports in or near the SPC slight-risk area (i.e., from Oklahoma into south Wisconsin; Fig. 11c), but few wind reports, which compose most of the storm reports over the 2022 HWT SFE period. The 13 May case, therefore, provides an example of the overprediction of UH SSPFs compared to Wind and Total OSRs (Figs. 4b,c,e,f). An SPC convective outlook update noted that inadequate low-level lapse rates diminished severe wind threats in Oklahoma given more stable boundary layers suppressing downbursts near the surface, although buoyancy still supported hail threats. A convective line spanning from Wisconsin to south-central Missouri spawned tornadoes and hail, so the small number of wind reports might reflect population bias as well as storm mode/environment limitations. We repeat here that wind reports can also be affected by the presence of wind damage markers. The 10-m *U* ensemble SSPFs demonstrate modest confidence of severe weather in the highest-probability SPC wind outlook area, with the most confidence in the Trans-Pecos and west-central regions of Texas (Fig. 9f). While this does indicate a false alarm when verified with wind reports in this region (Fig. 11d), the lower confidence in the Midwest United States prevents wind hazard overprediction biases relative to UH SSPFs (Fig. 4).

On 23 May, UH ensemble SSPFs generally have higher confidence of severe weather in Texas than in the Southeast United States (Figs. 9g,h), in agreement with the SPC threat assessment. UH nonzero SSPFs generally agree with the SPC convective outlook with potentially more skill, as there are a few instances of storm reports just outside the slight-risk threat area in the Texas Panhandle and southern Texas (Fig. 11e) that ensemble SSPFs forecast with confidence (greater than 0.5). Still, storm reports in the Trans-Pecos region might be more limited than in the Texas Panhandle or southern Texas, complicating

Brought to you by NOAA Library | Unauthenticated | Downloaded 05/29/25 05:25 PM UTC

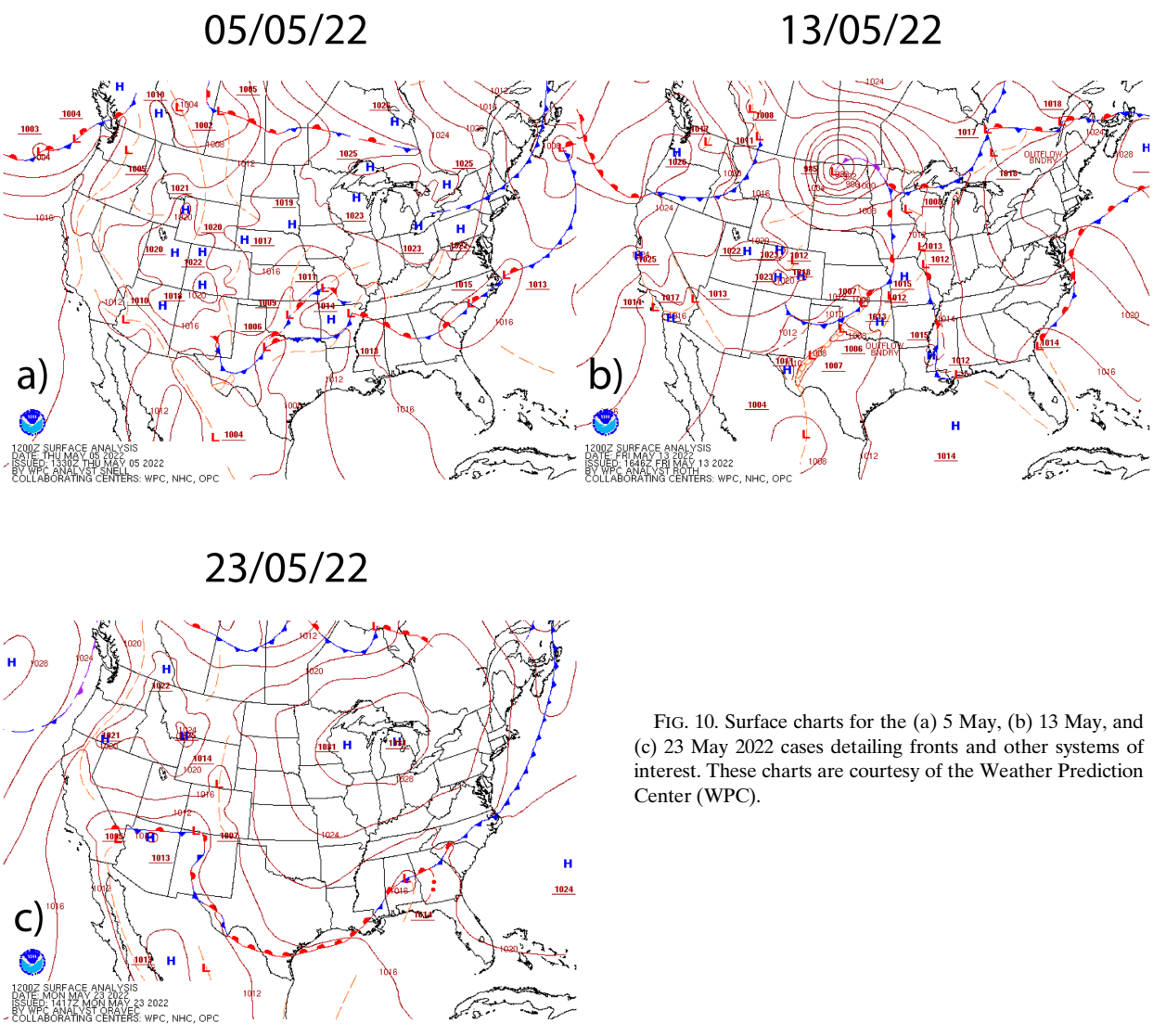

FIG. 10. Surface charts for the (a) 5 May, (b) 13 May, and (c) 23 May 2022 cases detailing fronts and other systems of interest. These charts are courtesy of the Weather Prediction Center (WPC).

verification in this area. Of particular interest on 23 May is the large number of wind reports received over the southeastern United States (Fig. 11f), despite the low confidence of UH SSPFs and the marginal-risk threat from the SPC convective outlook. Subsequent updates to the SPC day 1 convective outlook upgraded the risk threat from marginal to slight in this area (not shown) as subsequent observations reduced uncertainty regarding the amount of destabilization that would occur in the region given preexisting cloud coverage. We also note that the storm reports themselves might be affected by population bias, local wind damage indicators, and/or differences in tree/vegetation type that have different tolerances for wind damage (e.g., subsevere winds). Therefore, a single wind threshold that does not account for these factors may lead to regional verification biases. Finally, 10-m *U* SSPFs are very small in the southeastern United States, which is not entirely unexpected as the SPC wind outlook in this region has a probability of only 5% (Fig. 9i). Given how the event unfolded (Figs. 11e,f), it is clear that this event is an example of underprediction of UH and 10-m *U* ensemble SSPFs compared to OSRs in this region (Figs. 4b,e,g). The 10-m *U* ensemble SSPFs contain a similar coverage as UH SSPFs from the Texas Panhandle to southern Texas and qualitatively match observed wind reports with potentially more skill than the SPC wind outlook. Therefore, the qualitative skill of 10-m *U* ensemble SSPFs is not as deficient for this case as the 5 and 13 May cases.

5. Summary and discussion

In this study, we evaluate deterministic and ensemble convection-allowing forecasts of the CAPS FV3-LAM 10-member ensemble which was run during the 2022 NOAA HWT SFE period to evaluate its skill in forecasting severe weather hazards from 1200 to 1200 UTC (i.e., 12–36 h of forecast time). Hazards are predicted using model forecasts of 0–3- and 2–5-km updraft helicity (UH) and 10-m wind speed *U*. Instances of these model variables exceeding thresholds

Brought to you by NOAA Library | Unauthenticated | Downloaded 05/29/25 05:25 PM UTC

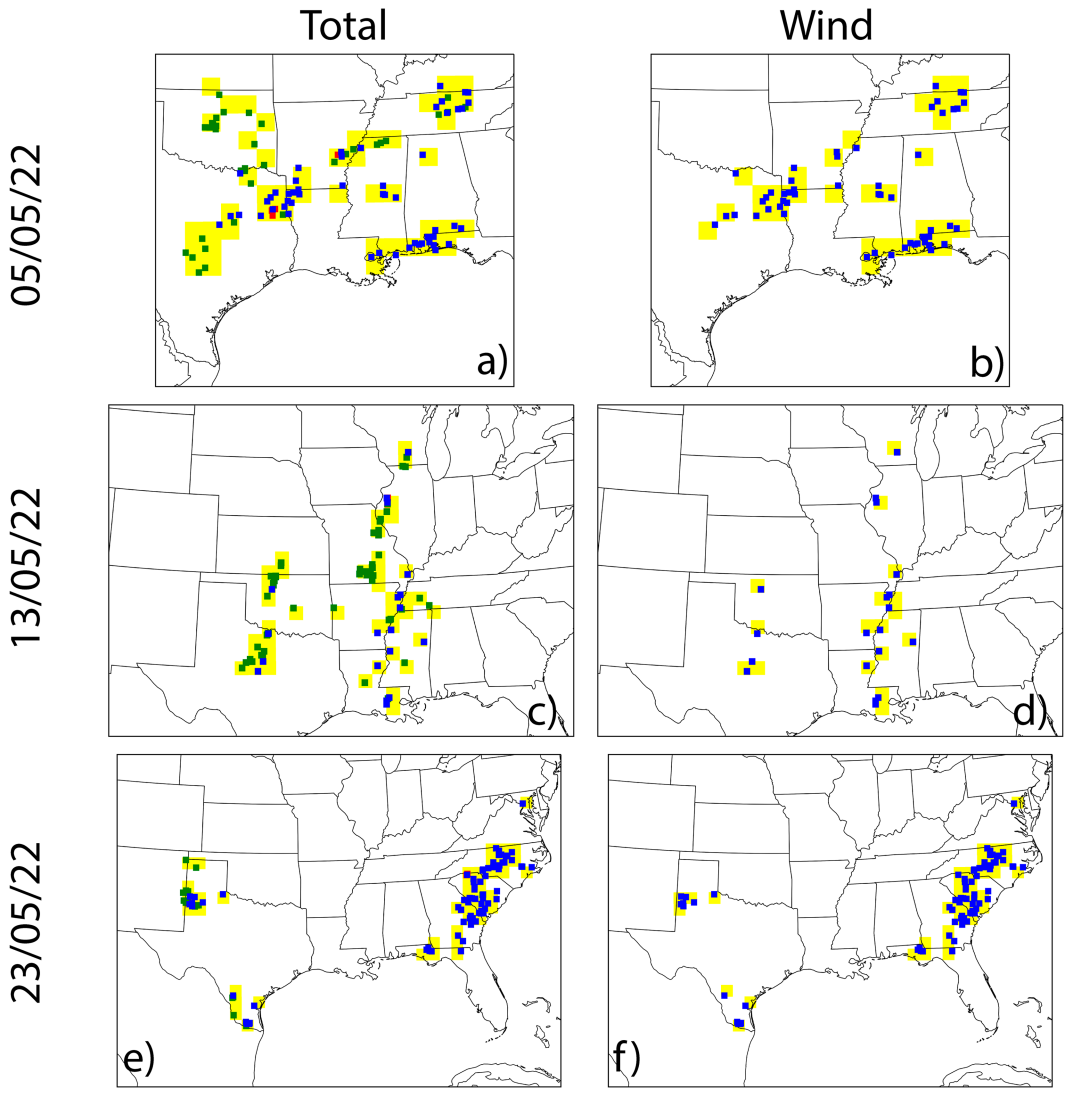

FIG. 11. OSRs (yellow boxes) from storm reports valid from 1200 to 1200 UTC for the (a),(b) 5 May, (c),(d) 13 May, and (e),(f) 23 May 2022 cases. Overlaid on the OSRs are tornado (red), hail (green), and windstorm (blue) reports.

(selected during preliminary testing to minimize overall frequency bias) are upscaled to the National Centers for Environmental Prediction Grid 211 with ~80-km grid spacing to create surrogate severe weather reports (SSRs). Similarly, observed storm reports (OSRs) are created from tornado and hail combined ("TorHail"), wind ("Wind"), and all three types ("Total") of local storm reports. Ensemble probabilistic forecasts and their observed analog are generated by applying a Gaussian smoother to the SSRs and OSRs, creating surrogate severe probability forecasts (SSPFs) and observed severe probabilistic fields (OSPFs), respectively. In addition to OSR/OSPF verification, SSPFs are compared to SPC day 1 convective outlooks issued by 0600 UTC and valid from 1200 to 1200 UTC to provide a point of comparison to an operational forecast made by skilled severe weather forecast specialists.

When nonbiased thresholds (i.e., resulting in a similar ensemble mean number of SSRs and number of OSRs over the 2022 NOAA HWT SFE period) are utilized to define SSRs, geographic biases when utilizing UH arise relative to Wind and Total OSRs. There tends to be an overprediction of UH SSRs spanning the southern Great Plains to the upper Midwest and an underprediction in the mid-Atlantic/southeastern United States. These geographic biases are similar for 10-m U SSRs when verified with Wind OSRs, with the overprediction further offset west. UH SSR verification with TorHail OSRs does not exhibit any strong geographic biases. Cases from 13 to 23 May indicate that these differences might be attributable to, for example, uncertainty in the evolution of convective environment or reporting biases. In other words, there are no clear biases solely attributable to regional convective patterns.

Skill differences among ensemble members with differing physics are small. In general, members with NSSL microphysics, MYNN PBL, and RUC LSM contain relatively higher frequency bias and skill than members with Thompson

Brought to you by NOAA Library | Unauthenticated | Downloaded 05/29/25 05:25 PM UTC

microphysics, TKE-EDMF PBL, and Noah-MP LSM members, respectively, when considering the skill of forecasts constructed using UH. Still, the ensemble means of SSRs are generally more skillful than the forecasts of individual members. Individual member forecasts constructed using 10-m U show very little consistency between physics configurations, at least in part due to their poor overall forecast skill. Attributes diagrams demonstrate that reliability for UH SSPFs is greatest for the prediction of all severe hazards compared to wind or tornado/hail hazards; overall reliability is also higher for larger spatial smoothing lengths. Similarly, these ensemble SSPFs show greater discrimination between observed and nonobserved events when verifying against all types of severe hazards. Brier score components and skill scores elaborate on these forecast characteristics and indicate that reliability is optimized at larger length scales (~180–210 km), while resolution and overall skill are optimized at smaller scales (~60–120 km). Sharpness is especially difficult to ascertain given the rareness of these severe events. Fractions skill score is typically largest when verified with all types of severe hazards, although 2–5-km UH SSPFs produce larger FSS when verified with tornado/hail hazards at large smoothing lengths. In contrast to low- and midlevel UH, 10-m U as a forecast proxy for severe wind events typically produces low-skill forecasts with few desirable qualities.

Generally, UH SSPFs using a small smoothing radius (σ = 30 km) qualitatively agree well with SPC day 1 (1200–1200 UTC) convective outlooks issued by 0600 UTC. For two of the three cases examined in detail (5 and 13 May 2022), larger SSPFs are clustered near the highest SPC risk areas (enhanced in the ArkLaTex region and slight in the Great Plains/Midwest regions, respectively), with higher probabilities in the enhanced risk on 5 May than in the slight risk on 13 May. Although there is inconsistent ensemble confidence in terms of UH SSPFs relative to the largest SPC convective outlook threat level on 23 May, observed storm reports of any severe hazard are concentrated closer to high SSPFs in the Texas Panhandle and southern Texas. Both UH SSPFs and the day 1 convective outlook underforecast a primarily severe wind event on this date in the Southeast United States. The 10-m U SSPFs display a similar spatial pattern as UH for this case near the highest wind outlook probabilities and have some consistency with the SPC day 1 wind outlook. Of the three cases, it seems the 10-m U SSPFs are most helpful for issuing threat levels corresponding to SPC wind outlook probabilities for the 23 May case.

Severe wind events may be diagnosed via updraft helicity interpretation of storm evolution, but direct prognosis (i.e., simulated wind speed) of such events is less accurate for the forecasts evaluated in this paper. There may be several reasons for this: The horizontal resolution of this regional model is 3 km. This resolution might not be fine enough to resolve small-scale wind events, such as downbursts. UH might not require as fine of a grid resolution to produce useful forecast skill, since this quantity represents the storm's overall potential for severe weather (i.e., rotating updraft) during subsequent evolution. As such, this indirect proxy remains popular for severe weather diagnosis, especially given that tornado and hail verification at this scale can be problematic. The 10-m U might capture severe wind in the absence of deep convection (which motivated the Z threshold employed in this paper). Storm reports still suffer from underreporting in sparsely populated regions, motivating the spatial upscaling employed in this paper to address the deficiency. The NSSL's Multi-Radar Multi-Sensor system has developed postprocessing diagnoses to derive physically meaningful quantities from radar data, such as maximum hail size, azimuthal shear, etc. These data generally have higher spatiotemporal resolutions than local storm reports but were developed with their own set of assumptions and parameterizations. As such, severe weather verification should include both well documented and newer experimental datasets to attempt to fill in the gap of missing observations.

It is important to continue model evaluations as operational weather models continue to evolve, especially as regional models can offer sufficient resolution to allow the explicit prediction of convective weather. In this paper, the NSSL microphysics members typically had more skill than the Thompson members when forecasting UH (i.e., higher critical success index when verifying with similarly upscaled total storm reports), but the entire ensemble with initial condition perturbations and mixed physics generally outperformed individual members. Assessing deterministic and ensemble performance can optimize the physics membership of the ensemble, as well as inform IC perturbation strategies, both of which are often needed to increase ensemble spread. Further, such ensemble performance evaluations can help clarify the effectiveness of different forecast products in operational decision-making when issuing risks and warnings for potential weather hazards (e.g., UH vs 10-m U forecasts). Subsequent model system refinements will continue to best inform hazardous weather prediction and warning guidance.

Acknowledgments. This project was supported by UFS R2O Grant NA16OAR4320115 and NOAA/OAR/OWAC Testbed Grants NA19OAR4590141 and NA22OAR4590522. The forecasts examined in this paper were run on the Frontera supercomputer (Stanzione et al. 2020) at the Texas Advanced Computing Center (TACC) at the University of Texas at Austin. TACC is supported by the National Science Foundation. We thank three anonymous reviewers for improving the quality of this manuscript.

Data availability statement. FV3-LAM hourly maximum forecasts that are evaluated in this study are available at Harvard Dataverse via https://doi.org/10.7910/DVN/RQCRSO limited to reviewer access until publication (Johnson 2024). Storm reports are archived at the Storm Prediction Center (SPC): https://www.spc.noaa.gov/climo/reports/today.html. Day 1 convective and wind outlooks are archived at the SPC: https://www.spc.noaa.gov/archive/. Surface charts are archived at the Weather Prediction Center (WPC): https://www.wpc.ncep.noaa.gov/archives/web_pages/sfc/sfc_archive.php.

Brought to you by NOAA Library | Unauthenticated | Downloaded 05/29/25 05:25 PM UTC

REFERENCES

Anderson, C. J., C. K. Wikle, Q. Zhou, and J. A. Royle, 2007: Population influences on tornado reports in the United States. *Wea. Forecasting*, **22**, 571–579, https://doi.org/10.1175/WAF997.1.

Benjamin, S. G., J. M. Brown, and T. G. Smirnova, 2016: Explicit precipitation-type diagnosis from a model using a mixed-phase bulk cloud–precipitation microphysics parameterization. *Wea. Forecasting*, **31**, 609–619, https://doi.org/10.1175/WAF-D-15-0136.1.

Brooks, H. E., M. Kay, and J. A. Hart, 1998: Objective limits on forecasting skill of rare events. *19th Conf. on Severe Local Storms*, Minneapolis, MN, Amer. Meteor. Soc., 552–555.

Brown, B., and Coauthors, 2021: The Model Evaluation Tools (MET): More than a decade of community-supported forecast verification. *Bull. Amer. Meteor. Soc.*, **102**, E782–E807, https://doi.org/10.1175/BAMS-D-19-0093.1.

Chen, F., and J. Dudhia, 2001: Coupling an advanced land surface–hydrology model with the Penn State–NCAR MM5 modeling system. Part I: Model implementation and sensitivity. *Mon. Wea. Rev.*, **129**, 569–585, https://doi.org/10.1175/1520-0493(2001)129<0569:CAALSH>2.0.CO;2.

Clark, A. J., and Coauthors, 2018: The Community Leveraged Unified Ensemble (CLUE) in the 2016 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Bull. Amer. Meteor. Soc.*, **99**, 1433–1448, https://doi.org/10.1175/BAMS-D-16-0309.1.

Clough, S. A., M. W. Shephard, E. J. Mlawer, J. S. Delamere, M. J. Iacono, K. Cady-Pereira, S. Boukabara, and P. D. Brown, 2005: Atmospheric radiative transfer modeling: A summary of the AER codes. *J. Quant. Spectrosc. Radiat. Transfer*, **91**, 233–244, https://doi.org/10.1016/j.jqsrt.2004.05.058.

Ebert, E. E., 2008: Fuzzy verification of high-resolution gridded forecasts: A review and proposed framework. *Meteor. Appl.*, **15**, 51–64, https://doi.org/10.1002/met.25.

Edwards, R., G. W. Carbin, and S. F. Corfidi, 2015: Overview of the storm prediction center. *13th History Symp.*, Phoenix, AZ, Amer. Meteor. Soc., 1.1, https://ams.confex.com/ams/95Annual/webprogram/Paper266329.html.

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.

Gallo, B. T., and Coauthors, 2021: Exploring convection-allowing model evaluation strategies for severe local storms using the Finite-Volume Cubed-Sphere (FV3) model core. *Wea. Forecasting*, **36**, 3–19, https://doi.org/10.1175/WAF-D-20-0090.1.

Gallus, W. A., Jr., N. A. Snook, and E. V. Johnson, 2008: Spring and summer severe weather reports over the midwest as a function of convective mode: A preliminary study. *Wea. Forecasting*, **23**, 101–113, https://doi.org/10.1175/2007WAF2006120.1.

Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting*, **24**, 1416–1430, https://doi.org/10.1175/2009WAF2222269.1.

Han, J., and C. S. Bretherton, 2019: TKE-based moist eddy-diffusivity mass-flux (EDMF) parameterization for vertical turbulent mixing. *Wea. Forecasting*, **34**, 869–886, https://doi.org/10.1175/WAF-D-18-0146.1.

Hitchens, N. M., H. E. Brooks, and M. P. Kay, 2013: Objective limits on forecasting skill of rare events. *Wea. Forecasting*, **28**, 525–534, https://doi.org/10.1175/WAF-D-12-00113.1.

Janjić, Z., 2005: A unified model approach from meso to global scales. *Geophysical Research Abstracts*, Vol. 7, Abstract 05582, https://meetings.copernicus.org/www.cosis.net/abstracts/EGU05/05582/EGU05-J-05582.pdf.

Johnson, M., 2024: 2022 NOAA HWT FV3-LAM forecasts. Harvard Dataverse, https://doi.org/10.7910/DVN/RQCRSO.

——, M. Xue, and Y. Jung, 2023: Biases and skill of four two-moment bulk microphysics schemes in convection-allowing forecasts for the 2018 Hazardous Weather Testbed Spring Forecasting Experiment period. *Wea. Forecasting*, **38**, 1621–1642, https://doi.org/10.1175/WAF-D-22-0171.1.

Lin, S.-J., 2004: A "vertically Lagrangian" finite-volume dynamical core for global models. *Mon. Wea. Rev.*, **132**, 2293–2307, https://doi.org/10.1175/1520-0493(2004)132<2293:AVLFDC>2.0.CO;2.

Loken, E. D., A. J. Clark, and C. D. Karstens, 2020: Generating probabilistic next-day severe weather forecasts from convection-allowing ensembles using random forests. *Wea. Forecasting*, **35**, 1605–1631, https://doi.org/10.1175/WAF-D-19-0258.1.

Long, P. E., 1986: An economical and compatible scheme for parameterizing the stable surface layer in the medium range forecast model. NCEP Office Note 321, 24 pp., https://repository.library.noaa.gov/view/noaa/11489.

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.

Mason, I. B., 1982: A model for the assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291–303.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600, https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

——, 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293, https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

Nakanishi, M., and H. Niino, 2006: An improved Mellor–Yamada Level-3 model: Its numerical stability and application to a regional prediction of advection fog. *Bound.-Layer Meteor.*, **119**, 397–407, https://doi.org/10.1007/s10546-005-9030-8.

Niu, G.-Y., and Coauthors, 2011: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *J. Geophys. Res.*, **116**, D12109, https://doi.org/10.1029/2010JD015139.

Olson, J. B., T. Smirnova, J. S. Kenyon, D. D. Turner, J. M. Brown, W. Zheng, and B. W. Green, 2021: A description of the MYNN surface-layer scheme. NOAA Tech. Memo. OAR GSL-67, 26 pp., https://repository.library.noaa.gov/view/noaa/30605.

Roberts, B., I. L. Jirak, A. J. Clark, S. J. Weiss, and J. S. Kain, 2019: Postprocessing and visualization techniques for convection-allowing ensemble. *Bull. Amer. Meteor. Soc.*, **100**, 1245–1258, https://doi.org/10.1175/BAMS-D-18-0041.1.

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.

Brought to you by NOAA Library | Unauthenticated | Downloaded 05/29/25 05:25 PM UTC

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608, https://doi.org/10.1175/2008WAF2222159.1.

Schwartz, C. S., 2017: A comparison of methods used to populate neighborhood-based contingency tables for high-resolution forecast verification. *Wea. Forecasting*, **32**, 733–741, https://doi.org/10.1175/WAF-D-16-0187.1.

Shin, H. H., and S.-Y. Hong, 2015: Representation of the subgrid-scale turbulent transport in convective boundary layers at gray-zone resolutions. *Mon. Wea. Rev.*, **143**, 250–271, https://doi.org/10.1175/MWR-D-14-00116.1.

Smirnova, T. G., J. M. Brown, S. G. Benjamin, and J. S. Kenyon, 2016: Modifications to the Rapid Update Cycle land surface model (RUC LSM) available in the Weather Research and Forecasting (WRF) Model. *Mon. Wea. Rev.*, **144**, 1851–1865, https://doi.org/10.1175/MWR-D-15-0198.1.

Snook, N., F. Kong, A. Clark, B. Roberts, K. A. Brewster, and M. Xue, 2020: Comparison and verification of point-wise and patch-wise localized probability-matched mean algorithms for ensemble consensus precipitation forecasts. *Geophys. Res. Lett.*, **47**, e2020GL087839, https://doi.org/10.1029/2020GL087839.

Sobash, R. A., J. S. Kain, D. R. Bright, A. R. Dean, M. C. Coniglio, and S. J. Weiss, 2011: Probabilistic forecast guidance for severe thunderstorms based on the identification of extreme phenomena in convection-allowing model forecasts. *Wea. Forecasting*, **26**, 714–728, https://doi.org/10.1175/WAF-D-10-05046.1.

——, C. S. Schwartz, G. S. Romine, K. R. Fossell, and M. L. Weisman, 2016: Severe weather prediction using storm surrogates from an ensemble forecasting system. *Wea. Forecasting*, **31**, 255–271, https://doi.org/10.1175/WAF-D-15-0138.1.

——, ——, ——, and M. L. Weisman, 2019: Next-day prediction of tornadoes using convection-allowing models with 1-km horizontal grid spacing. *Wea. Forecasting*, **34**, 1117–1135, https://doi.org/10.1175/WAF-D-19-0044.1.

Stanzione, D., J. West, R. T. Evans, T. Minyard, O. Ghattas, and D. K. Panda, 2020: Frontera: The evolution of leadership computing at the National Science Foundation. *PEARC'20: Practice and Experience in Advanced Research Computing 2020: Catch the Wave*, Portland, OR, Association for Computing Machinery, 106–111, https://dl.acm.org/doi/10.1145/3311790.3396656.

Supinie, T. A., J. Park, N. Snook, X.-M. Hu, K. A. Brewster, M. Xue, and J. R. Carley, 2022: Cool-season evaluation of FV3-LAM-based CONUS-scale forecasts with physics configurations of experimental RRFS ensembles. *Mon. Wea. Rev.*, **150**, 2379–2398, https://doi.org/10.1175/MWR-D-21-0331.1.

Thompson, G., and T. Eidhammer, 2014: A study of aerosol impacts on clouds and precipitation development in a large winter cyclone. *J. Atmos. Sci.*, **71**, 3636–3658, https://doi.org/10.1175/JAS-D-13-0305.1.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction*. Academic Press, 467 pp.

Zhang, C., and Coauthors, 2019: How well does an FV3-based model predict precipitation at a convection-allowing resolution? results from CAPS forecasts for the 2018 NOAA Hazardous Weather Test Bed with different physics combinations. *Geophys. Res. Lett.*, **46**, 3523–3531, https://doi.org/10.1029/2018GL081702.

Brought to you by NOAA Library | Unauthenticated | Downloaded 05/29/25 05:25 PM UTC