Biases and Skill of Four Two-Moment Bulk Microphysics Schemes in Convection-Allowing Forecasts for the 2018 Hazardous Weather Testbed Spring Forecasting Experiment Period

MARCUS JOHNSON,[a,b] MING XUE,[a,b] AND YOUNGSUN JUNG[a,b]

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

(Manuscript received 11 October 2022, in final form 22 June 2023, accepted 23 June 2023)

ABSTRACT: A proof-of-concept systematic evaluation of convective hazards is applied to short-term (1–6 h) forecasts using the Morrison, National Severe Storms Laboratory (NSSL), Predicted Particle Properties (P3), and Thompson two-moment microphysics schemes for the 2018 NOAA Hazardous Weather Testbed Spring Forecasting Experiment (HWT SFE) period (hereafter "MORR," "NSSL," "P3," and "THOM" experiments, respectively). Four convective line cases are highlighted to elaborate on relative experiment biases/skill. Composite reflectivity and 1-h accumulated precipitation are examined to determine storm coverage/precipitation biases/skill utilizing point-based verification with a neighborhood. Simulated 1–6-km updraft helicity and observed 3–6-km azimuthal shear and MESH are examined to consider simulated rotation and hail core prediction with object-based scores. Over the full season, MORR displays little overall storm coverage bias relative to NSSL, P3, and THOM underprediction. The equitable threat score (ETS) and fractions skill score (FSS) of P3 are lower than the other experiments. P3 and THOM underpredict convective regions with intense reflectivity relative to MORR and NSSL overprediction. All experiments underpredict precipitation amounts. P3 light precipitation FSS is lower than other experiments. Rotation object verification exhibits sensitivity to microphysics experiments, as microphysics has an indirect influence on storm dynamics. While P3 has the largest hail object underprediction, all experiments grossly overpredict the number of hail objects in convective line cases despite forecast objects defined with the same product (MESH) and threshold as observations. The importance of microphysics ice parameterization and ongoing scheme updates highlight the need to apply this verification framework to optimal/updated schemes before optimizing ensemble design.

KEYWORDS: Cloud microphysics; Forecast verification/skill; Numerical weather prediction/forecasting; Short-range prediction

1. Introduction

The National Oceanic and Atmospheric Administration (NOAA) Warn-on-Forecast (WoF; Stensrud et al. 2009, 2013) research program attempts to improve warning lead times of hazardous convective weather such as tornadoes and hail through high-resolution data assimilation and short-term numerical weather prediction (NWP), rather than predominantly relying on hazardous weather observation or detection. Consistent with this vision, the Center for Analysis and Prediction of Storms (CAPS) has collaborated with the NOAA Hazardous Weather Testbed (HWT) since 2007 to test and evaluate the latest convection-allowing ensemble forecasting capabilities and to provide guidance on the optimal design of such systems by running real-time Storm Scale Ensemble Forecasts (SSEFs). Different dynamic cores, physics parameterization advances including microphysics schemes (including process complexity and the number of moments predicted), data assimilation, and initial perturbation strategies were tested and evaluated (e.g., Kong 2018). Starting from 2016, HWT provided guidance under the Community Leveraged Unified Ensemble (CLUE; Clark et al. 2018) framework to coordinate contributions from different organizations and help systematically answer ensemble design questions. Microphysics (MP) processes and parameterizations have profound impact in convective storm forecasting, but as parameterization of MP still contains large uncertainties (e.g., Johnson et al. 2016; Fan et al. 2017; Putnam et al. 2017b; Johnson et al. 2019; Morrison et al. 2020), it is important to use as skillful as possible MP schemes within convective-scale ensembles. Additionally, such systems may benefit from the use of multiple reasonably skillful MP schemes (e.g., Johnson and Wang 2017; Loken et al. 2019). The assessment of the performance of MP schemes in practical settings is therefore important.

Bulk microphysics parameterization schemes (BMPs) that predict moments of the particle size distribution (PSD; rather than the discretized PSD as in spectral bin microphysics [SBM]) are typically used in operational NWP models because of their much lower computational cost and lack of evidence that SBMs will outperform BMPs (e.g., Xue et al. 2017; Morrison et al. 2020). The steady increase in computational power has enabled the use of BMPs in real-time NWP that predict two moments instead of one moment of hydrometeor PSDs. The improvement of two-moment BMPs over their one-moment counterparts due to their PSD flexibility (particularly the unconstrained

Jung's current affiliation: Office of Science and Technology Integration, Silver Spring, Maryland.

Corresponding author: Ming Xue, mxue@ou.edu

DOI: 10.1175/WAF-D-22-0171.1

intercept parameter and its role in facilitating reasonable cold pool structure, differential sedimentation) at the convective scale is well-documented (e.g., Dawson et al. 2010; Jung et al. 2012). Specifically, several studies (e.g., Yussouf et al. 2013; Putnam et al. 2014; Wheatley et al. 2014; Putnam et al. 2017a; Wang and Wang 2017) have noted the benefits of improved prediction of mesoscale convective system (MCS) and supercell storm structures (i.e., short-range hazardous storms of interest to systems such as WoF) with the use of two-moment schemes.

Given the benefits of two-moment microphysics, operational NWP models have steadily increased microphysics complexity in this direction. Operational models at the NOAA National Centers for Environmental Prediction (NCEP) include the prior Rapid Update Cycle (RUC; Benjamin et al. 2004a,b), the Rapid Refresh (RAP; Benjamin et al. 2016b), and the High-Resolution Rapid Refresh (HRRR; Benjamin et al. 2016a). Before its replacement by the RAP, the RUC employed Reisner et al. (1998) and Thompson et al. (2004) microphysics that prognose two moments of cloud ice. Early versions of the RAP adopted the Thompson et al. (2008) BMP that included two prognostic moments of the rain category. Currently, the RAP and HRRR both employ the Thompson and Eidhammer (2014) aerosol-aware BMP, which updated the cloud water category to two prognostic moments. The High Resolution Deterministic Prediction System (HRDPS) became operational for Canada in 2014 using two-moment Milbrandt–Yau (MY2; Milbrandt and Yau 2005a,b) microphysics (Milbrandt et al. 2016), before adopting the Predicted Particle Properties (P3; Morrison and Milbrandt 2015; Milbrandt and Morrison 2016) scheme in 2018 (Jouan et al. 2020). The National Severe Storms Laboratory (NSSL) Experimental Warn-on-Forecast System for ensembles (NEWS-e) expanded microphysics complexity from the partially two-moment Thompson et al. (2008) BMP in 2016 to the fully two-moment NSSL BMP (Mansell et al. 2010) with an additional hail category in 2017 to more accurately simulate supercells (Skinner et al. 2018). While earlier CAPS SSEFs were primarily one- and partially two-moment microphysics (Xue et al. 2007, 2008), later SSEF configurations overwhelmingly used partially and full two-moment microphysics (Kong 2018).

Microphysics performance in convection-allowing models (CAMs) may be verified using different forecast evaluation techniques. Traditional, point-based verification at the convective scale can include a neighborhood relaxation given the difficulty for microphysics schemes within CAMs to prognose small-scale structures/precipitation at the exact correct time and location (e.g., Schwartz 2017). Scale-separation (e.g., Yu et al. 2020) and field deformation (e.g., Keil and Craig 2009) techniques can also mitigate detailed space and time errors. Neighborhood relaxations may include neighborhood consideration in forecasts and/or observations (e.g., Clark et al. 2010; Ben Bouallègue and Theis 2014). Another increasingly popular and promising approach for convective-scale forecast evaluation is object-based verification (Skinner et al. 2018; Flora et al. 2019), which emphasizes storm structure/feature validation (such as forecasted hail cores simulated by microphysics) while additionally allowing a more robust verification dataset than sparse local storm reports (e.g., Brooks et al. 2003; Sobash et al. 2011). User-tunable parameters intrinsic to object verification (Davis et al. 2006) allow users to filter/match objects based on their attributes (e.g., centroid distance, area) and facilitate comparison between separate forecast and observation fields.

This study aims to help improve short-range, small-scale hazardous weather forecasts by presenting a proof-of-concept systematic evaluation of convective hazards in forecasts initialized with convective-scale data assimilation that differ in microphysics only. This evaluation concept is applied to forecasts produced using four (fully or partially) two-moment BMP schemes, with configurations of the ensemble forecasts produced by CAPS during the 2018 NOAA HWT Spring Forecasting Experiment (SFE) at 3-km grid spacing over the contiguous United States (CONUS). Among the BMPs is the P3 scheme that has a nontraditional representation of ice-phase hydrometeors by not using predefined categories. This novel scheme has been evaluated for idealized and real case studies featuring orographic and stratiform precipitation, and deep convection among others (e.g., Morrison et al. 2015; Hou et al. 2020; Naeger et al. 2020; Milbrandt et al. 2021). The evaluation of hazards is performed in terms of composite radar reflectivity, precipitation, rotation feature, and hail core forecasts. Forecasts are verified using Multi-Radar Multi-Sensor (MRMS; Smith et al. 2016; Zhang et al. 2016) observations: composite reflectivity, accumulated precipitation, azimuthal shear, and maximum expected size of hail (MESH; Witt et al. 1998). Both point-based neighborhood and object-based verifications are employed to assess relative forecast skill and biases. Along with full season forecasts over the entire 2018 HWT SFE period, four convective line cases within the season are highlighted to elaborate on full season BMP relative behaviors. Such a study can help improve short-range forecasting by providing a proof-of-concept evaluation of simulated hazards, and the verification framework can be applied to forecasts using the latest version of the microphysics schemes to inform optimal ensemble design. Because some of the schemes have undergone significant modifications/enhancements, the performance of the schemes observed in this paper should not be used to make recommendations for future forecasting systems.

The rest of this paper is organized as follows. Section 2 details forecast design, the BMP schemes tested, the Model Evaluation Tools (MET) package utilized to populate contingency tables, and the four convective line cases. Section 3 examines relative microphysics skill using point-based verification with a neighborhood in the forecasts, section 4 evaluates the forecasts using simulated rotation and hail objects, and section 5 provides a summary and conclusions.

2. Forecast system design and verification

a. 2018 CAPS HWT ensemble forecasts

The forecasts to be evaluated in this paper were initialized from analyses produced using the Gridpoint Statistical Interpolation (GSI) ensemble Kalman filter (EnKF) and CAPS EnKF (e.g., Xue et al. 2006; Jung et al. 2012) data assimilation systems valid at 0000 UTC of each day. Based upon the

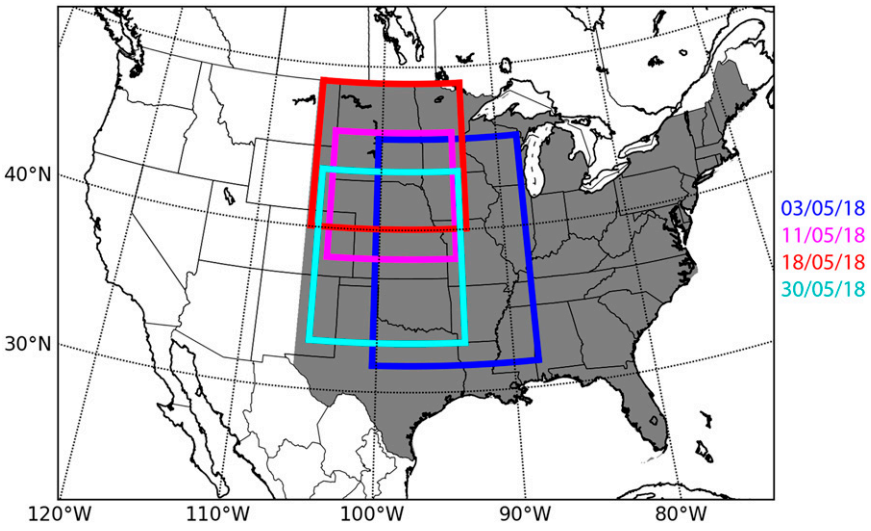

FIG. 1. The 2018 CAPS HWT SSEF computational domain over CONUS. The verification domain for the full season is shaded, while the verification domains for the four convective line cases are indicated by colored boxes.

ensemble configuration described in Labriola et al. (2021), the EnKF used 40 members. Thompson (Thompson et al. 2008) microphysics is used during model integration in the data assimilation (DA) period. The ensemble was initialized at 1800 UTC from North American Mesoscale (NAM) analysis plus perturbations derived from 3-h forecasts of the 1500 UTC cycle Short-Range Ensemble Forecast (SREF). Conventional observations were assimilated hourly between 1900 and 0000 UTC, and radar (Z and V_r) observations were assimilated every 15 min in the last hour using CAPS EnKF to arrive at the final analyses at 0000 UTC. The 10-member ensemble forecasts were run with varied microphysics and planetary boundary layer (PBL) schemes (see Table 4 in Kong 2018) at 0000 UTC from the 40-member ensemble mean and 8 EnKF member analyses. The lateral boundary conditions came from NAM and perturbations derived from SREF forecasts. The Weather Research and Forecasting (WRF) Model version 3.9.1.1 (Skamarock et al. 2008) is used at a 3-km grid spacing with 51 vertical levels over CONUS (Fig. 1). The forecasts initiated 0000 UTC Monday through 0000 UTC Friday (in GMT) from 30 April to 1 June 2018 plus 27 April. The forecasts were run overnight and available in the morning of weekdays.

Microphysics schemes used in the 10-member ensemble forecasts include the Morrison (Morrison et al. 2009), the NSSL (Mansell et al. 2010), the P3 (Morrison and Milbrandt 2015; Milbrandt and Morrison 2016), and the Thompson (Thompson et al. 2008) schemes. The PBL schemes include the Mellor–Yamada–Janjić (MYJ; Janjić 1994), Yonsei University (YSU; Hong et al. 2006), and the Mellor–Yamada–Nakanishi–Niino (MYNN; Nakanishi and Niino 2006). For the purpose of this study, forecasts over the 2018 HWT SFE period are run in post–real time for the Morrison, NSSL, P3, and Thompson schemes, using the EnKF ensemble mean initial conditions and NAM forecast lateral boundary conditions, and Noah land surface model (LSM) and MYJ PBL schemes as in members 9 and 10 (Kong 2018). These four forecasts (hereafter "MORR," "NSSL," "P3," and "THOM" experiments, respectively) therefore differ only in the BMP scheme used during the 2018 HWT SFE. We also note here that verification would likely be different at smaller horizontal grid spacings as convective updrafts might be better resolved at higher resolutions.

b. *Microphysics schemes*

The Thompson scheme examined here is partially two-moment, prognosing mass mixing ratio q and number concentration N_t for rain and cloud ice while prognosing only the mass mixing ratio q for cloud water, snow, and graupel. The Morrison scheme is partially two-moment, prognosing q and N_t for rain, cloud ice, snow, and graupel and q for cloud water. The NSSL scheme is two-moment for cloud water, rain, cloud ice, snow, graupel, and hail, and also prognoses the bulk volume of graupel and hail, allowing for the prediction of rimed ice bulk density. Finally, the P3 BMP with one ice category is unique in that it has one ice category spanning multiple ice habits. These ice habits are partitioned within the P3 scheme's PSDs by virtue of mass consistency and riming history. The scheme used here is partially two-moment, prognosing q and N_t for rain and ice and q for cloud water. The scheme additionally prognoses rime mass mixing ratio q_{rim} and rime volume mixing ratio B_{rim}. The reader is referred to Morrison and Milbrandt (2015) for more information on the configuration of the scheme used in this study; recent P3 scheme updates are discussed below. As the EnKF ensemble mean analyses obtained using Thompson microphysics during the model integration are used to initialize these experiments, initial N_t and bulk volume of rimed ice in the NSSL scheme not prognosed by the Thompson scheme are constant or diagnosed based on assumptions such as fixed intercept parameters and

ice densities. The Morrison scheme diagnoses initial N_t regardless of the Thompson scheme's prognosis, and the P3 scheme does not set or diagnose rime variables.

All microphysics schemes examined in this paper have exhibited skill simulating deep convection (e.g., Morrison et al. 2015; Johnson et al. 2016, 2019), and the P3 and Thompson schemes are used operationally as previously mentioned. Snook et al. (2019) found the Morrison member contained the highest bias in 3-h accumulated precipitation (3-h AP) exceeding 0.01 in. during the 2016 Hydrometeorology Testbed (HMT) Flash Flood and Intense Rainfall (FFaIR) experiment among a subensemble differing in microphysics only (i.e., Morrison, MY2, P3, and Thompson schemes). The members using Morrison and P3 microphysics contained high bias in 3-h AP exceeding 0.5 in., relative to low bias in the member using the Thompson scheme. Zhang et al. (2019) found that forecasts using the Thompson scheme had typically higher neighborhood equitable threat score (NETS) and fractions skill score (FSS) than forecasts using the NSSL scheme for moderate (~0.1–0.4 in. observed) hourly precipitation during the 2018 HWT SFE using the Finite-Volume Cubed-Sphere (FV3) model.

Recent P3 scheme updates (not in this paper) include multiple ice categories, which can lessen the "dilution" of properties of an existing ice distribution (Milbrandt and Morrison 2016). Furthermore, it can include prognostic liquid fraction (Cholette et al. 2019) and three prognostic moments for rain (Paukert et al. 2019) and ice (Milbrandt et al. 2021). Notably, the third prognostic moment removes the need for an ice size limiter, originally utilized to reduce over size sorting of hydrometeors, and improves reflectivity/hail, etc. We also point out here that an updated version of the Thompson scheme that additionally prognoses cloud water, and "water friendly" and "ice friendly" aerosol number (Thompson and Eidhammer 2014) is currently used operationally in the RAP and HRRR models, but not in these forecasts. Recently, the Thompson scheme added prognostic number and bulk volume to its graupel–hail hybrid category (Jensen et al. 2023), similar to the NSSL and P3 schemes. Further, the graupel category in the Morrison scheme may be modified to be more "hail-like" (i.e., bulk density and fall speed) through a WRF namelist parameter; this feature is not used here. These rimed ice updates/optimizations may improve forecasts of deep convection, so their performance should be carefully evaluated in future studies before any microphysics recommendations can be made.

c. *Model Evaluation Tools (MET) and verification datasets*

Forecast evaluation statistics are computed using the MET (Brown et al. 2021) package Version 6.0. The "grid_stat" tool of the package is employed for traditional, point-based verification as forecast model output and its corresponding observed verification data are on the same CONUS grid. As only short-term forecasts (1–6 h) are evaluated, a square neighborhood of length 18 km is utilized to populate contingency tables based on the "neighborhood maximum" approach (e.g., Sobash et al. 2011): a hit at any grid point is defined as both model and observation fields exceeding the given threshold anywhere inside the neighborhood, a false alarm is defined as the model field exceeding the given threshold anywhere inside the neighborhood but not observations, and a miss is defined as the observed field exceeding the given threshold anywhere inside the neighborhood but not the model field. Neighborhood statistics exclude data points that do not have 100% "valid" data (i.e., outside the analysis domain/masking regions). We note here that the analysis domain is limited to CONUS land east of 106°W longitude (Fig. 1) as the observed radar coverage, from which verification products are derived, is more limited over the oceans compared to forecasts.

Forecast and observed objects for verification are created using the Method for Object-Based Diagnostic Evaluation (MODE; Davis et al. 2006, 2009) tool. MODE utilizes both a convolution filter to smooth forecast and observed data and object-defining thresholds for the tool's fuzzy logic engine. The interest function for object matching employed in MODE is defined as

$$T(\alpha) = \frac{\sum_i w_i C_i(\alpha) I_i(\alpha_i)}{\sum_i w_i C_i(\alpha)}, \tag{1}$$

where α contains object attributes (e.g., object area), w_i are user-defined scalar weights, $C_i(\alpha)$ are attribute confidence maps, and I_i are user-defined interest maps. This interest function can accommodate more object attributes and flexible interest maps compared to that used in Skinner et al. (2018), but leverages the following user-defined parameters to emulate their function. Object matching only considers centroid and boundary distance attributes, both with user-defined scalar weights of 1. We define centroid and boundary distance interest maps linearly as equal to 1 at 0-km distances, and 0 at the maximum allowed distances (39 km). Further, rotation and hail forecast and observed datasets are hourly maxima. There is no temporal penalty in object matching if objects occur at any point within the hourly forecast period, which alleviates the need for a temporal component in the interest function. Objects are considered matched when total interest exceeds 0.2, and can only match with one object in the opposite field. Forecast and observed objects are merged within their respective fields if boundary distance does not exceed 12 km. Objects are filtered if their area is less than 144 km^2. Each user-defined parameter controlling object matching represents a degree of freedom, the sensitivity of which is discussed in Skinner et al. (2018) and Flora et al. (2019). It is easy to imagine how decreasing the merging distance may increase false alarms, or increasing the maximum allowed distance may increase hits. Given that this paper explores both point-based and object verification, sensitivity to MODE parameters is outside its scope.

All observation datasets are provided by the NSSL's MRMS system (e.g., Smith et al. 2016; Zhang et al. 2016). Point-based verification utilizes simulated and observed composite (i.e., column-maximum) reflectivity Z and 1-h accumulated precipitation (1-h AP). Model Z is calculated internally

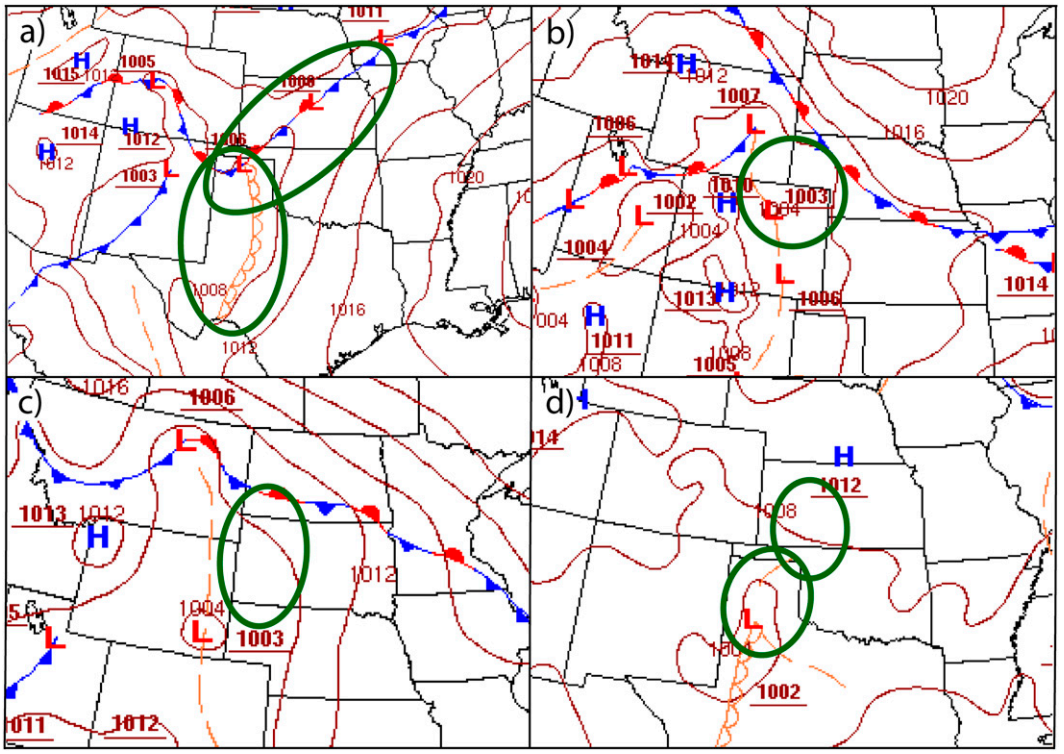

FIG. 2. Surface charts for the (a) 3 May, (b) 11 May, (c) 18 May, and (d) 30 May convective line cases valid at 1800 UTC the previous day. Synoptic and mesoscale areas of interest are denoted by green circles. Surface charts are courtesy of the Weather Prediction Center (WPC).

in microphysics code using the Rayleigh scattering approximation, and may differ among microphysics due to assumptions by the schemes (e.g., spherical versus nonspherical snow, rimed ice density). Rotation objects are verified using simulated 1–6-km updraft helicity and observed 3–6-km azimuthal wind shear, two popular rotation verification proxies (e.g., Skinner et al. 2016, 2018; Flora et al. 2019). Hail objects employ the revised MESH formulation by Murillo and Homeyer (2019), which is derived from a more robust dataset (5954 hail reports versus 147 original) and contains greater correlation between the forecast and observed MESH fields in the forecasts than the original MESH formulation (Witt et al. 1998). We note here the limitations of MESH to define hail core objects, as its calculation is reflectivity-based and therefore might be influenced by rain particles, although it is also weighted using height levels of subfreezing temperatures. We also explore the utilization of maximum hail size (MHS) derived from simulated ice PSDs in section 4a. Given the lack of observation PSD precision, MESH remains the best approximation for observed hail size. Still, it is utilized to define hail objects representing hail cores, not to verify hail size.

d. Convective line cases

In addition to the "full season" forecasts over the entire HWT SFE period, four convective line cases (of 3, 11, 18, and 30 May 2018) are further examined to elaborate on relative systematic biases/skill. Each case uses a more limited verification domain compared to the full-season eastern CONUS domain (Fig. 1) to limit verification to the primary convective systems of interest.

The 3 May case contained two convective lines from two different synoptic/mesoscale structures. A stationary front situated over the Oklahoma Panhandle (Fig. 2a) forced convection between 1600 and 1700 UTC 2 May, and exhibited squall line structure over central Kansas by 2100 UTC. South of the stationary front was a north–south-oriented dryline that extended into west Texas. Convection along the dryline initiated between 1800 and 1900 UTC 2 May, and spawned storms with supercellular and other discrete structures. By 0100 UTC 3 May, the discrete cellular structures initialized along the dryline continued to merge upstream into a convective line over central Oklahoma that extended into northern Texas, and eventually merged with the northern squall line near 0400 UTC. For the 11 May case, a combination of a shortwave trough, terrain-following flow, and diurnal heating (not shown) initiated convection in southeast Wyoming between 1900 and 2000 UTC 10 May (Fig. 2b), and continued in western Nebraska and northeast Colorado shortly after. Strong, discrete storms at 2300 UTC in western Nebraska merged into a convective line that slowly traversed the state over the forecast period.

For the 18 May case, diurnal heating coupled with a prior outflow boundary (not shown) initiated convection in western South Dakota between 1700 and 1800 UTC 17 May (Fig. 2c).

Near 2300 UTC, these storms began merging with existing storms spanning the North–South Dakota border. During the forecast period, the storms organized into a prominent mesoscale convective system (MCS) from southern North Dakota to northwestern Nebraska. Near 0600 UTC, the convective line of the MCS was primarily in Nebraska, while its stratiform region encompassed northwestern Nebraska and central South Dakota. Convection initiated near a prior outflow boundary (not shown) on the Kansas–Oklahoma border (Fig. 2d) between 1800 and 1900 UTC 29 May. While early storms were more discrete, the storms merged upstream into a convective line by 0300 UTC 30 May centered over eastern Kansas. The dryline in the Texas Panhandle additionally spawned a discrete supercell that initiated between 2000 and 2100 UTC in the Texas Panhandle. This storm moved to the east into Oklahoma, and was isolated from the convective line over the forecast period.

3. Point-based forecast verification

a. *Full season forecasts*

The Morrison, NSSL, P3, and Thompson BMP schemes as present in the 2018 HWT SFE are examined (as a proof-of-concept forecast evaluation of convective hazards) among the four sets of short-term (t = 1–6 h) forecasts using the same Noah LSM and MYJ PBL schemes, EnKF ensemble mean initial conditions, and NAM forecast boundary conditions, as described in section 2a. WRF v3.9.1.1 is compiled with Intel compilers and forecasts are produced on Texas Advanced Computing Center's (TACC) Stampede2 supercomputer consisting of nodes with single 68-core Intel Xeon Phi 7250 CPUs, utilizing 420 cores over 8 computing nodes. Wall clock times for the four experiments (up to 6 h) for the 11 May case are 1.29, 1.37, 1.38, and 1.63 h using the Morrison, Thompson, P3, and NSSL schemes, respectively. Forecasted storm coverage in these experiments are evaluated using composite reflectivity thresholds of $Z \geq 15$ dBZ for the overall areal extent including stratiform and convective precipitation and $Z \geq 40$ dBZ for convective regions following Putnam et al. (2017b). Similarly, light and heavy precipitation forecasts are compared using 1-h AP thresholds of 0.01 and 0.5 in., which roughly correspond with the prior two Z thresholds based on the default Z–R relationship of operational WSR-88D radars ($Z = 300R^{1.4}$).

With the performance diagrams, the MORR experiment has small overall storm coverage bias (near the dashed bias = 1 line; Fig. 3a), indicating that its forecast misses and false alarms are comparable. However, the NSSL, P3, and THOM experiments display underprediction of this coverage, as each experiment's forecasts generally contain bias below 1. Critical success index (CSI) predictably drops from overall to convective region coverage (Fig. 3c), given that smaller-scale structures are more difficult to successfully predict than larger structures. There is a clear experiments demarcation between the MORR/NSSL overprediction of convective region coverage relative to complementing P3/THOM underprediction. Experiment biases of predicted storm coverage for the four convective lines are similar to the full season biases (Figs. 3b,d), indicating that the four convective line cases can represent reasonable analogs to link bias with microphysics scheme used in these forecasts. Convective line storm coverage CSI is typically larger than those over the full season, likely due to a smaller verification domain centered over significant convection.

As equitable threat score (ETS) is a flavor of CSI that removes the influence of random hits, storm coverage ETSs are generally consistent with prior CSIs (Figs. 4a,e). The NSSL experiment has the highest overall storm coverage ETS until 0200 UTC, after which the MORR experiment exceeds the NSSL (and other experiments). MORR overpredicts overall storm coverage compared to NSSL at t = 1 h (Fig. 4c), indicating that the NSSL's higher ETS at this time might be due to fewer false alarms. The P3 experiment has the lowest overall storm coverage ETS. Both the MORR and NSSL experiments have the highest full season convective region coverage ETSs over the forecast period. This higher ETS reflects the higher bias over the forecast period for the MORR and NSSL experiments, which are generally greater than 1 (Fig. 4g). While full season ETS is reflected in the ETSs of the convective lines (e.g., low overall storm coverage ETS in the P3 experiment; Figs. 4b,f), there is still high temporal and across-case ETS variance among the experiments. This indicates that the convective line ETSs are unreliable to elaborate on full-season storm coverage ETS trends due to individual storm evolution. Similar to CSI, convective line ETS is generally larger than ETS over the full season.

Overall storm coverage fractions skill score (FSS) is well above the skill line [$0.5 + (f_0/2)$, where f_0 is observed fractional coverage] both for the full season and across the four cases at nearly all neighborhood half-widths for the MORR, NSSL, and THOM experiments (Figs. 5a,b). While we retain constant storm coverage and precipitation thresholds when calculating FSS (consistent with performance diagrams and ETS), we note here that using percentiles can focus FSS more on spatial pattern verification without the influence of bias (Roberts and Lean 2008). While the P3 experiment is skillful across all scales for the full season, it requires larger neighborhood scales for skillful forecasts for the convective line cases. The THOM experiment typically produces the largest FSS across the four cases, while the P3 generally produces the smallest FSS, which is consistent with full season forecasts. Lower convective region coverage FSSs demonstrate less skillful forecasts: only the P3 and THOM experiments produce skillful forecasts over the season at larger verification scales (Fig. 5c). The NSSL experiment contains the smallest convective region coverage FSS over the season, which is generally reflected (along with the high THOM FSS) in the convective line FSSs (Fig. 5d). Overall storm coverage FSS is similar to or lower than full season FSS, while MORR and NSSL experiments display a noticeable increase in FSS from their low full season convective region coverage FSS. This increase is likely due to narrowing the verification area, given their convective region coverage overprediction.

Light/heavy precipitation (1-h AP ≥ 0.01, 0.5 in.) performance diagrams reveal an underprediction bias across the full season and the convective lines for each experiment (Figs. 3e–h).

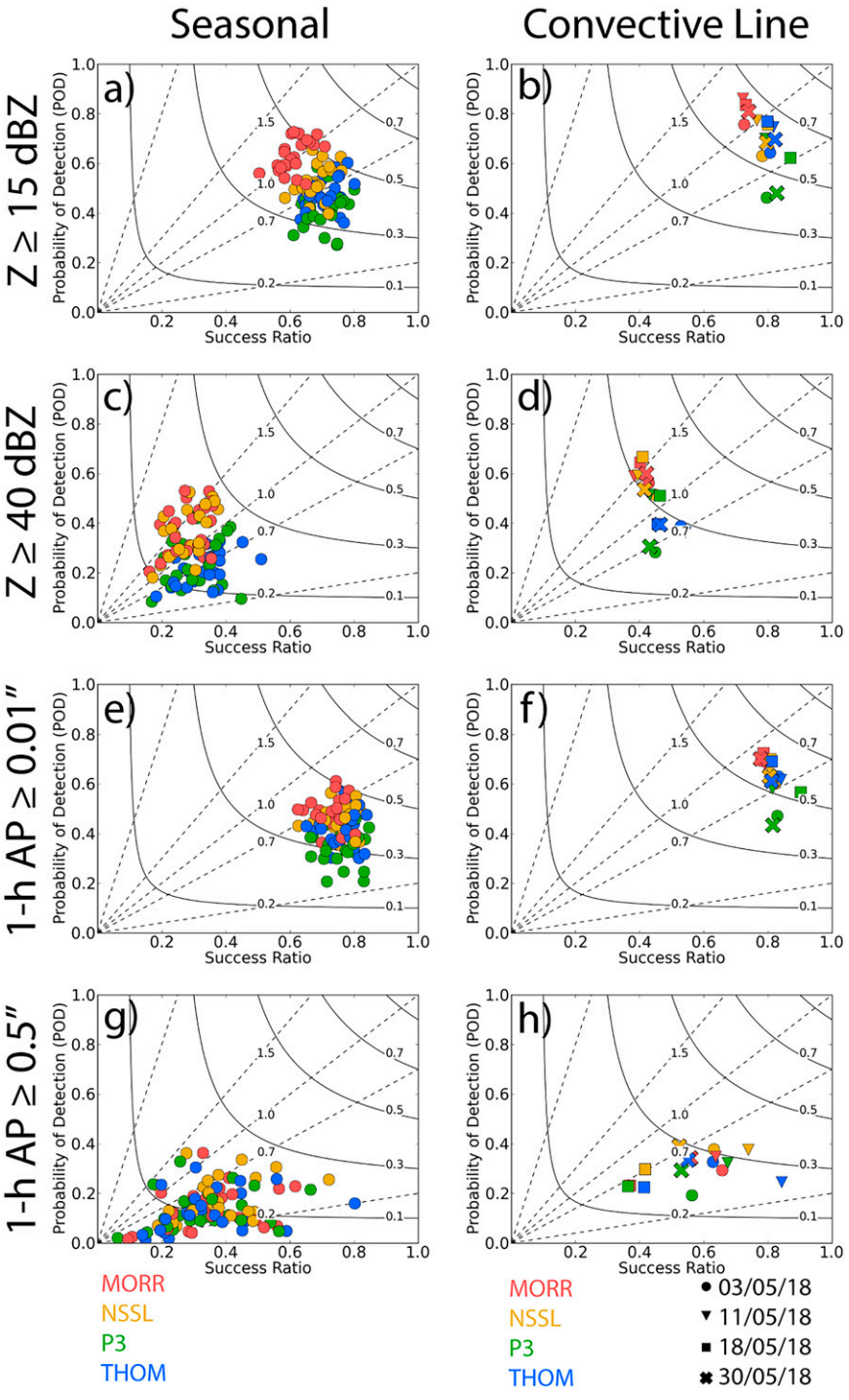

FIG. 3. Performance diagrams for full season and convective line forecasts of composite reflectivity Z (a),(b) ≥15 and (c),(d) ≥40 dBZ and 1-h accumulated precipitation (e),(f) ≥0.01 and (g),(h) ≥0.5 in. The MORR, NSSL, P3, and THOM experiments are represented by red, gold, green, and blue markers, respectively, while the convective line marker shape corresponds to convective line case as denoted in the legend below the subplots. Dashed lines represent constant bias, while solid lines are constant CSI.

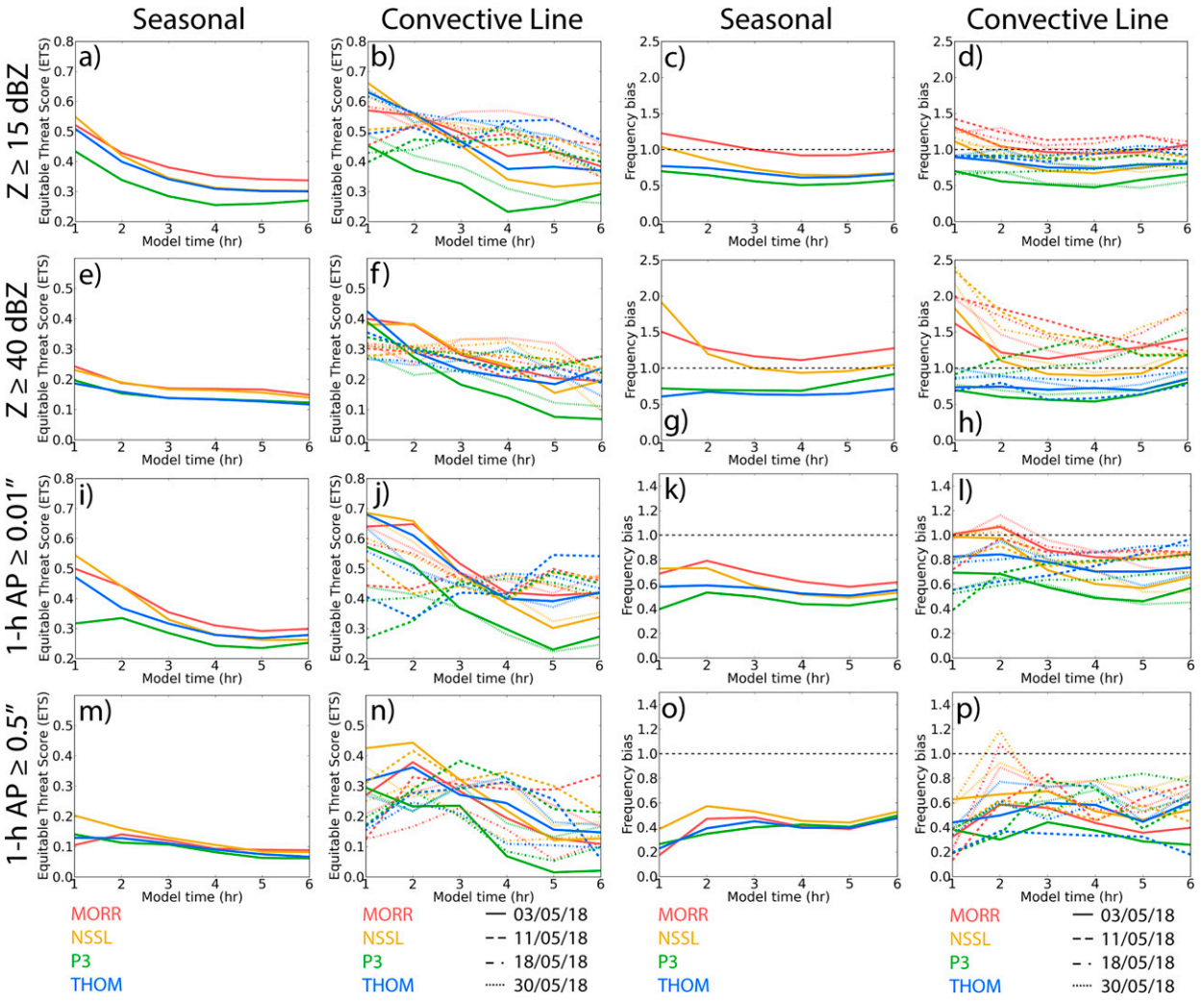

FIG. 4. Equitable threat scores (ETSs) in the left two columns and frequency bias in the right two columns for full season and convective line forecasts of composite reflectivity Z (a)–(d) ≥15 and (e)–(h) ≥40 dBZ and 1-h accumulated precipitation (i)–(l) ≥0.01 and (m)–(p) ≥0.5 in. The MORR, NSSL, P3, and THOM experiments are represented by red, gold, green, and blue lines, respectively, while convective line line styles correspond to convective line case as denoted in the legend below the subplots.

The relative BMP experiment stratification of full season and convective line light precipitation biases weakly resembles that of overall storm coverage in terms of composite reflectivity, linking the two but not entirely as precipitation is more dependent on the low-level hydrometeors (reflective of numerous factors affecting hydrometeor sedimentation) rather than the column maximum. The heavy precipitation low CSI and convective line groupings imply that these forecasts are more sensitive to the cases than microphysics scheme used in the experiment. Similar to storm coverage, convective line light and heavy precipitation CSI is typically larger than that over the full season.

Full season precipitation ETSs are similar in both value and span across the microphysics experiments examined (Figs. 4i,m), although the P3 experiment consistently contains the smallest light precipitation ETS over the forecast period. The NSSL experiment produces the highest light precipitation ETS and bias (Fig. 4k) at 0100 UTC, while the MORR experiment improves later with larger bias in the forecast, similar to overall storm coverage ETS. Heavy precipitation ETSs among the experiments are small for much of the forecast period. Similar to convective region coverage, the MORR and NSSL experiments contain slightly larger ETS than the P3 and THOM. Heavy precipitation bias in the NSSL experiment is largest over the period (Fig. 4o), increasing hits and subsequently ETS. Full season precipitation ETS is somewhat reflected in the convective lines (e.g., high initial light precipitation ETS in the NSSL experiment; Figs. 4j,n), but seem more sensitive to mesoscale/synoptic situations given the high temporal ETS variances across the cases. Convective line precipitation ETS is generally larger than over the full season.

The MORR experiment generally produces the highest light precipitation FSS across the convective lines even though the NSSL experiment has the highest full season FSS

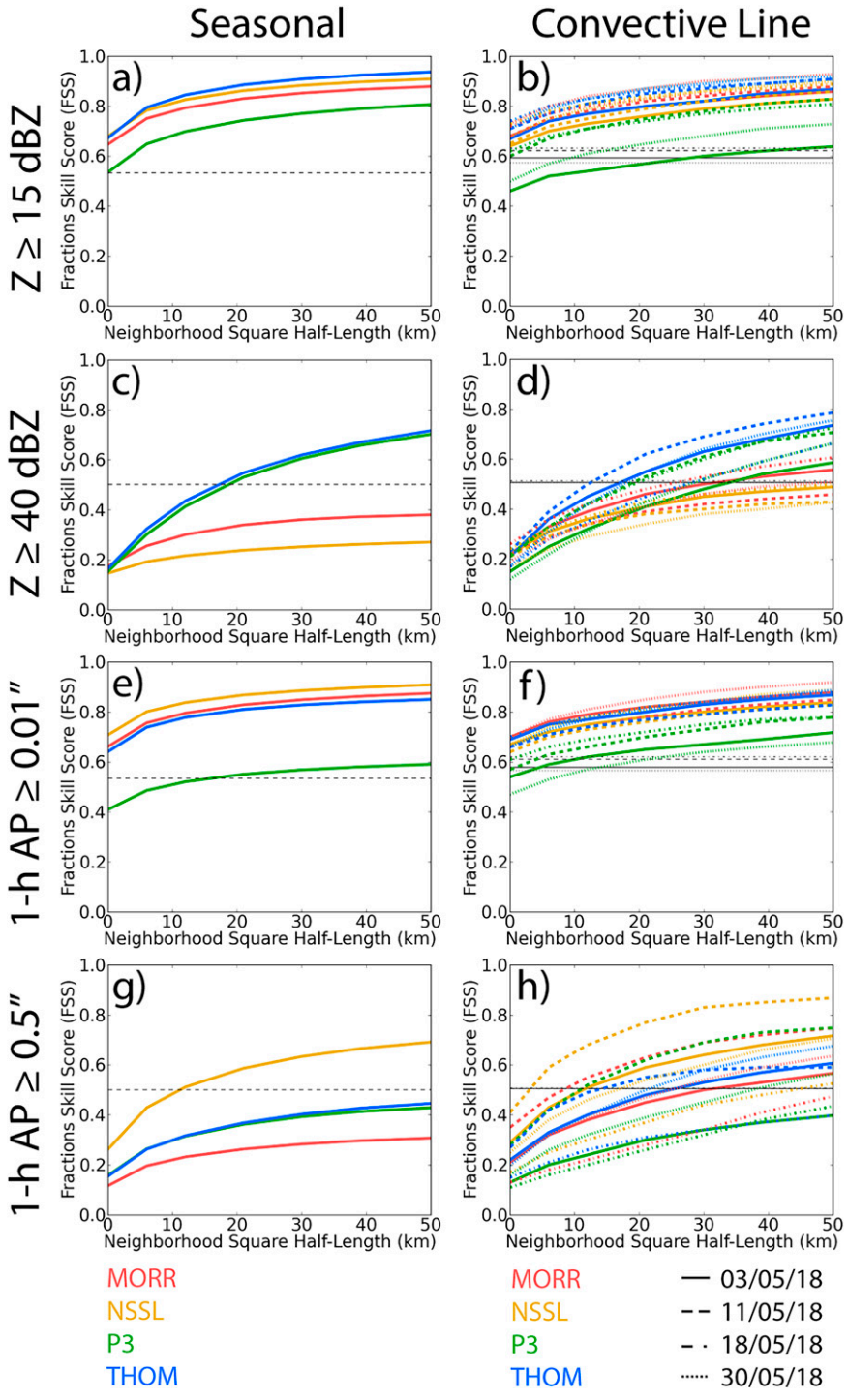

FIG. 5. Fractions skill score (FSS) for full season and convective line forecasts of composite reflectivity Z (a),(b) ≥15 and (c),(d) ≥40 dBZ and 1-h accumulated precipitation (e),(f) ≥0.01 and (g),(h) ≥0.5 in relative to neighborhood square half-length. The MORR, NSSL, P3, and THOM experiments are represented by red, gold, green, and blue lines, respectively, while convective line FSS line style corresponds to convective line case as denoted in the legend below the subplots. FSS exceeding its corresponding black dashed/line style line indicates a forecast with skill.

(Figs. 5e,f). Still, light precipitation FSSs over the full season are generally skillful, with the exception of the P3 experiment at small neighborhood scales. This is reflected in the convective lines. Both full season and convective line heavy precipitation FSSs indicate that NSSL best matches the observed heavy precipitation distribution, as the experiments struggle to forecast with skill (Figs. 5g,h). This is expected, as heavy precipitation areas (much like convective regions) are localized structures and require large verification scales. Convective line light precipitation FSS is improved for the P3 experiment, but generally similar for the remaining experiments compared to full season FSS. The MORR and THOM experiments have improved convective line heavy precipitation FSS. Convective line precipitation FSS might be larger than full season FSS when the verification domain is reduced to regions with strong convection, especially given the persistent precipitation underprediction bias.

b. *Selected convective line cases*

Paintball plots (i.e., filled contours matching verification thresholds that vary color with experiment) are examined for the convective line cases at 0100 UTC, when full season ETS is generally highest, to reconcile experiment biases and behavior from the forecasts (Fig. 6). We note here that areal coverage in the plots and its link to over/underprediction is not definitive because of the use of a neighborhood for verification, but it does provide a helpful estimate of relative biases. The overall storm coverage in MORR and NSSL experiments exceeds the areal extent in P3 and THOM at 0100 UTC (Figs. 6a–d). This areal overprediction by both experiments diminishes at later times although MORR better sustains this storm coverage compared to NSSL (not shown), consistent with full season and convective line biases (Figs. 3a,b and 4c,d), and full season ETSs (Fig. 4a) given the latter's sensitivity to hits. Overall storm coverage is smallest in the P3 experiment, which often simulates fragmented overall storm coverage not seen in observations and relative to the other experiments (e.g., Fig. 6d). This is consistent with the experiment's largest full season and convective line underprediction bias, and lowest full season ETS. MORR's largest relative overall storm coverage is consistent with its scheme's single slow-falling (and therefore, more prone to faster updraft ejection and horizontal advection) graupel-like rimed ice category used here, compared to the faster-falling NSSL scheme's two rimed ice categories and the Thompson scheme's graupel–hail "hybrid" rimed ice category. The P3 scheme's single ice category parameterization is more complicated by virtue of rime history, but its inability to simulate larger storm coverage may be due to the restrictive mean ice size limiter in this version of the scheme resulting in lower reflectivity compared to other experiments. While a lower rimed ice size has greater potential for horizontal advection as previously mentioned, an imposed upper limit of ice size can limit particle growth by riming and subsequently, reflectivity. This limiter has been removed from the scheme more recently, resulting in a general increase of composite reflectivity for an idealized supercell, especially in its forward flank (Milbrandt et al. 2021). We also note here the potential impact of snow on overall storm coverage in NSSL and THOM relative to other experiments (not shown): relatively small snow terminal velocity is likely enhancing overall storm coverage in NSSL due to favorable horizontal advection, while high snow production at the expense of cloud ice is a known bias with the Thompson scheme (e.g., Van Weverberg et al. 2013). Given the small stratification of overall storm coverage FSSs among the MORR, NSSL, and THOM experiments, areal overpredictions by MORR and NSSL at early forecast times (along with NSSL relative underprediction at later times) result in the THOM experiment typically producing higher FSS (Figs. 5a,b).

The convective region coverage overpredictions by MORR and NSSL experiments (Figs. 6e–h) lessen during the forecast period, but persist relative to P3 and THOM as well as observations (not shown). This is consistent with full season and convective line bias stratification (Figs. 3c,d and 4g,h), and full season ETSs (Fig. 4e). Such experiment behavior is not unexpected, given the Morrison scheme's excessive hydrometeor size sorting tendency (linked to the two-moment scheme's zero shape PSD parameter; see Johnson et al. 2016) and small graupel fall speed, the NSSL scheme's modeling of hail as a large, dense rimed-ice category facilitated by graupel wet growth (e.g., Johnson et al. 2016, 2019) enhancing reflectivity, and the P3 scheme's ice size limiter's potential to inhibit particle growth through riming. The Thompson scheme's modeling of rimed ice as a single-moment graupel–hail hybrid category combined with fast rimed ice fall speed for large particles precludes large hail reflectivity coverage, as well as enhanced graupel horizontal advection in its experiment. Newer versions of the P3 and Thompson schemes that improve rimed ice representation might address this underprediction bias, and therefore should be tested to fully inform microphysics selection. Regardless, the excessive areal convective region coverage evident in the MORR and NSSL experiments lower their FSS compared to THOM and P3 (Figs. 5c,d).

Systematic light precipitation areal underprediction is most obvious across the convective lines at 0100 UTC for the P3 and THOM experiments (Figs. 6i–l), consistent with their biases (Figs. 3f and 4l at t = 0100 UTC) and P3's lowest FSS (Figs. 5e,f). The Thompson scheme's known high snow production bias would lower its ice flux relative to other schemes, and therefore precipitation in its experiment. After initial overprediction for three of the four convective line cases, the MORR and NSSL experiments shift toward underprediction (not shown). Light precipitation coverage in the NSSL experiment diminishes more quickly than in MORR, consistent with full season light precipitation ETS of MORR exceeding NSSL at later times (Fig. 4i). This also explains why the MORR experiment generally produces the highest FSSs for these cases (Fig. 5f) considering persistent underprediction. While the light precipitation underprediction is consistent with overall storm coverage underprediction in the NSSL, P3, and THOM experiments, it seemingly contradicts the lack of MORR overall storm bias. However, the Morrison scheme's graupel fall speed is typically smaller than those in the NSSL and Thompson schemes (not shown), lowering its flux relative to the other schemes. Enhancing downward precipitation flux

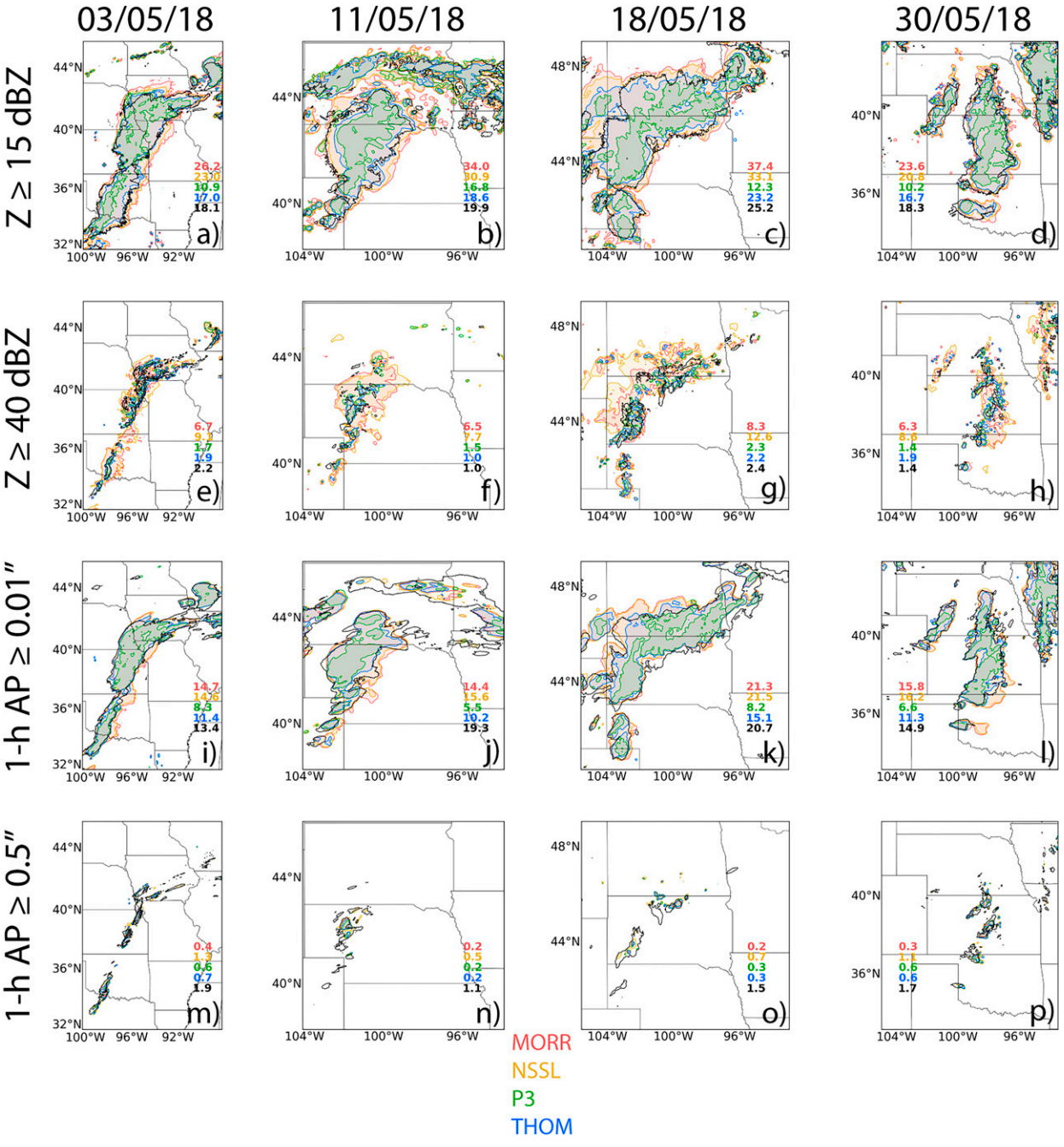

FIG. 6. (a)–(d) Overall storm and (e)–(h) convective region coverage ($Z \geq 15, 40$ dBZ), and (i)–(l) light and (m)–(p) heavy precipitation (1-h accumulated precipitation ≥ 0.01, 0.5 in., respectively) convective line paintball plots at t = 0100 UTC. MORR, NSSL, P3, and THOM experiments are represented by red, gold, green, and blue filled contours, respectively, while corresponding observations are black lines. Each subplot contains relevant storm or precipitation areal coverage (as a percentage) for each microphysics experiment and observations.

is an area for improvement for all microphysics schemes examined in their experiments, specifically fall speed and rime representation (graupel-like and/or hail-like categories or within the P3 scheme's ice PSD, snow versus rimed ice mass production), although modification might diminish storm coverage prediction. We note here that recent P3 and Thompson scheme updates (and the Morrison scheme's hail-like rimed ice) might improve downward hydrometeor fluxes (e.g., Morrison et al. 2015; Milbrandt et al. 2021).

Heavy precipitation is substantially underpredicted (Figs. 6m–p), in agreement with full season and convective line biases (Figs. 3g,h and 4o,p) and low ETS (Figs. 4m,n). The NSSL experiment

forecasts the largest heavy precipitation areal coverage at 0100 UTC, but not over the majority of the forecast period for the 11 and 18 May cases. This can likely be attributed to the scheme's hail flux, which sediments faster than rimed ice in the Morrison and Thompson schemes. The MORR experiment (which typically has the largest areal coverage for these cases) produces heavy precipitation false alarms for much of the 11 May case, resulting in the NSSL experiment producing the highest heavy precipitation FSSs for all cases (Fig. 5h). 18 May heavy precipitation evolution is more complicated, but the higher NSSL FSS might be linked to early MORR overprediction as MORR and NSSL underprediction are similar at other forecast times.

4. Object-based forecast verification

a. *Object dataset validity*

Rotation and hail (both hazards of interest to forecasters) proxy variable choice for object creation is evaluated by comparing maximum values after convolution smoothing at every model output time through correlation coefficient and bivariate KDEs (Fig. 7). Smoothing convolution radii for rotation and hail datasets are 3 and 12 km, respectively, with the smaller rotation radius to prevent oversmoothing of the convective-scale rotating updraft. Rotation KDEs reveal little difference among the four experiments, with fairly narrow KDE contours (i.e., linearity) and correlation coefficients ranging from 0.51 to 0.56 (Figs. 7a,d,g,j). This indicates that simulated 1–6-km updraft helicity and observed 3–6-km azimuthal shear are appropriate proxies for rotation object comparison. While microphysics can influence updraft helicity (such as latent heat release modifying buoyancy and pressure perturbations), these effects are secondary and more subtle than thermodynamic/dynamic storm environment, etc.

Hail objects representing hail cores are created from simulated and observed MESH (itself derived from reflectivity and subfreezing height levels), resulting in larger hail correlation and more linear KDEs than rotation in the NSSL, P3, and THOM experiments (Figs. 7e,h,k), but not the MORR experiment (Fig. 7b). Rimed ice in the Morrison scheme is represented by default as medium-density, slow-falling graupel. Compared to the Morrison scheme, the NSSL scheme contains graupel-like and hail-like rimed ice categories with predicted rimed ice density, the P3 scheme contains rimed ice (both fully and partially when applicable) in its ice PSD that varies density (spanning medium and high densities similar to graupel-like and hail-like in other bulk schemes) based on riming history, and the Thompson scheme's graupel–hail hybrid category with diagnosed intercept parameter contains rimed ice fall speeds similar to graupel in the NSSL scheme with density = 500 kg m^{-3} for large particles. This shifts the MORR experiment's greatest concentration of maximum simulated MESH to smaller sizes (due to quicker updraft ejection and subsequently, less rime growth) and reduces the span of simulated MESH sizes relative to those observed. As MORR MESH correlation coefficient is still 0.48 and its KDE reveals linearity, forecast and observed MESH across the experiments allow reasonable hail object comparison. Although simulated and observed MESH are similarly diagnosed, we also include KDEs of maximum hail size (MHS) similar to Eq. (3) in Milbrandt and Yau (2006), which is calculated from simulated rimed ice PSDs (i.e., graupel in the MORR and THOM experiments, hail with constant bulk density = 900 kg m^{-3} in the NSSL experiment) or diagnosed from ice in the P3 experiment using an assumed exponential PSD and constant bulk density = 500 kg m^{-3} (Figs. 7c,f,i,l). MHS can also be utilized to define hail cores, but the undesirable narrow range of maximum simulated MHS in the MORR and P3 experiments make simulated MESH a more appropriate hail core proxy for this set of forecasts.

Although a tail-end percentile may be employed for object definition (Skinner et al. 2018; Potvin et al. 2019), subsequent object creation and matching between simulated and observed fields might be suboptimal given the differing rotation distributions among the experiments (Fig. 7m). The observed 3–6-km azimuthal wind shear span near and below the 80th percentile is greater relative to simulated 1–6-km updraft helicity, resulting in a shortened distribution tail relative to the updraft helicity distribution (as seen above the 80th percentile). While this is reflective of the rotation proxies employed (i.e., extreme updraft helicity values are reasonable for isolated strong updrafts with well-defined rotation while azimuthal wind shear does not require vertical motion), it also motivates a trial-and-error object threshold approach, which is employed in this study. Hail distributions based on simulated MESH (Fig. 7n) or potentially, MHS (Fig. 7o), are more suited than rotation toward a percentile-based object definition. In this paper, hail objects created from simulated MESH also utilize a trial-and-error object threshold definition for consistency with rotation objects. Although each microphysics experiment might forecast objects with higher skill by individually optimizing object creation thresholds and/or matching/merging parameters outlined in section 2c, we apply the same thresholds and interest functions across the experiments for a fair comparison.

b. *Microphysical performance*

Forecast rotation objects are defined as 1–6-km updraft helicity $\geq$ 28 m^2 s^{-2}, while observed objects are defined as 3–6-km azimuthal wind shear $\geq$ 0.0035 or $\geq$0.0039 s^{-1} (Fig. 8). The lower observed threshold is chosen given the similar number of forecast and observed objects in the MORR experiment, while the higher threshold creates observed objects closer in number to forecasted objects in the NSSL, P3, and THOM experiments. These two thresholds alleviate biases across the microphysics experiments, as well as provide insight into the microphysics experiment span of forecast rotation. There is little experiment skill stratification when considering the performance diagrams, as CSI typically ranges between 0 and 0.3 across the experiments (Figs. 8a,b). Biases are largely contingent on the rotation object threshold, as the MORR experiment shifts from little bias to overprediction bias, the NSSL and THOM experiments shift from underprediction to overprediction bias (but less than MORR), and the P3 experiment

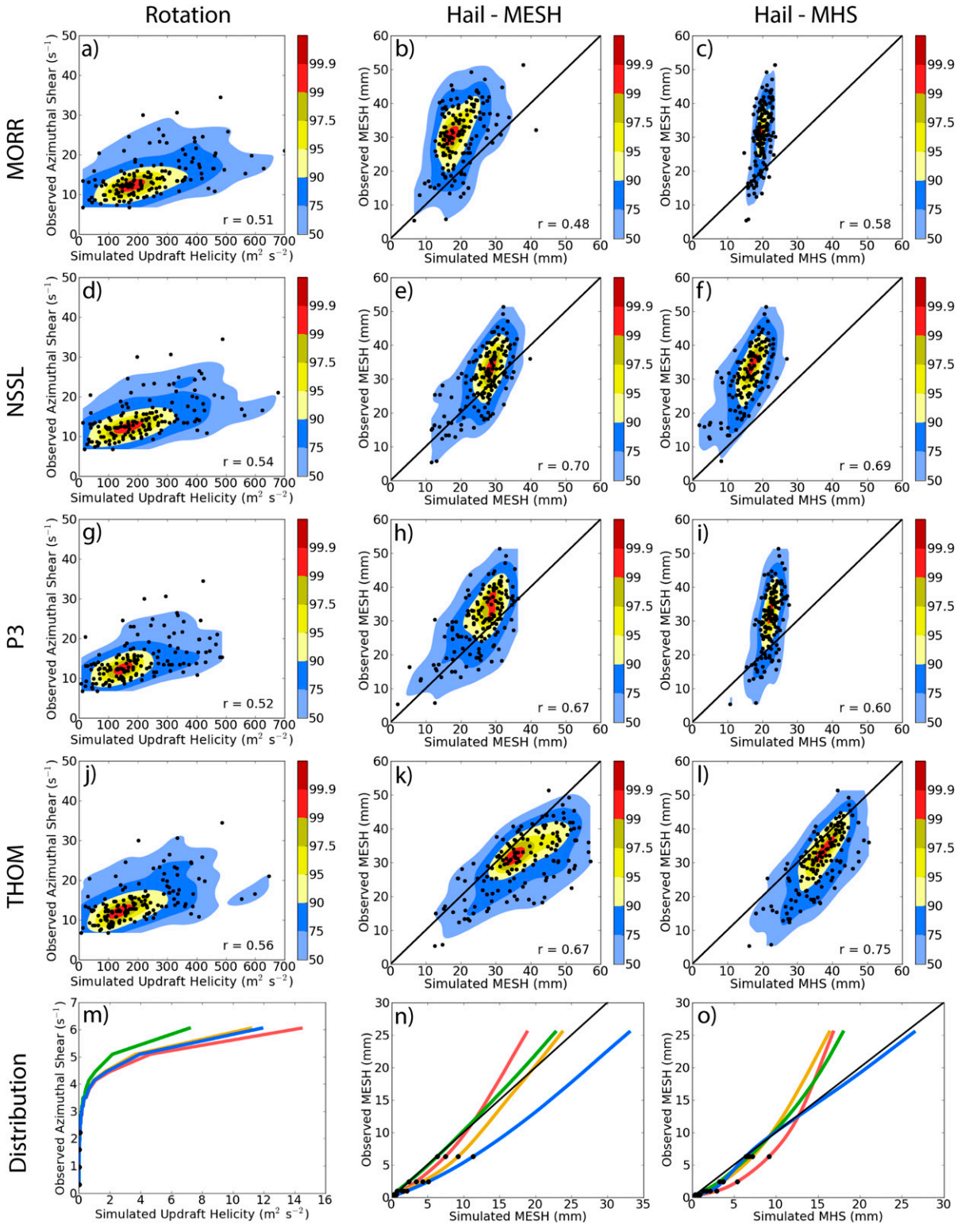

FIG. 7. Model and observed (left) rotation and hail object using simulated (center) MESH and (right) MHS dataset maxima at each model output time and their corresponding kernel density estimate (KDE) percentiles for the (a)–(c) MORR, (d)–(f) NSSL, (g)–(i) P3, and (j)–(l) THOM experiments, with correlation coefficients also shown. Observed 3–6-km azimuthal shear values have magnitude of 10^{-3}. Also plotted are model and observed distributions spanning the 1st–99th percentiles for (m) rotation and (n),(o) hail nonzero datasets. The MORR, NSSL, P3, and THOM experiments are denoted by red, gold, green, and blue lines, respectively, and black dots indicate 20th, 40th, 60th, and 80th percentiles. Diagonal lines in the hail plots designate perfect correspondence between simulated MESH and MHS, and observed MESH.

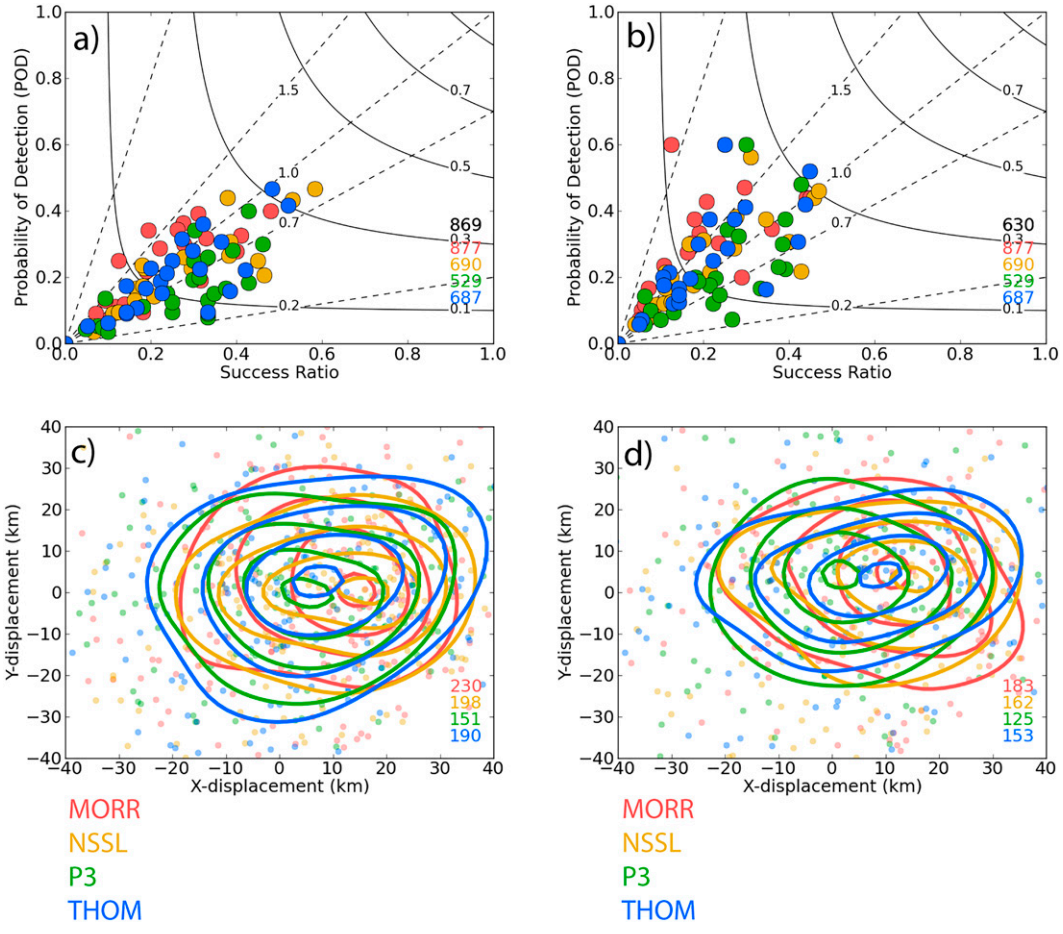

FIG. 8. Rotation object performance diagrams and displacements using thresholds of simulated 1–6-km updraft helicity $\geq 28\ m^2\ s^{-2}$ and observed 3–6-km azimuthal wind shear $\geq$ (a),(c) 0.0035 and (b),(d) 0.0039 s^{-1}, with MORR, NSSL, P3, and THOM experiments represented by red, gold, green, and blue, respectively. Performance diagrams additionally note the number of forecast and (black) observed objects, while displacement diagrams indicate matched objects. Dashed lines in performance diagrams represent constant bias, while solid lines are constant CSI.

shifts from underprediction to neutral-underprediction bias when the observation object threshold increases from 0.0035 to 0.0039 s^{-1} (Figs. 8a,b). This is reflective of the number of rotation objects that the threshold choice creates (i.e., as the number of forecast and observation objects get closer to each other, bias decreases), but also reveals the importance of optimal rotation object threshold choice across experiments. Rotation object displacement (defined as forecast centroid minus observed centroid) reveals an eastward bias with little to northward meridional bias, although the eastward bias is lowest for P3 (Figs. 8c,d). This displacement in the expected climatological storm motion for the CONUS domain in Fig. 1 is in agreement with other convective-scale forecasts (e.g., Yussouf et al. 2015; Skinner et al. 2016). However, previous object-based studies (Skinner et al. 2018; Flora et al. 2019) have been unable to conclusively isolate a positive storm motion bias. Different spatial and temporal evolutions between predicted and observed storms can result in displaced object centroids independent of advection biases. The inconsistent displacement shift (but still in the expected climatological storm motion direction) by increasing the observation object threshold highlights the storm morphology effect on displacement as only observed object geometry is modified (Figs. 8c,d).

Hail objects are defined in a similar manner as rotation objects, with forecast MESH threshold of ≥8 or ≥10 mm, and an observed MESH threshold of ≥8 mm (Fig. 9). Like rotation objects, there is little hail object skill stratification among the experiments. The P3 experiment contains an underprediction bias, MORR contains little bias, and the NSSL and THOM experiments contain overprediction biases at the forecast MESH threshold of ≥8 mm (Fig. 9a). Increasing the forecast threshold to ≥10 mm reduces the number of MORR hail objects more than the other experiments, shifting the experiment's hail object bias closer to P3's underprediction bias (Fig. 9b). THOM hail objects are less affected and display a neutral/slight overprediction bias for the forecast threshold of ≥10 mm. As seen in Figs. 7b, 7e, 7h, 7k, MORR's greatest concentration of maximum simulated MESH is at smaller sizes relative to other experiments. This is due to the design of the

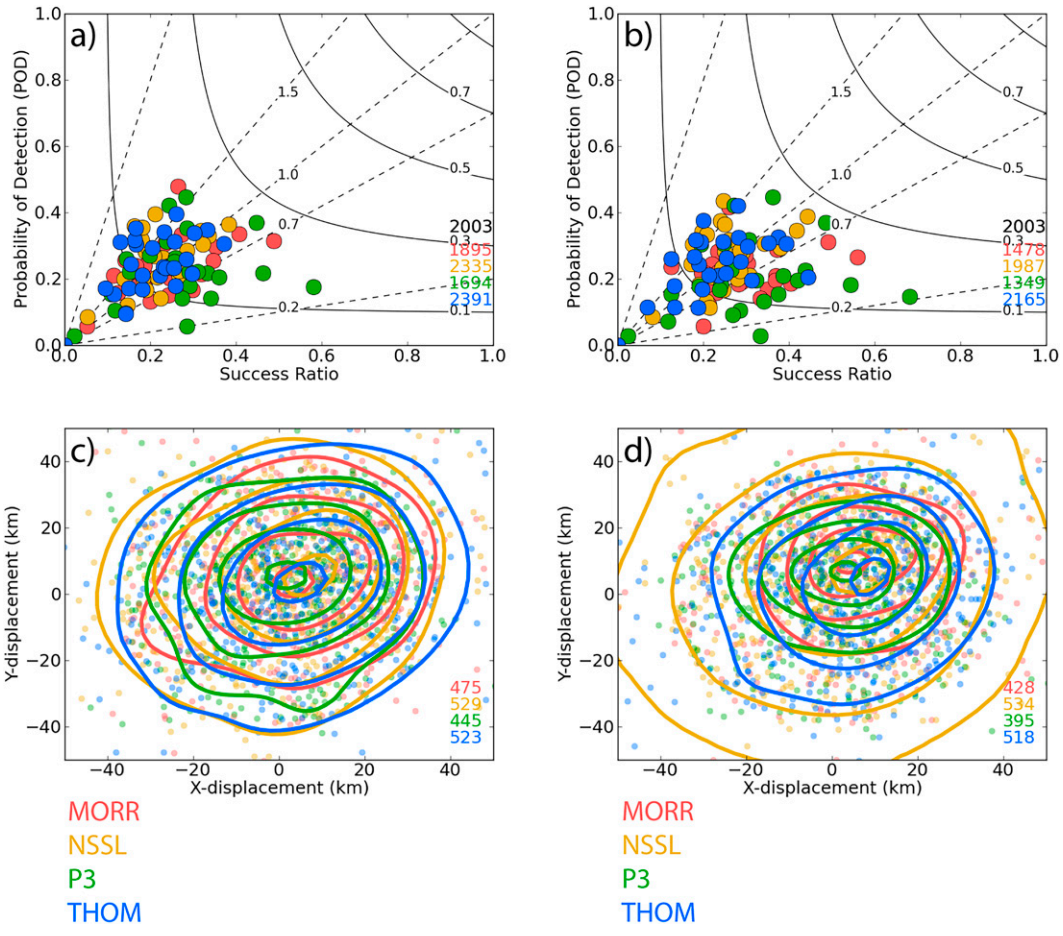

FIG. 9. As in Fig. 8, but for hail objects using thresholds of simulated MESH ≥ (a),(c) 8 and (b),(d) 10 mm, and observed MESH ≥ 8 mm.

Morrison scheme's slow-falling, graupel-like rimed-ice category, which quickly ejects particles out of updrafts, while graupel wet growth to hail in the NSSL scheme and the Thompson scheme's hybrid graupel–hail category with diagnostic intercept parameter are more conducive to rimed ice growth for the MESH sizes utilized to define hail objects. Similar to rotation object displacement, hail object displacement is in the expected storm climatological direction (Figs. 9c,d). Again, the extent of this north and east bias generally inconsistently shifts as the forecast object threshold increases.

Rotation and hail objects sampled in the convective line cases at 0100 UTC are shown to supplement prior object verifications (Figs. 10, 11). The lower rotation observation object threshold is utilized given that objects are reasonably expected to contract or disappear as this threshold increases, while the lower hail forecast object threshold represents a similar hail core definition (MESH ≥ 8 mm) between forecasts and observations. With the exception of MORR and NSSL for the 18 May case, the number of rotation objects are underforecasted by each experiment at 0100 UTC, owing to suboptimal object merging/larger forecast objects (e.g., Figs. 10a,e,i,m,q along the north Texas/Oklahoma convective line), and inability to organize rotating updrafts (e.g., Figs. 10d,h,l,p,t in the central Kansas organizing convective line). While the coverage of forecast objects may be similar or encompass observed objects, the goal of object verification is to match features of interest (i.e., rotating updraft, hail core). Thus, inability to resolve individual objects (either over- or undermatching) reveals a forecast skill deficiency regardless of object overlap. Further from 0100 UTC, experiment biases and differences are highly case dependent. Generally, experiments start underpredicting the number of rotation objects (Fig. 10), then overpredict by the end of the forecast period (not shown). MORR (3 and 18 May) and P3 (18 and 30 May) experiments simulate the number of rotation objects closest to observations over the forecast period; the reasons are inconsistent among the cases. This is in agreement with MORR but not P3 experiment full season biases (Fig. 8a), as only the 3 May case demonstrates the large P3 underprediction bias. MORR subtly increases convection north and throughout the squall line compared to THOM in the middle of the forecast period, and better maintains rotating convection than the NSSL experiment on 3 May (not shown). Suboptimal merging of forecast rotation objects may also play a role. Storms in the MORR-simulated MCS weaken relative to the NSSL, P3, and THOM experiments nearing the end of the 18 May case, although the total number of rotation objects are

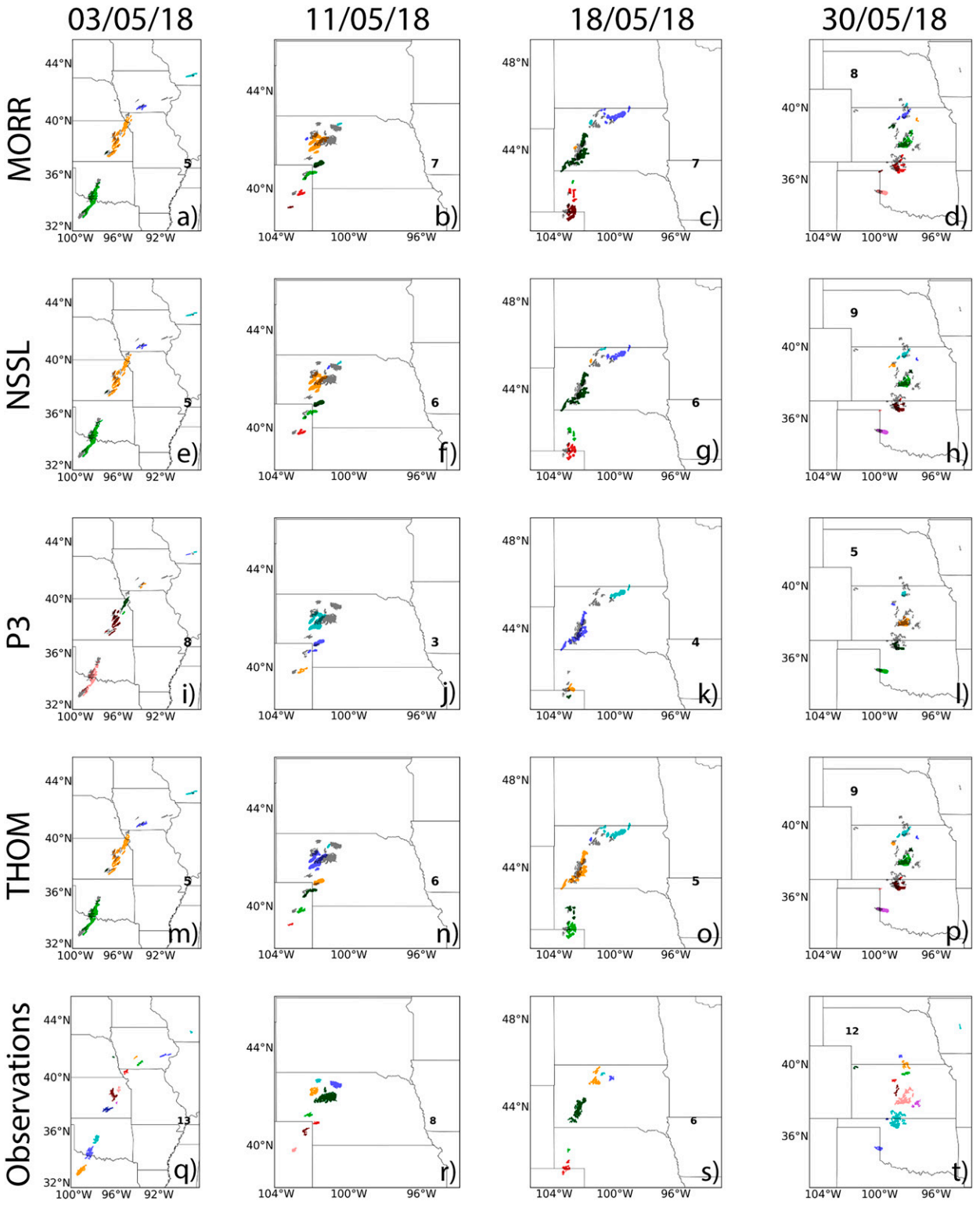

FIG. 10. Rotation object plots using a simulated 1–6-km updraft helicity threshold of 28 $m^2 s^{-2}$, and an observed 3–6-km azimuthal wind shear threshold of 0.0035 s^{-1} for the (a)–(d) MORR, (e)–(h) NSSL, (i)–(l) P3, and (m)–(p) THOM experiments, and (q)–(t) corresponding observations for the 3, 11, 18, and 30 May cases at t = 0100 UTC. Colors denote distinct objects in forecast and observation plots, while shaded observation objects are additionally present in forecast object plots. Each subplot indicates the number of objects.

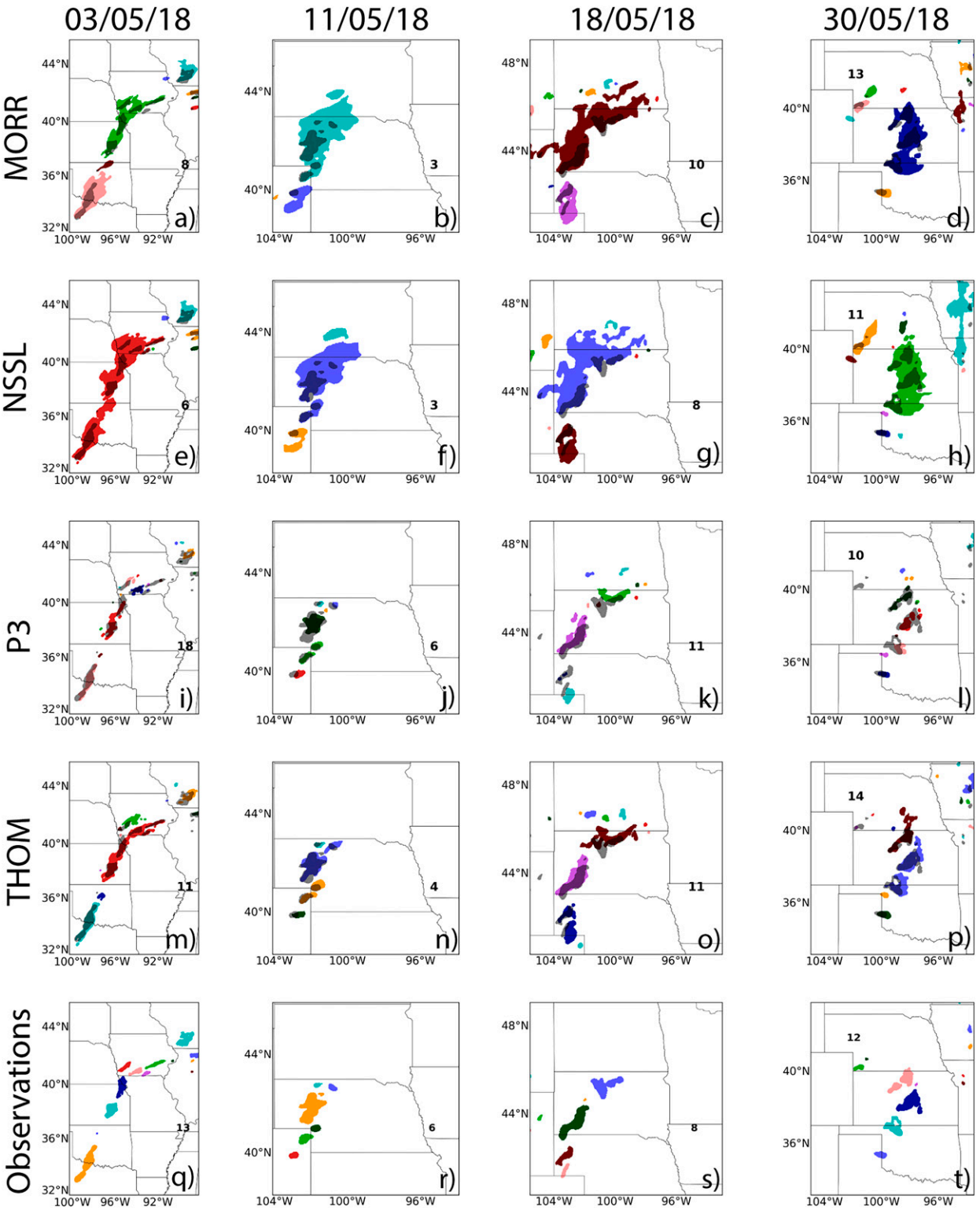

FIG. 11. As in Fig. 10, but for hail objects with simulated and observed object threshold definitions of MESH ≥ 8 and 8 mm, respectively.

highly similar among the experiments. P3 avoids rotation object overpredictions seen in other experiments on 30 May by simulating less rotation objects throughout the convective lines.

The number of hail objects are both over- and underpredicted at $t = 0100$ UTC. The underprediction is due to large/merged forecast objects (3 and 11 May; Fig. 11) and misses (e.g., P3 experiment on 30 May; Fig. 11l). As the forecast period advances, the number of hail objects are grossly overpredicted across all experiments (not shown), despite hail objects defined with identical hail size formulations and thresholds (simulated and observed MESH ≥ 8 mm): for the 11, 18, and 30 May cases, the number of simulated hail objects closest to observations is the experiment with the smallest overprediction. This has important implications when forecasting hail cores defined using MESH products: the number of simulated hail cores might be overforecast in convective lines compared to observations. This overprediction is greatest for the 18 May case: hail objects may extend to North Dakota and Minnesota, which contain no observed hail objects in the analysis domain over the forecast period. With a systematic gross overprediction in these cases, full season experiment biases are not obviously present. In fact, the THOM experiment typically simulates the fewest objects, while each experiment individually simulates the most hail objects in the four cases.

Finally, object plots reveal prominent eastward displacements of simulated rotation (e.g., northeastern Colorado for the 11 May case; Figs. 10b,f,j,n,r) and hail (e.g., Texas/Oklahoma for the 30 May case; Figs. 11d,h,l,p,t). Northward displacements are additionally present for rotation (e.g., northern Kansas for 30 May; Figs. 10d,h,l,p,t) and hail (e.g., southwestern Nebraska in the MORR and NSSL experiments, southwestern South Dakota in the P3 and THOM experiments for the 18 May case; Figs. 11c,g,k,o,s) objects. Figure 11 illustrates the forecasted northward and eastward displacement bias despite differing object geometries and pathlengths among the experiments for the 11 and 18 May cases.

5. Summary and discussion

We present a proof-of-concept systematic evaluation of convective hazards applied to four deterministic forecasts differing in microphysics only (specifically, the Morrison ["MORR"], NSSL, P3, and Thompson ["THOM"] partially or fully two-moment schemes as they were available) over approximately a one-month period of the 2018 NOAA Hazardous Weather Testbed (HWT) Spring Forecasting Experiment (SFE). We employ point-based verifications with a neighborhood (equitable threat score ETS and fractions skill score FSS) to assess relative microphysical storm coverage (both overall and convective regions, defined as composite reflectivity $Z \geq 15$ and 40 dBZ, respectively) and 1-h accumulated precipitation (both light and heavy, defined as 1-h AP ≥ 0.01 and 0.5 in., respectively) skill and biases. The evaluation also includes object-based verification for simulated updraft rotation and hail cores, two proxies for hazards of interest to forecasters. Four convective line cases are highlighted in addition to elaborate on relative microphysical biases/skill over the entire 2018 HWT SFE period. This proof-of-concept is useful for forecast systems because it can reveal relative storm coverage, precipitation, rotation, and hail biases and skill among forecasts that differ in microphysics only, and subsequently inform microphysics population in an ensemble forecast. The systematic evaluation would need to be applied to the most current/optimal versions of the schemes before any such recommendations can be made.

The MORR experiment contains little overall storm coverage bias relative to NSSL, P3, and THOM underprediction. MORR has the highest full season overall storm coverage ETS over much of the forecast period (t = 1–6 h). The full season ETS and FSS of the P3 experiment are lowest. For more intense convective region coverage, the MORR and NSSL experiments overpredict relative to P3 and THOM underprediction, resulting in larger full season ETS. Convective region coverage FSSs for the MORR and NSSL experiments are reduced due to this overprediction. There is a systematic underprediction bias for both light and heavy precipitation across experiments. The MORR experiment generally forecasts larger light precipitation ETS. P3 light precipitation FSS is much lower than the other experiments. Full season heavy precipitation ETS is low across all experiments considering their substantial underprediction biases. The NSSL experiment produces forecasts with the highest heavy precipitation FSS.

When increasing the threshold of observed rotation object definition, rotation prediction by each experiment shifts toward positive bias, as the neutral bias threshold for the MORR experiment is closer to $\geq 0.0035\ s^{-1}$, in between ≥ 0.0035 and $\geq 0.0039\ s^{-1}$ for the NSSL and THOM experiments, and closer to $\geq 0.0039\ s^{-1}$ for the P3 experiment. In other words, there are separate rotation object thresholds with minimum bias across experiments, indicating their effects on storm dynamics through thermodynamic processes. Similarly, increasing the forecast hail object threshold (MESH ≥ 8–10 mm) shifts biases toward underprediction, with the largest underprediction bias shift occurring with the MORR experiment. The P3 experiment simulates the largest hail underprediction bias in terms of the number of hail objects relative to other experiments. Even with identical products and thresholds (simulated and observed MESH ≥ 8 mm) defining hail core objects, the number of forecast hail objects are largely overpredicted across the convective lines. Object displacement and convective line object plots reveal a general simulated positive storm motion bias (northward and eastward bias).

Our forecast verification framework reveals that ice representation in microphysics remains a challenging, but crucial component in forecasts. By default, the Morrison scheme represents rimed ice as a medium-density, slow-falling particle, which can lead to quicker updraft ejection (limiting rime growth) and preferential horizontal advection relative to other schemes. This also limits the number of hail core objects defined using MESH simulated by its experiment. The Thompson scheme has a known snow production bias at the expense of cloud ice relative to other schemes, reducing ice flux and subsequent precipitation coverage in its experiment. The NSSL scheme simulates large rimed ice as graupel undergoing wet

growth to hail, with relatively slow-falling snow additionally contributing to overall storm coverage via enhanced horizontal advection in its experiment. In the version of the P3 scheme examined in this study, particle growth by riming can be limited by the mean ice size limiter, reducing storm coverage and the number of hail objects created by its experiment. Representing observed ice modes with a finite number of hydrometeor categories inevitably leads to ice parameterization differences among microphysics schemes, which have a large effect on subsequent storm evolution.

Compared to the other three BMPs, the P3 scheme uses a nontraditional design where ice category particle properties are predicted in lieu of multiple distinct predefined ice categories. Because it is relatively new, systematic evaluation of its relative performance for spring storm-season convection prediction and comparison with other schemes has not been documented much in the literature. The flavor of the P3 scheme evaluated here is the two-moment single-ice category with two-moment rain microphysics. Our proof-of-concept evaluation applied to the P3 version available in 2018 HWT SFE forecasts reveal fragmented overall storm coverage and large light precipitation underprediction bias in its experiment. Since the inception of P3 microphysics, more improvements have been added as the scheme develops. The P3 scheme with one ice category has the potential to introduce ice property error per the "dilution" effect, where riming ice can be stunted by a newly initiated/mixing with an ice distribution primarily composed of small spherical and nonspherical ice. The resulting distribution could then be less reflective of previous riming history. This motivated adding additional ice categories and merging categories with similar properties to separate ice modes. The P3 scheme has also incorporated prognostic liquid fraction (Cholette et al. 2019) and has recently evolved to predict three moments of ice (Milbrandt et al. 2021), which can help grow rimed ice within the single ice category more physically. We note here that prognosing three moments for rain has been explored in Paukert et al. (2019), but the feature is currently not implemented in the "official" P3 microphysics [see Cholette et al. (2023) for a description of P3 microphysics development]. Additionally, its mean ice size limiter has been removed. These improvements (particularly ice representation) will likely address some of the reflectivity coverage, precipitation, and hail core deficiencies found in the 2018 HWT SFE forecasts (e.g., Cholette et al. 2023). Similarly, the Thompson microphysics scheme has been updated to include prognostic graupel number and bulk volume to improve hail representation (Jensen et al. 2023). Given such continued updates and potential optimizations (e.g., hail-like rimed ice in the Morrison scheme) of ice representation, it is essential to systematically test and evaluate the schemes' performances and identify the potential biases and areas needing further improvements.

Our evaluation showed that all microphysics schemes examined in these deterministic forecasts over the 2018 HWT SFE period display convective hazard biases and deficiencies relative to each other. Employing an ensemble with multiple microphysics is one way to account for model error related to microphysical processes and ice representation. The verification framework employed here could be applied to an ensemble's mean, or to generate ensemble probabilities by considering the number of members exceeding a threshold or within an object. Again, we want to emphasize that to make recommendations for the choice of schemes toward the optimal design of future convective-scale ensemble forecasting systems, similar systematic evaluations should be performed based on forecasts of sufficient sample sizes that use the latest version of the candidate schemes. Further, optimal mixed-physics ensemble design in the context of microphysics membership should be explored in future studies.

Acknowledgments. This research was primarily supported by the NOAA Warn-on-Forecast (WoF) Grant NA16OAR4320115. The real-time and retrospective forecasts were produced on TACC supercomputer Stampede 2 via NSF Xsede allocation. We also thank Jason Milbrandt and two anonymous reviewers who helped to greatly improve this paper.

Data availability statement. 2D forecast and observed products are available at Harvard Dataverse via https://doi.org/10.7910/DVN/X74UQJ limited to reviewer access until publication (Johnson 2022). Surface charts are archived at the Weather Prediction Center (WPC): https://www.wpc.ncep.noaa.gov/archives/web_pages/sfc/sfc_archive.php.

REFERENCES

Ben Bouallègue, Z., and S. E. Theis, 2014: Spatial techniques applied to precipitation ensemble forecasts: From verification results to probabilistic products. *Meteor. Appl.*, **21**, 922–929, https://doi.org/10.1002/met.1435.

Benjamin, S. G., G. A. Grell, J. M. Brown, T. G. Smirnova, and R. Bleck, 2004a: Mesoscale weather prediction with the RUC hybrid isentropic-terrain-following coordinate model. *Mon. Wea. Rev.*, **132**, 473–494, https://doi.org/10.1175/1520-0493(2004)132<0473:MWPWTR>2.0.CO;2.

——, and Coauthors, 2004b: An hourly assimilation-forecast cycle: The RUC. *Mon. Wea. Rev.*, **132**, 495–518, https://doi.org/10.1175/1520-0493(2004)132<0495:AHACTR>2.0.CO;2.

——, J. M. Brown, and T. G. Smirnova, 2016a: 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.

——, and Coauthors, 2016b: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, https://doi.org/10.1175/MWR-D-15-0242.1.

Brooks, H. E., C. A. Doswell III, and M. P. Kay, 2003: Climatological estimates of local daily tornado probability for the United States. *Wea. Forecasting*, **18**, 626–640, https://doi.org/10.1175/1520-0434(2003)018<0626:CEOLDT>2.0.CO;2.

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.

Cholette, M., H. Morrison, J. A. Milbrandt, and J. M. Thériault, 2019: Parameterization of the bulk liquid fraction on mixed-phase particles in the Predicted Particle Properties (P3)

scheme: Description and idealized simulations. *J. Atmos. Sci.*, **76**, 561–582, https://doi.org/10.1175/JAS-D-18-0278.1.

——, J. A. Milbrandt, H. Morrison, D. Paquin-Ricard, and D. Jacques, 2023: Combining triple-moment ice with prognostic liquid fraction in the P3 microphysics scheme: Impacts on a simulated squall line. *J. Adv. Model. Earth Syst.*, **15**, e2022MS003328, https://doi.org/10.1029/2022MS003328.

Clark, A. J., W. A. Gallus, and M. L. Weisman, 2010: Neighborhood-based verification of precipitation forecasts from convection-allowing NCAR WRF Model simulations and the operational NAM. *Wea. Forecasting*, **25**, 1495–1509, https://doi.org/10.1175/2010WAF2222404.1.

——, 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.

Davis, C., B. Brown, and R. Bullock, 2006: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772–1784, https://doi.org/10.1175/MWR3145.1.

——, ——, ——, and J. Halley-Gotway, 2009: The Method for Object-Based Diagnostic Evaluation (MODE) applied to numerical forecasts from the 2005 NSSL/SPC spring program. *Wea. Forecasting*, **24**, 1252–1267, https://doi.org/10.1175/2009WAF2222241.1.

Dawson, D. T., M. Xue, J. A. Milbrandt, and M. K. Yau, 2010: Comparison of evaporation and cold pool development between single-moment and multimoment bulk microphysics schemes in idealized simulations of tornadic thunderstorms. *Mon. Wea. Rev.*, **138**, 1152–1171, https://doi.org/10.1175/2009MWR2956.1.

Fan, J., and Coauthors, 2017: Cloud-resolving model intercomparison of an MC3E squall line case: Part I—Convective updrafts. *J. Geophys. Res. Atmos.*, **122**, 9351–9378, https://doi.org/10.1002/2017JD026622.

Flora, M. L., P. S. Skinner, C. K. Potvin, A. E. Reinhart, T. A. Jones, N. Yussouf, and K. 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.

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

Hou, T., H. Lei, Z. Hu, J. Yang, and X. Li, 2020: Simulations of microphysics and precipitation in a stratiform cloud case over northern China: Comparison of two microphysics schemes. *Adv. Atmos. Sci.*, **37**, 117–129, https://doi.org/10.1007/s00376-019-8257-0.

Janjić, Z. I., 1994: The step-mountain eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. *Mon. Wea. Rev.*, **122**, 927–945, https://doi.org/10.1175/1520-0493(1994)122<0927:TSMECM>2.0.CO;2.

Jensen, A. A., G. Thompson, K. Ikeda, and S. Tessendorf, 2023: Improving the representation of hail in the Thompson microphysics scheme. *Mon. Wea. Rev.*, https://doi.org/10.1175/MWR-D-21-0319.1, in press.

Johnson, A., and X. Wang, 2017: Design and implementation of a GSI-based convection-allowing ensemble data assimilation and forecast system for the PECAN field experiment. Part I: Optimal configurations for nocturnal convection prediction using retrospective cases. *Wea. Forecasting*, **32**, 289–315, https://doi.org/10.1175/WAF-D-16-0102.1.

Johnson, M., 2022: Deterministic and ensemble forecasts. Harvard Dataverse, https://doi.org/10.7910/DVN/X74UQJ.

——, Y. Jung, D. T. Dawson II, and M. Xue, 2016: Comparison of simulated polarimetric signatures in idealized supercell storms using two-moment bulk microphysics schemes in WRF. *Mon. Wea. Rev.*, **144**, 971–996, https://doi.org/10.1175/MWR-D-15-0233.1.

——, ——, J. A. Milbrandt, H. Morrison, and M. Xue, 2019: Effects of the representation of rimed ice in bulk microphysics schemes on polarimetric signatures. *Mon. Wea. Rev.*, **147**, 3785–3810, https://doi.org/10.1175/MWR-D-18-0398.1.

Jouan, C., J. A. Milbrandt, P. A. Vaillancourt, F. Chosson, and H. Morrison, 2020: Adaptation of the Predicted Particles Properties (P3) microphysics scheme for large-scale numerical weather prediction. *Wea. Forecasting*, **35**, 2541–2565, https://doi.org/10.1175/WAF-D-20-0111.1.

Jung, Y., M. Xue, and M. Tong, 2012: Ensemble Kalman filter analyses of the 29–30 May 2004 Oklahoma tornadic thunderstorm using one- and two-moment bulk microphysics schemes, with verification against polarimetric radar data. *Mon. Wea. Rev.*, **140**, 1457–1475, https://doi.org/10.1175/MWR-D-11-00032.1.

Keil, C., and G. C. Craig, 2009: A displacement and amplitude score employing an optical flow technique. *Wea. Forecasting*, **24**, 1297–1308, https://doi.org/10.1175/2009WAF2222247.1.

Kong, F., 2018: 2018 CAPS Spring Forecast Experiment program plan. NOAA/NWS/SPC, 29 pp.

Labriola, J., Y. Jung, C. Liu, and M. Xue, 2021: Evaluating forecast performance and sensitivity to the GSI EnKF data assimilation configuration for the 28–29 May 2017 mesoscale convective system case. *Wea. Forecasting*, **36**, 127–146, https://doi.org/10.1175/WAF-D-20-0071.1.

Loken, E. D., A. J. Clark, M. Xue, and F. Kong, 2019: Spread and skill in mixed- and single-physics convection-allowing ensembles. *Wea. Forecasting*, **34**, 305–330, https://doi.org/10.1175/WAF-D-18-0078.1.

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.

Milbrandt, J. A., and M. K. Yau, 2005a: A multimoment bulk microphysics parameterization. Part I: Analysis of the role of the spectral shape parameter. *J. Atmos. Sci.*, **62**, 3051–3064, https://doi.org/10.1175/JAS3534.1.

——, and ——, 2005b: A multimoment bulk microphysics parameterization. Part II: A proposed three-moment closure and scheme description. *J. Atmos. Sci.*, **62**, 3065–3081, https://doi.org/10.1175/JAS3535.1.

——, and ——, 2006: A multimoment bulk microphysics parameterization. Part III: Control simulation of a hailstorm. *J. Atmos. Sci.*, **63**, 3114–3136, https://doi.org/10.1175/JAS3816.1.

——, and H. Morrison, 2016: Parameterization of cloud microphysics based on the prediction of bulk ice particle properties. Part III: Introduction of multiple free categories. *J. Atmos. Sci.*, **73**, 975–995, https://doi.org/10.1175/JAS-D-15-0204.1.

——, S. Bélair, M. Faucher, M. Vallée, M. L. Carrera, and A. Glazer, 2016: The pan-Canadian high resolution (2.5 km) deterministic prediction system. *Wea. Forecasting*, **31**, 1791–1816, https://doi.org/10.1175/WAF-D-16-0035.1.

——, H. Morrison, D. T. Dawson II, and M. Paukert, 2021: A triple-moment representation of ice in the Predicted Particle

Properties (P3) microphysics scheme. *J. Atmos. Sci.*, **78**, 439–458, https://doi.org/10.1175/JAS-D-20-0084.1.

Morrison, H., and J. A. Milbrandt, 2015: Parameterization of cloud microphysics based on the prediction of bulk ice particle properties. Part I: Scheme description and idealized tests. *J. Atmos. Sci.*, **72**, 287–311, https://doi.org/10.1175/JAS-D-14-0065.1.

——, G. Thompson, and V. Tatarskii, 2009: Impact of cloud microphysics on the development of trailing stratiform precipitation in a simulated squall line: Comparison of one- and two-moment schemes. *Mon. Wea. Rev.*, **137**, 991–1007, https://doi.org/10.1175/2008MWR2556.1.

——, J. A. Milbrandt, G. H. Bryan, K. Ikeda, S. A. Tessendorf, and G. Thompson, 2015: Parameterization of cloud microphysics based on the prediction of bulk ice particle properties. Part II: Case study comparisons with observations and other schemes. *J. Atmos. Sci.*, **72**, 312–339, https://doi.org/10.1175/JAS-D-14-0066.1.

——, and Coauthors, 2020: Confronting the challenge of modeling cloud and precipitation microphysics. *J. Adv. Model. Earth Syst.*, **12**, e2019MS001689, https://doi.org/10.1029/2019MS001689.

Murillo, E. M., and C. R. Homeyer, 2019: Severe hail fall and hailstorm detection using remote sensing observations. *J. Appl. Meteor. Climatol.*, **58**, 947–970, https://doi.org/10.1175/JAMC-D-18-0247.1.

Naeger, A. R., B. A. Colle, N. Zhou, and A. Molthan, 2020: Evaluating warm and cold rain processes in cloud microphysical schemes using OLYMPEX field measurements. *Mon. Wea. Rev.*, **148**, 2163–2190, https://doi.org/10.1175/MWR-D-19-0092.1.

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.

Paukert, M., J. Fan, P. J. Rasch, H. Morrison, J. A. Milbrandt, J. Shpund, and A. Khain, 2019: Three-moment representation of rain in a bulk microphysics model. *J. Adv. Model. Earth Syst.*, **11**, 257–277, https://doi.org/10.1029/2018MS001512.

Potvin, C. K., and Coauthors, 2019: Systematic comparison of convection-allowing models during the 2017 NOAA HWT Spring Forecasting Experiment. *Wea. Forecasting*, **34**, 1395–1416, https://doi.org/10.1175/WAF-D-19-0056.1.

Putnam, B. J., M. Xue, Y. Jung, N. A. Snook, and G. Zhang, 2014: The analysis and prediction of microphysical states and polarimetric radar variables in a mesoscale convective system using double-moment microphysics, multinetwork radar data, and the ensemble Kalman filter. *Mon. Wea. Rev.*, **142**, 141–162, https://doi.org/10.1175/MWR-D-13-00042.1.

——, ——, ——, ——, and ——, 2017a: Ensemble probabilistic prediction of a mesoscale convective system and associated polarimetric radar variables using single-moment and double-moment microphysics schemes and EnKF radar data assimilation. *Mon. Wea. Rev.*, **145**, 2257–2279, https://doi.org/10.1175/MWR-D-16-0162.1.

——, ——, ——, G. Zhang, and F. Kong, 2017b: Simulation of polarimetric radar variables from 2013 CAPS spring experiment storm-scale ensemble forecasts and evaluation of microphysics schemes. *Mon. Wea. Rev.*, **145**, 49–73, https://doi.org/10.1175/MWR-D-15-0415.1.

Reisner, J., R. M. Rasmussen, and R. T. Bruintjes, 1998: Explicit forecasting of supercooled liquid water in winter storms using the MM5 mesoscale model. *Quart. J. Roy. Meteor. Soc.*, **124**, 1071–1107, https://doi.org/10.1002/qj.49712454804.

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., 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.

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., L. J. Wicker, D. M. Wheatley, and K. H. Knopfmeier, 2016: Application of two spatial verification methods to ensemble forecasts of low-level rotation. *Wea. Forecasting*, **31**, 713–735, https://doi.org/10.1175/WAF-D-15-0129.1.

——, 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.

Smith, T. M., and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) severe weather and aviation products: Initial operating capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 1617–1630, https://doi.org/10.1175/BAMS-D-14-00173.1.

Snook, N., F. Kong, K. A. Brewster, M. Xue, K. W. Thomas, T. A. Supinie, S. Perfater, and B. Albright, 2019: Evaluation of convection-permitting precipitation forecast products using WRF, NMMB, and FV3 for the 2016–2017 NOAA hydrometeorology testbed flash flood and intense rainfall experiments. *Wea. Forecasting*, **34**, 781–804, https://doi.org/10.1175/WAF-D-18-0155.1.

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.

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.

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.

——, R. M. Rasmussen, and K. Manning, 2004: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part I: Description and sensitivity analysis. *Mon. Wea. Rev.*, **132**, 519–542, https://doi.org/10.1175/1520-0493(2004)132<0519:EFOWPU>2.0.CO;2.

——, P. R. Field, R. M. Rasmussen, and W. D. Hall, 2008: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Mon. Wea. Rev.*, **136**, 5095–5115, https://doi.org/10.1175/2008MWR2387.1.

Van Weverberg, K., and Coauthors, 2013: The role of cloud microphysics parameterization in the simulation of mesoscale convective system clouds and precipitation in the tropical western Pacific. *J. Atmos. Sci.*, **70**, 1104–1128, https://doi.org/10.1175/JAS-D-12-0104.1.

Wang, Y., and X. Wang, 2017: Direct assimilation of radar reflectivity without tangent linear and adjoint of the nonlinear observation operator in the GSI-based EnVar system: Methodology and experiment with the 8 May 2003 Oklahoma City tornadic supercell. *Mon. Wea. Rev.*, **145**, 1447–1471, https://doi.org/10.1175/MWR-D-16-0231.1.

Wheatley, D. M., N. Yussouf, and D. J. Stensrud, 2014: Ensemble Kalman filter analyses and forecasts of a severe mesoscale convective system using different choices of microphysics schemes. *Mon. Wea. Rev.*, **142**, 3243–3263, https://doi.org/10.1175/MWR-D-13-00260.1.

Witt, A., M. D. Eilts, G. J. Stumpf, E. DeWayne Mitchell, J. T. Johnson, and K. W. Thomas, 1998: Evaluating the performance of WSR-88D severe storm detection algorithms. *Wea. Forecasting*, **13**, 513–518, https://doi.org/10.1175/1520-0434(1998)013<0513:ETPOWS>2.0.CO;2.

Xue, L., and Coauthors, 2017: Idealized simulations of a squall line from the MC3E field campaign applying three bin microphysics schemes: Dynamic and thermodynamic structure. *Mon. Wea. Rev.*, **145**, 4789–4812, https://doi.org/10.1175/MWR-D-16-0385.1.

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

——, and Coauthors, 2007: CAPS realtime storm-scale ensemble and high-resolution forecasts as part of the NOAA Hazardous Weather Testbed 2007 Spring Experiment. *22nd Conf. on Weather Analysis and Forecasting/18th Conf. on Numerical Weather Prediction*, Park City, UT, Amer. Meteor. Soc., 3B.1, https://ams.confex.com/ams/22WAF18NWP/techprogram/paper_124587.htm.

——, and Coauthors, 2008: CAPS realtime storm-scale ensemble and high-resolution forecasts as part of the NOAA Hazardous Weather Testbed 2008 Spring Experiment. *24th Conf. on Severe Local Storms*, Savannah, GA, Amer. Meteor. Soc., 12.2, https://ams.confex.com/ams/24SLS/techprogram/paper_142036.htm.

Yu, B., K. Zhu, M. Xue, and B. Zhou, 2020: Using new neighborhood-based intensity-scale verification metrics to evaluate WRF precipitation forecasts at 4 and 12 km grid spacings. *Atmos. Res.*, **246**, 105117, https://doi.org/10.1016/j.atmosres.2020.105117.

Yussouf, N., E. R. Mansell, L. J. Wicker, D. M. Wheatley, and D. J. Stensrud, 2013: The ensemble Kalman filter analyses and forecasts of the 8 May 2003 Oklahoma City tornadic supercell storm using single- and double-moment microphysics schemes. *Mon. Wea. Rev.*, **141**, 3388–3412, https://doi.org/10.1175/MWR-D-12-00237.1.

——, D. C. Dowell, L. J. Wicker, K. H. Knopfmeier, and D. M. Wheatley, 2015: Storm-scale data assimilation and ensemble forecasts for the 27 April 2011 severe weather outbreak in Alabama. *Mon. Wea. Rev.*, **143**, 3044–3066, https://doi.org/10.1175/MWR-D-14-00268.1.

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 Testbed with different physics combinations. *Geophys. Res. Lett.*, **46**, 3523–3531, https://doi.org/10.1029/2018GL081702.

Zhang, J., and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimation: Initial operating capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 621–638, https://doi.org/10.1175/BAMS-D-14-00174.1.