UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

EVALUATION OF THE EXPERIMENTAL WARN-ON-FORECAST SYSTEM AND WOF-HYBRID 3DENVAR SYSTEM ON SHORT-TERM FORECASTS FOR 2021 REAL-TIME CASES

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE IN METEOROLOGY

By

NOAH T. CARPENTER
Norman, Oklahoma
2022

EVALUATION OF THE EXPERIMENTAL WARN-ON-FORECAST SYSTEM AND WOF-HYBRID 3DENVAR SYSTEM ON SHORT-TERM FORECASTS FOR 2021 REAL-TIME CASES

A THESIS APPROVED FOR THE
SCHOOL OF METEOROLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Adam J. Clark, Chair

Dr. Jidong Gao

Dr. Kelvin K. Droegemeier

Dr. Steven M. Cavallo

Dr. Xuguang Wang

Acknowledgements

First, I would like to thank my advisors Patrick Burke, and Dr. Adam Clark. They both believed in my abilities to complete and defend a thesis in less than a year. In addition, both Patrick and Adam were very accommodating and supportive, which helped me stay motivated throughout the research process during the pandemic. Thanks also to Adam for the countless hours helping me with many coding issues. I would also like to thank Dr. Jidong Gao for helping me understand data assimilation processes and providing feedback on my figures and analysis. I also want to thank my other committee members, Dr. Kelvin Droegemeier, Dr. Steven Cavallo, and Dr. Xuguang Wang for reading my thesis and providing extensive feedback. Special thanks to Dr. Patrick Skinner for providing code and helping me understand the object-based methodology used in this study. Another thank you to Brian Matilla for teaching me how to interpolate my data onto the WoFS grid at different resolutions. I would also like to thank the OU SoM, NOAA/NSSL, and the Warn-on-Forecast project for the opportunity to perform this research.

I am also very thankful for all the friendships I made at OU and will cherish the memories made and those we will continue to make. I also appreciate all my undergrad friends for providing much needed study breaks, rant sessions, and trip opportunities to get my mind off grad school. I also want to thank my parents and my sisters for allowing me to ramble on about my research and classes without context and for supporting me every step of the way. Finally, I would like to thank my boyfriend, Matthew, for the endless support and sticking with me through all the ups and downs, both personal and professional, that come with being in grad school. I appreciate all of you and everything you have done for me over the past 2.5 years; there is no way I could have done this without y’all.

Table of Contents

Abstract

Over the last few decades, it has become more important than ever to provide accurate forecasts for severe hazards that have become more common due to climate change effects. Over this time, several forecasting experiments have been performed with increasing computer power to better our understanding of these hazards. Currently, severe thunderstorms are diagnosed through a Warn-on-Detection paradigm, which bases severe warnings on storm reports or live radar data. To increase severe warning lead times, a Warn-on-Forecast has been developed, which focuses on the forecast evolution of ensemble systems to focus on probabilistic guidance of individual thunderstorm hazards. From this, the Warn-on-Forecast System (WoFS, defined in Section 1.3) and Warn-on-Forecast Hybrid System (WoF-Hybrid, defined in Section 1.5) have been developed.

The purpose of this study is to demonstrate and evaluate the capability of WoFS and WoF-Hybrid for predicting short-term severe weather forecasts that occurred during 2021 and help to identify room for future improvement. In addition to the 3 km grid-spacing, 18-member Experimental Warn-on-Forecast System (WoFS), which uses a Gridpoint Statistical Interpolation- Ensemble Kalman Filter (GSI-EnKF) data assimilation method, a 1.5 km grid-spacing, single member hybrid three-dimensional ensemble–variational data assimilation (3DEnVar System; referred to hereafter as WoF-Hybrid) has been tested for several years in NOAA Hazardous Weather Testbed (HWT) Spring Forecasting Experiments (SFE). Whether WoF-Hybrid exhibits attributes that complement WoFS is an open question. This is addressed by applying a spatial verification method, Fractions Skill Score (FSS), and object-based verification measures to 31 cases between April–December 2021. For the spatial verification method, WoF-

Hybrid reflectivity FSSs are significantly lower than the WoFS member average at forecast initialization. Scores converge after one hour and are higher for WoF-Hybrid through the rest of the period. The difference in skill at initialization may be attributed to a higher reflectivity bias and a greater number of spurious convective cells in WoF-Hybrid.

The object-based method applied to reflectivity shows similar probability of detections (PODs) and false alarm ratios (FARs) for both systems. For updraft helicity (UH), however, WoF-Hybrid yields better PODs at all lead times. WoF-Hybrid better resolves UH swaths seen in plan-view analysis of forecast output, which tend to be narrow in the verification dataset. Partitioning cases based on event severity per the number of local storm reports, this study shows WoF-Hybrid has higher PODs and lower FARs than the WoFS member average for both reflectivity and UH in eight high-end events. WoF-Hybrid forecast reflectivity objects are both closer in size and location to observed objects than are WoFS objects.

Three severe weather events of 26 May and 27 May (multi-mode severe and heavy rain), and 10 December (rare winter tornado outbreak) are selected for detailed investigation. In each May case, PODs and FARs for both prediction systems were similar; however, for 10 December, WoF-Hybrid far outperformed WoFS in forecasts of the single most impactful thunderstorm in the dataset (quad-state tornadic supercell). These results suggest that using both WoF-Hybrid and WoFS forecast guidance may better support NWS forecasters' warning and forecast decisions.

1. Introduction and Literature Review

1.1 Motivation for Improving Severe Weather Forecasts

Two of the most common billion-dollar disasters in the United States are severe thunderstorms and flash flooding, with 152 and 36 events, respectively since 1980 (Smith 2022). Flash flood disasters are responsible for an average of 18 deaths per event, while severe thunderstorm disasters yield an average of 13 deaths per event (NOAA NCEI 2022). A 2020 report from the United Nations states that when comparing 1980-1999 to 2000-2019, major flash flooding events have more than doubled ,and the frequency of severe thunderstorm events has risen 40%, with climate change being the main culprit (UNDRR 2020). As these events are occurring more frequently, it is important to accelerate our understanding to better alert the public when their lives are in danger. One way to improve forecasts is to verify model output against observational data to understand a model's strengths and weaknesses in predicting hazardous weather. This may increase the accuracy and confidence of forecasts made by model users/forecasters, thereby increasing trust between meteorologists and the general populace. In addition, this allows model developers to identify areas for model improvement. Currently, National Weather Service (NWS) severe thunderstorm forecasts include watches up to 6 hours before the event and warnings within the hour before a hazard occurs in a given area. This "watch-to-warning" practice has been mainly provided by extrapolations from existing radar and other in-situ observations until recent years. However, there is a growing desire for efficient and effective guidance from convection-allowing models (CAMs) (Stensrud et al. 2009) to improve forecasts at these short-range lead times. It is important to acknowledge the history of how CAMs came to be to better understand how these systems operate, improve systematic errors, and occasionally provide very accurate storm-scale forecasts.

1.2 Research Leading to the Development of WoFS

In the 1960s, computer resources were not as powerful and readily available as they are today. Only idealized two-dimensional simulations were produced to further understand conditionally unstable environments through the thermal convection of cumulus clouds (Ogura and Charney 1962). Enabled by an increase in computing power during the late 60s and early 70s, three-dimensional model simulations could be produced with coarse grid spacing. A study by Wilhelmson (1974) showed that the addition of vertical wind shear compared to a two-dimensional model produced longer-lasting, taller, more realistic cloud representations agreeing with Steiner (1973) that only three-dimensional models can create accurate convective simulations. Various modeling studies throughout the late 70s and 80s focused on a small range of convective environmental conditions, such as unidirectional/directionally varying vertical wind shear (Weisman and Klemp 1982, 1984) and multicellular storm environments (Wilhelmson and Chen 1982), which helped better our understanding of the dynamics of convective storms. A more comprehensive review of convective, storm-scale modeling studies from the 60s through the 80s can be found in Wilhelmson and Wicker (2001).

The initial focus of researchers was on improving storm-scale models and increasing computing power to produce the most accurate simulations. With computing power increasing through the 1980s, it became possible for researchers at the Center for Analysis and Prediction of Storms (CAPS, established in 1989 one of the National Science Foundation's first 11 Science and Technology Centers) to evaluate the evolution of a cloud model starting from three-dimensional observations for the first time (Lilly 1991; Droegemeier 1997). Lilly successfully evaluated a forecast run with mesoscale resolution from a single Doppler radar network, however, the results were far from useful operationally. Lilly (1991) determined that in order for these data to be of greater use, models need to be able to deal with highly intermittent variables while also increasing the utility of Doppler radar data. In 1993, CAPS began a yearly testing of its forecasting system (Advanced Regional Prediction System; ARPS). During the first two experiments, ARPS was run on a small 115 km x 115 km domain, which could not fully predict the exact location of storms but provided useful information regarding storm mode (Janish et al. 1995). By the 1995 iteration, a 336 km x 336 km grid with 3 km grid spacing was nested within a 1200 km x 1200 km grid with 15 km grid spacing (Xue et al. 1996). This experiment performed well on three cases that were analyzed even without convection present in the initial state (Droegemeier et al. 1996). However, it was determined that these systems need to establish a reliable track record over many years to give confidence to forecasters in its predictions (Xue et al. 1996). Another forecasting project known as the Storm and Mesoscale Ensemble Experiment (SAMEX) took place in May 1998 consisting of the first real-time operation of four ensembles of mesoscale models, one of which was ARPS. The major takeaway from this study was that analyses coming from an ensemble containing multiple models outperforms the performance of individual ensemble systems (Hou et al. 2001). In addition, the perturbations of initial and boundary conditions and microphysics schemes are important for regional forecasting models and should be further examined (Hou et al. 2001).

Also ongoing in the 90s was another experiment by Brooks et al. (1993) which used the 1991 Storm Type Operational Research Model Test Including Predictability Evaluation (STORMTIPE-1991). This model input one representative sounding into a three-dimensional cloud model, which triggered convection and used these results to determine convective mode and evolution. This methodology was computationally inexpensive; however, these simulated storms did not interact with fronts, drylines, or outflow boundaries, which is quite unrealistic (Wicker et al. 1995). Brooks et al. (1993) found that numerical model forecasts performed better than traditional sounding analysis in the prediction of storm modes, with six of the 12 forecast days producing good forecasts, two of which were significant tornado events. A later iteration of this experiment, STORMTIPE-1995 (Wicker et al. 1997), used a similar methodology and over the six-day period, consistently produced a high bias in convection over the six-day period. One positive aspect of this result was that it allowed forecasters to understand the worst-case scenario of severe convection; however, this could lead to higher errors in model forecasts (Wicker et al. 1997). These two experiments helped exhibit the potential for convective-scale operational forecasting.

By the early 2000s, research into storm-scale models was primarily focused on determining if an increase in model resolution led to a significant increase in model skill and on increasing model domain size as computing power increased. A 4 km grid mesh was most commonly used in the larger domain models, as a study by Bryan et al. (2003) showed that this was the coarsest resolution necessary to accurately depict midlatitude convective systems**.** One of the earlier real-time CAM experiments, which was created by the National Center for Atmospheric Research (NCAR), occurred in the early-to-mid 2000s using the Weather Research and Forecasting (WRF) model (Skamarock et al. 2005). Around the same time, there was a lot of interest in creating CAM guidance for forecasters, which could facilitate tests of new verification and visualization techniques. A framework, NSSL-WRF, was created by a collaborative effort between the

National Severe Storms Laboratory (NSSL), the Global Systems Laboratory (GSL), the Storm Prediction Center (SPC), and the Environmental Modeling Center (EMC) in 2006 and was the predecessor to many current-day CAMs (Heinselman et al. 2022). NSSL-WRF was a sophisticated, yet computationally efficient model that generated real-time, daily one-to-36-hour forecasts of convection using 4-km grid spacing on a CONUS domain (NOAA NSSL, 2022a). The two systems assessed in this study currently use many of the experimental diagnostics from NSSL-WRF.

While initial CAMs were quite useful in depicting convective storm mode, sampling errors in the observations led to error-prone forecasts. There were errors due to uncertainties in the forward operators and uncertainties of radial velocity and reflectivity measurements which led to an incorrect initialization of convection (Zhang et al. 2004). Furthermore, systematic errors in CAMs produced inaccurate forecasts in time, location, size, and intensity in a variety of model variables. In a study by Zhang et al. (2006), it was found, especially in finer-resolution CAMs, that initial analysis errors can cause large forecast errors over a 36-hour forecast period. In addition, small-scale initial errors in random noise can grow rapidly, leading to spurious convection in short-term mesoscale analyses and forecasts. More accurate forecasts could not be produced without better data assimilation techniques or enhanced observational data to improve initial analysis (Zhang et al. 2006). Even when observational data had decent accuracy, if the background covariances were not accurately estimated, an increase in observational accuracy could not increase the accuracy of the analysis and model forecast (Zhang et al. 2004). These findings demonstrated the importance of correcting systematic errors while maintaining accurate observational data, to improve CAM guidance.

Multiple studies have shown that increasing model resolution from ~30 km to ~10 km shows a great increased skill in verification metrics (Colle et al. 2003; Rife and Davis 2005). When decreasing to ~5 km grid spacing, storm structure is better depicted than with coarser models; however, there are fewer, if any, increases in model skill (Mass et al. 2004). An increase in computing power and model resolution throughout the mid 2000s allowed for more research surrounding the analysis of storm structures with higher-resolution CAMs. Stensrud et al. (2009) argued that models now had the capability to forecast storm-scale attributes, focusing on individual thunderstorms to help improve lead time for warnings of their severe hazards. It was also noted that the Warn-on Detection method of severe weather warnings based on radar reflectivity signatures and observer reports couldn’t provide sufficient lead times, especially to those located near a report. Stensrud et al. (2009) suggested a shift to a Warn-On-Forecast paradigm to employ CAMs to explore the potential of increasing warning lead times. One idea to implement forecasting focused on individual storms was to create model output with a CAM that could produce an ensemble of convective-scale forecasts using a probabilistic approach (Roebber et al. 2004; Stensrud et al. 2009). This probabilistic guidance would allow for end users to analyze different output possibilities and how they compare to one another using one system. Work on such a system, which is now known as the Warn-On Forecast System (referred to hereafter as WoFS), funded by The National Oceanic and Atmospheric Administration (NOAA), began in 2009 with an aim to bring researchers and forecasters together to bridge the gap between severe weather watches and warnings, using a rapidly updating, high-performance, real-time CAM ensemble (Stensrud et al. 2013).

By the early 2010s, the Warn-on-Forecast project was established. Shortly thereafter, several studies were able to begin assimilating high-resolution satellite and Multi-Radar Multi-Sensor (MRMS) velocity and reflectivity data (Kong et al. 2006; Wheatley et al. 2015; Jones et al. 2016; Yussouf et al. 2015, 2016). Wheatley et al. (2015) and Kong et al. (2006) showed that an ensemble based on these data assimilation methods had the potential to forecast various storm modes; however, a higher-resolution model would be necessary to determine if the ensemble could correctly anticipate mesovortices. In addition, Yussouf et al. (2015) produced a retrospective ensemble forecast of the 27 April 2011 Alabama tornado outbreak in which the ensemble continuously assimilated MRMS data every five minutes for six hours, producing one-hour forecasts every 15 minutes. The ensemble was able to consistently predict the mesocyclones for all the isolated supercell tornadoes examined. The study stated errors in storm motion and spurious cell production were observed, but overall, evidence for promising skillful probabilistic updraft helicity (UH) output for this event was shown, which could have provided on average up to 18 minutes of lead time for the subset of 11 tornadoes examined.

With computing power increasing tremendously by 2016, NSSL began real-time WoFS experiments. WoFS is initialized every 30 minutes producing output every 5 minutes while continuously assimilating observations every 15 minutes to provide rapid updates and convergence toward solutions, combatting the uncertain nature of individual thunderstorms. Notably, forecasts produced at this pace and projected out 3–6 hours, reached forecasters within a few tens of minutes of the model initial time, making them viable for use in watch-to-warning operations. During experimental real-time runs, forecasts are made from 1700 UTC to 0300 UTC, and this is repeated for multiple dates each spring in the NOAA Hazardous Weather

Testbed Spring Forecasting Experiment (SFE). Over the last seven years, these experiments have been evaluated by forecasters, researchers, and students to determine WoFS forecast skill as well as how to best develop WoFS for operational forecast use (Heinselman et al. 2022).

1.3 WoFS Configuration

WoFS is an experimental, real-time, convection-allowing ensemble that aims to fill the temporal gap between convective outlook and thunderstorm warning administration (Heinselman et al. 2022). WRF is used with the Advanced Research WRF dynamic solver core as the base model for WoFS (Skamarock 2008). In addition, WoFS uses the NSSL double-moment cloud microphysics scheme, which has been shown to produce a better representation of convective precipitation than a single-moment scheme (Mansell et al. 2010; Yussouf et al. 2013). WoFS consists of 36 members with 3-km grid spacing over a 900 km by 900 km domain where storms are predicted to develop for a given day (Heinselman et al. 2022). All the members use a GSI-EnKF data assimilation system (Liu et al. 2018) implemented by the High-Resolution Rapid Refresh Ensemble (HRRRE) and assimilate WSR-88D radar (reflectivity and radial velocity), GOES-16, -17 satellite data (clear sky radiance and cloud water path), and ASOS surface data interpolated to 5-km grid spacing every 15 minutes (Figure 1a; Burke 2019; Clark et al. 2021a; Dowell et al. 2011,2021). Every ensemble member has a unique amalgamation of initial conditions, boundary conditions, radiation schemes, and planetary boundary layer physics schemes which provide unique six-hour forecasts to provide a metric for forecast uncertainty (Skinner et al. 2018). The first 18 members, which are analyzed in this study, produce 0–6-hour forecasts that are initialized on the hour while the final 18 members produce 0–3-hour forecasts initialized 30 minutes after the hour (Clark et al. 2021a). Initial background fields for WoFS are derived from a one-hour forecast by a 36-member, hourly cycled High-Resolution Rapid Refresh Data Assimilation System (HRRRDAS) initialized at 1400 UTC, while initial conditions for each initialization come from the model analysis fields from initial data assimilation cycles. The Boundary conditions are given by the HRRRE forecasts with stochastic physics initialized from the HRRRDAS analysis at 1200 UTC (Clark et al. 2021a).

1.4 Research Leading to the Development of WoF-Hybrid

In addition to forecast ensembles, researchers at the Center for Analysis and Prediction of Storms (CAPS) and NSSL were looking into deterministic, and ensemble forecast improvements throughout the late 1990s and 2000s (Kong et al. 2006, 2007). They pursued 3DVAR data assimilation methods which aim to solve the analysis problem by minimizing a so-called cost function by assimilating radar and satellite data (Gao et al. 1999, 2004; Stensrud and Gao 2010). The cost function is defined in Gao et al. (1999) as the summation of all squared errors due to the mismatch between model analyses and observations. The minimization of the cost function can be done in a single step, which depicts one of the major benefits to this approach: it is much more cost-effective than an ensemble of forecasts (Gao et al. 2013b). In addition, forecasters currently use deterministic guidance to issue warnings and forecasts, and therefore it would be easier to incorporate into operations, whereas a rapidly updating storm-scale ensemble will require new training. Gao et al. (2013a) analyzed four real-time tornado cases from the spring of 2010 and showed that analyses from the 1-km 3DVAR system could depict detailed mesocyclone structure as well as strong mid-level vertical vorticity, both of which have been quantitatively verified. By 2012, after model improvements, SFE participants noted 3DVAR successes with regard to mesocyclone detection, and this increased their confidence in their own forecasts (Gao et al. 2013a).

In a study by Gao et al. (2013b), a new hybrid 3DVAR-EnKF (3DEnVAR) data assimilation algorithm was introduced. The hybrid method yields lower root-mean-square errors than EnKF and 3DVAR methods for hydrometeor related variables. Gao et al. (2013b) also determined that the hybrid 3DEnVAR better fits the observations at forecast initialization which can decrease storm spin-up time. Within data assimilation systems, there is always a balance between resolution and computational cost. The hybrid system ingests ensemble-derived covariances and produces a single deterministic forecast, which decreases computational cost and allows for an increase in resolution, while gaining benefits of EnKF-initialized ensemble forecasts (Gao et al. 2013b; Gao and Stensrud 2014).

The system using the 3DEnVAR algorithm, WoF-Hybrid, was tested in the 2017 SFE and produced accurate forecasts for several severe weather events (Gao et al. 2017). One major drawback of the WoF-Hybrid system in the 2017 SFE, however, was that many spurious convective cells were produced, which may be attributed to excessive moisture inputs from the cloud analysis package (Wang et al. 2018). Two studies, Wang et al. (2019) and Pan et al. (2021), analyzed five cases in May 2017 with multiple configurations of WoF-Hybrid to determine potential forecast utility. In Wang et al. (2019), WoF-Hybrid resolution differences (1.5 km vs. 3 km), the inclusion of a cloud analysis scheme, and the ensemble mean were compared. The ensemble mean and WoF-Hybrid systems produced similar results when comparing reflectivity output, but WoF-Hybrid had more accurate UH tracks in shape, size, and location. In addition, the inclusion of a cloud analysis scheme in WoF-Hybrid produced too large of an MCS, less skillful forecasts, and more spurious cells than the observed data. WoF-Hybrid at 1.5 km resolution generally improves forecasts compared to the 3 km WoF-Hybrid. This suggests that the inclusion of a 1.5 km deterministic system in tandem with the 3 km WoFS ensemble may provide useful information, but this hypothesis has not been systematically tested and serves as the motivation for the current study. It was determined that the WoF-Hybrid 1.5 km system without a cloud analysis scheme provided the best forecasts both subjectively and quantitatively, motivating the continued work on this system (Wang et al. 2019). In Pan et al. (2021), WoF-Hybrid was tested with three different sets of assimilated data: radar; both radar and surface; and radar, surface, and satellite-derived products. In general, when adding satellite data, reflectivity and UH forecasts improved in skill due to improvements in water vapor distribution, the saturation of the moisture fields, and low-level vertical wind shear. Upper layer precipitable water observations and increased spurious cell production with the addition of satellite data were determined to be causes of lower skill. It was suggested that better background error covariances and a quality control process for precipitable water inputs may provide solutions to these problems (Pan et al. 2021). This study aims to help determine the strengths and shortcomings of WoF-Hybrid as a companion piece to WoFS to further improve forecast output.

1.5 WoF-Hybrid Configuration

WoF-Hybrid uses the flow-dependent background error covariances derived from the 3 km WoFS background analysis and gives one deterministic forecast at 1.5 km resolution derived from 3DEnVAR methods over the WoFS 900 km by 900 km domain (Wang et al. 2019). The alpha control method (Wang et al. 2007) to efficiently integrate ensemble information from

WoFS was implemented in WoF-Hybrid. Similar to WoFS, WoF-Hybrid assimilates WSR-88D data interpolated to a 1 km grid mesh and GOES-16, -17 data every 15 minutes producing 0–6-hour forecasts initialized every hour from 1700 UTC to 0300 UTC (Heinselman et al. 2022). The initial background fields for WoF-Hybrid are similarly derived from the 1400 UTC HRRRDAS 1-hr forecast, while each 0-6-hour forecast gets initial conditions from the WoF-Hybrid analysis fields. The initial boundary conditions, however, are given by only member one of the HRRRDAS 1200 UTC forecast (Clark et al. 2021a). WoF-Hybrid takes background covariances from the 3 km WoFS ensemble and produces a 1.5 km deterministic forecast using the WRF model. Figure 1b depicts the workflow of WoF-Hybrid one-way coupling from the WoFS ensemble.

1.6 Applying WoFS to operations

Over the course of the creation of these two systems, there have been many noteworthy WoFS successes and forecast improvements which support the goal of WoFS becoming operational by 2030. From the 2016 SFE, Lawson et al. (2018) found that WoFS 0-3-hour quantitative precipitation forecasts were better than the now operational High-Resolution Rapid Refresh model (HRRR), most notably at higher thresholds and smaller spatial scales. Skinner et al. (2018) found that system improvements in the HRRRE, used to generate WoFS initial conditions, and implementation of a NSSL double-moment microphysics scheme produced more skillful reflectivity forecasts with fewer false alarms in 2017 than in 2016. Looking at data from the 2017 SFE, Jones et al. (2018) made improvements to the NSSL double-moment microphysics scheme to create more accurate cloud areal coverage depictions while not lessening reflectivity object skill. From the 2018 experiment data, Choate et al. (2018) determined paintball plots, which show either reflectivity or UH objects for each member on a single plot, were by far the most used product by WoFS users. This discovery led to further development of these plots and probabilistic guidance for eventual operational forecast use. Wilson et al. (2019) discusses the importance of how to effectively display WoFS data and teach end users how to use new guidance depictions, which will be important for the correct application of WoFS. Aided by computing power increases, the WoFS domain was increased from a 750 km by 750 km grid to a 900 km by 900 km grid for the 2019 SFE (Clark et al. 2020). One particularly interesting find from Gallo et al. (2022) was that during the 2019 SFE, WoFS UH probabilities didn't consistently improve with decreasing lead time, while human-generated forecasts generally did improve. This implies that even though WoFS guidance may be erroneous at times, the human interpretation of guidance can still produce acceptable forecasts (Gallo et al. 2022). In addition, Wilson et al. (2021) discovered that participants in the 2019 SFE used reflectivity, UH, hail, and surface wind products most often. Users tended to view products in a similar order for events with higher tornado reports, which may be because higher-end events are more straightforward to forecast (Wilson et al. 2021). In the 2020 SFE, a nine-member 1.5 km WoFS ensemble was subjectively compared to the 18-member 3 km WoFS ensemble, and preliminary results show that the majority of participants stated that the 1.5 km WoFS ensemble "might or might not" provide value above the 3 km WoFS Ensemble (Clark et al. 2021b). Further testing the difference in subjective rating between WoFS ensembles of differing number of members (9, 13, and 18) in the 2022 SFE determined that the number of members did not yield a significant difference in forecast quality (Clark et al. 2022). This could be motivation for a smaller ensemble for future WoFS iterations to conserve computational resources, which could then be focused into improving model resolution or data assimilation frequency.

1.7 Previous WoFS Verification work

Forecast verification is the practice of measuring forecast quality and value (value is a separate issue that is not address in this thesis). This can be done both qualitatively, by comparing forecast and observed quantities visually to determine similarities or differences, or quantitatively, by determining forecast accuracy statistically. Qualitative verification metrics are a quick way to spatially assess forecast skill; however, this can be a time-consuming process fraught with interpretation biases. On the other hand, quantitative verification metrics can assess statistical significance without human biases and include calculation of various statistical metrics to see forecast skill through different lenses. This study uses both qualitative verification methods as well as spatial and object-based quantitative verification metrics to compare WoF-Hybrid and WoFS more comprehensively.

It is important to verify model forecasts to better determine model advantages and shortcomings, analyze the quality of the forecasts, and compare various forecast systems. To better understand the framework in which this study operates, it is important to discuss prior results that have used similar methods to analyze 2015-2019 WoFS and WoF-Hybrid real-time forecasts. One of the spatial methods used to analyze these forecasts is the fractions skill score (FSS; Roberts and Lean 2008; see Section 2.1). Lawson et al. (2018) used FSS to compare the operational HRRR to 2016 WoFS rainfall output. WoFS outperformed the HRRR at early lead times, but this difference diminished throughout the forecast period. Miller et al. (2022) determined 2015-2016 WoFS 3-hour forecasts were skillful for higher rainfall thresholds, with many areas directly overlapping with Stage IV verification data. This implies that some strong convection of a certain scale is sufficiently resolved within the ensemble. In addition, the probability of rainfall that meets or exceeds flash flood guidance levels is well matched to areas which received high rainfall amounts (Miller et al. 2022).

Object-based verification methods (discussed in Section 2.2) developed by Skinner et al. (2018) have become one of the most promising methods being applied to convective storm hazards within WoFS over the last few years. Initial results from over 60 WoFS forecasts from 2015-2017 showed that reflectivity objects are better forecast than UH objects due to the high bias of mesocyclone object occurrence (Skinner et al. 2018). In addition, reflectivity forecasts improved from 2016 to 2017, owing to a lower false alarm ratio. Skinner et al. (2018) also demonstrated that this improvement was due to upgrades to the HRRRE, which improved initial conditions in WoFS, and the NSSL double-moment cloud physics scheme, which lowered the high bias seen in 2016 forecasts, yielding more skillful early forecasts. Additionally, this study showed that WoFS produced more accurate forecasts for larger, more intense storms. This discovery is particularly motivating for the operational use of WoFS guidance, as the better prediction of high-end events is one of the goals of the system. In a study by Flora et al. (2019), most of the 2017-2018 WoFS matched UH objects' centroids were within 30 km of the observed objects up to 150 minutes after forecast initialization. Hence, most forecast objects are within a severe warning polygon distance of an observed object, which suggests WoFS output could provide useful guidance for warning operations.

Analyzing 2017 WoFS output further, Jones et al. (2018) presented that the choice of a cloud microphysics scheme has a large effect on forecast skill. Use of the Thompson microphysics scheme showed improvement in upper-level cloud coverage compared to the NSSL double-moment scheme used in the 2017 SFE. This result helped show that a change in cloud microphysics scheme has an impact on forecast skill (Jones et al. 2018). This should be considered in the improvement of WoFS going forwards as the upper-level cloud coverage has impacts on the heating of the boundary layer and placement of thermal boundaries. In analyzing 2017-2021 WoFS cases, Guerra et al. (2022) determined that if, at the time of data assimilation, the storm object is older than one hour, WoFS forecasts yield notably higher object detection probabilities throughout three hours of lead time. This shows that the assimilation of radar and satellite observations of ongoing thunderstorms for multiple model cycles is the key to producing more accurate forecasts of those storms.

A more recent study by Chen et al. (2022) applies object-based verification to 3DEnVAR for the first time. The forecasts produce the highest UH and reflectivity verification scores using the object-based methods for 2017 real-time cases. Forecasts using 3DVAR yielded the lowest scores for the same model variables. A major finding from Chen et al. (2022) was that the ensemble forecasts produced lower false alarm ratios for reflectivity and UH than forecasts generated from variational methods. This implies that rotation is more effectively analyzed, and there is less spurious convection in EnKF initializations, providing evidence to focus on the 3DEnVAR methodology rather than solely variational methods for future deterministic forecasts.

It is important to use multiple verification types to assess the skill of a model forecast, as every method has its limitations. For spatial verification metrics like FSS, if the number of points above a threshold in a forecast match that of the observations in the same area, the model is rewarded. The orientation and shape of the individual objects, however, are not known, and therefore cannot be penalized or rewarded. To combat this, it may be useful to subjectively verify objects to determine more information. Also, while computing FSS over an area, if the radius is too large or small compared to both the resolution or grid size of the model, scores can be skewed or smoothed (Jones 2014). Observed convection on the edges of the forecast domain can lead to errors in FSS due to the model not assimilating or incorrectly locating observed storm objects.

For object-based methods, one of the complications is the flexibility it provides. While this can be a positive since this method can easily be applied to a wide variety of severe hazards, attention must be given to ensure suitable thresholds are chosen for the question being answered (Skinner et al. 2018). Dramatic changes to results can occur if user-defined parameters are incorrectly chosen. It is also important to make sure these parameters are sensitive to the areal coverage of the phenomena being analyzed (Flora et al. 2019). This can be difficult when examining cases with differing storm modes, as mesoscale convective systems span a much larger scale than supercellular events. Another limitation of object-based methods is that contingency table metrics can only analyze ensembles by treating each member as a deterministic forecast (Skinner et al. 2018). This limitation can be overcome by using ensemble verification metrics, but their application is outside the scope of this study. The utilization of object-based methods and FSS provide sufficient avenues for comparing a deterministic forecast system with an ensemble of forecasts.

1.8 Research Questions and Hypothesis

The long-term goal is to help determine a) does the 1.5-km grid spacing WoF-Hybrid system generally provide a more accurate characterization of storm morphology than the 3-km grid spacing WoFS or b) if WoF-Hybrid performance is not generally significantly greater than WoFS performance, does it provide enough additional, useful information for operational forecasters to make its inclusion in a real-time operational modeling scheme worthwhile. To our knowledge, there has yet to be a comprehensive, quantitative analysis and comparison of WoFS and WoF-Hybrid of this scale. The hypothesis is that the WoF-Hybrid system will provide additional information that complements WoFS due to its higher resolution and inclusion of a hybrid variational data assimilation technique, yielding a better ability to predict smaller discrete cells. The verification methods used to address this hypothesis are described in Section 2. Section 3 discusses reflectivity, precipitation, and UH forecasts analyzed subjectively and quantitatively. Finally, Section 4 provides concluding remarks, limitations, and ideas for future work.

2. Data and Methods

Thirty-one cases for which both WoFS and WoF-Hybrid forecasts are available from April-December 2021 are analyzed. Table 1 breaks down each case by the number of storm reports in the forecast domain, the maximum SPC categorical outlook, event severity (which will be defined in detail at the end of this section), and which of the forecast initializations between 1700 and 0300 UTC were not available in both systems for this study. Two verification methods were used to analyze forecast output: a spatial verification method, Fractions Skill Score (FSS), and an object-based method. These methods, and the data used in each are described in further detail in the following sections.

2.1 Fractions Skill Score

2.1.1 Data Used

For the FSS calculation, the six-hour WoFS and WoF-Hybrid forecasts depicted in Table 1 are used. It is important that all data is the same resolution for the FSS calculation, so all observational data and WoF-Hybrid were interpolated to the 3 km WoFS grid with a Lambert Conformal Conical projection. The three variables that are analyzed include reflectivity, one-hour precipitation accumulation, and six-hour precipitation accumulation. For reflectivity, WoFS produces output every 5 minutes, while WoF-Hybrid produces output every 10 minutes. So, for this study, data every 10 minutes is considered for both observational and model reflectivity data. There is no direct output for one-hour and six-hour rainfall accumulation in the two systems. For WoFS (WoF-Hybrid), 5-minute (ten-minute) rainfall data were summed over one hour and six hours for 1- and six-hour forecasts, respectively. MRMS data every 10 minutes was used for observed composite reflectivity by preprocessing with a Cressman filter using a radius of influence of 1 km. For observed one-hour rainfall data, National Centers for Environmental Prediction/Environmental Modeling Center 4 km Stage IV, gauge-corrected one-hour precipitation accumulation data were used. Rainfall accumulation for the entire six-hour forecast period was calculated by summing six one-hour Stage IV gauge-corrected files for each initialization. The precipitation data was then interpolated to the 3 km grid using a neighborhood budget approach to conserve the domain mean of the accumulated rainfall.

2.1.2 FSS Definition

The FSS is a neighborhood spatial verification method that assesses the performance of a forecast against observed values in a predefined spatial window. To calculate the FSS, the fraction of grid points having a higher value than a specified threshold in a circular area within the domain having a predefined radius is determined in both the observed and model data. This is performed for 1- and six-hour precipitation as well as reflectivity. The thresholds and radii used for each are defined in Table 2. Once fractions have been determined, the Mean Squared Error (MSE) is calculated for the observed and forecast fractions

$$MSE = \frac{1}{N_{lon}N_{lat}} \sum_{i=1}^{N_{lon}} \sum_{i=1}^{N_{lat}} \left[O_{ij} - M_{ij}\right]^2, \tag{1}$$

where N_{lon} and N_{lat} are the number of unique latitudes and longitudes associated over all grid points, which are both equal to 300 for this calculation. O_{ij} and M_{ij} are the field of fractions at a specific radius and above a threshold for the observed and model data, respectively. In addition, Robert and Lean (2008) states that it is important to use a reference MSE in the FSS calculation, as the MSE depends strongly on the frequency of the event itself. The reference MSE can be thought of as the MSE of the model fractions added with the MSE of the observed fractions, which produces the largest possible error. The following depicts the final form of the FSS calculated for each threshold and radius:

$$FSS = 1 - \frac{MSE}{MSE_{ref}}. \tag{2}$$

The full derivation of Equation 2 can be seen in Roberts and Lean (2008). To demonstrate the benefit of the FSS, Figure 2 depicts a hypothetical rainfall forecast and corresponding observed rainfall field, both of which indicate rain at 4 of the 16 grid points. Zero of the grid points with rainfall match one-to-one between the two grids. Using traditional grid-based verification, this would yield a forecast with zero skill; however, when the FSS is used, both grids have rainfall occurring in one-fourth of the grid points, which would yield perfect forecast skill. The FSS is a useful statistic for rewarding spatial skill on the domain scale while not penalizing a forecast for grid-scale imperfections. This is valuable for variables that exhibit high spatial intermittency, like reflectivity and precipitation, where the model may correctly forecast a storm in shape and size, but be slightly displaced, yielding low skill in grid-based verification methods.

2.1.3 Bias Correction Method

A method was applied during the FSS calculation to correct for systematic biases within WoFS and WoF-Hybrid reflectivity and precipitation data, similar to the histogram method described in Piani et al. (2010). First, the percentage of model data grid points with a value above a predefined threshold at the first time step for all 31 cases is calculated. This is subtracted from 100% to determine the typical percent coverage of values exceeding this threshold in a sizable sample of the dataset. Then, the histogram at the first time step over all cases is calculated for the first WoFS ensemble member. The data value (reflectivity or precipitation) in the histogram associated with the percentile of the predefined threshold in the observed data is the new threshold for this member in the FSS calculation. This process is repeated for each ensemble member and WoF-Hybrid at the first time step. Then, the observed data threshold percentile, model percentile, and new bias-corrected threshold are calculated for the next time step, and so on, until the end of the forecast period. For reflectivity, this bias correction is applied at ten-minute time intervals for each of the three MRMS thresholds. For precipitation, this method is performed at both one-hour and six-hour time steps for each of the four one-hour and four six-hour Stage IV precipitation data thresholds, respectively.

2.2 Object-Based Verification

2.2.1 Data Used

For object-based verification methods, the same WoFS and WoF-Hybrid cases described for the FSS calculation were analyzed, focusing on composite reflectivity and 2-5 km UH (only mid-level UH is discussed in this study and will be further referred to as solely UH), both at ten-minute intervals. One difference from the FSS method is that WoFS and WoF-Hybrid are analyzed at their original respective resolutions. Ten-minute MRMS composite reflectivity is used to verify model reflectivity. Since UH, the integral of the vertical vorticity (s^{-1}) multiplied by the updraft velocity ($m\ s^{-1}$) between 2 to 5 km above ground level, is not an observable variable, it is verified using proxy, ten-minute MRMS azimuthal wind shear (AWS) values. Both variables were interpolated using a neighborhood budget approach to the 3 km WoFS grid and the 1.5 km WoF-Hybrid grid.

2.2.2 Reflectivity Object Definition

The object-based methodology provides additional insights, measuring the proximity of forecast closed-contour reflectivity and UH swaths of a certain strength to observed swaths of the same or similar strength (by proxy in the case of UH). If an object is deemed “close” (the meaning of this is described herein), it is considered a “matched pair” or hit. In addition, misses and false alarms are calculated, producing contingency table-based statistics. These can then be used as performance metrics for the verification of WoFS and WoF-Hybrid.

For reflectivity, an observed object is defined as a closed-contour composite reflectivity field with values greater than or equal to 40 dBZ. To diminish the impacts of bias as much as possible, reflectivity objects for WoFS and WoF-Hybrid were bias-corrected and defined using a value matching the percentile of 40 dBZ in the MRMS data used. This percentile, 99.30%, corresponds to 46.1 dBZ in WoFS and 47.1 dBZ in WoF-Hybrid, which indicates a high reflectivity bias in both systems. A slightly higher bias seen in WoF-Hybrid (3DEnVAR) compared to WoFS (EnKF) is consistent with results in Kong et al. (2020). Objects within the same time step are labeled with consecutive integers from west to east and have diagnostic information collected. Then, objects smaller than 144 km^2 are filtered out in both systems. This is to remove potential noise that is typically present in high-resolution CAMs. In addition to the requirement that observed objects exceed 40 dBZ, the *maximum* dBZ value within a closed contour must be greater than or equal to 45 dBZ (52.3 dBZ in WoFS and 53.3 dBZ in WoF-Hybrid) to filter out storm modes not related to convection. These storm modes have been seen in previous work by Skinner et al. (2018) with reflectivity values higher than the minimum thresholds defined above for both model and observed data.

In addition to the above filters, a clustering process is used to group objects in proximity into one. This process occurs when the smallest distance between two objects from the same dataset at the same time step is less than 15 km. These new clustered objects are then renumbered with a different integer, and diagnostic properties are recalculated. Clustering objects is especially important for squall line events as a line of convection may have multiple storm objects very close, separated by small areas in which object criteria are not met. Without this clustering process, objects within large convective systems would be over-counted, skewing statistical results.

2.2.3 2-5 km UH/AWS Object Definition

UH and AWS objects are defined in a slightly different manner than reflectivity objects. While reflectivity objects are meaningful at each individual time step, UH/AWS objects are more useful in 30-minute swaths to better detect the presence of a developed mesocyclone. To obtain these swaths, 2-5 km UH data is extracted from 30-minute WoFS and WoF-Hybrid aggregate files. For AWS objects, 5-minute MRMS AWS values are summed over 30 minutes to produce AWS swaths.

For this study, a closed-contour (a set of individual discrete points that exceed a threshold value with no breaks between them), AWS 30-minute swath is considered a noteworthy, persistent mesocyclone and, therefore an object if it has a value greater than 0.004 s^{-1}. This falls within the range of values used in similar research (Skinner et al. 2018; Chen et al. 2022; Miller et al. 2022). WoFS and WoF-Hybrid UH minimum thresholds are calculated similarly to reflectivity using the percentile of the observed AWS threshold. This was between 99.96% and 99.98% for both WoFS and WoF-Hybrid grid MRMS, which led to threshold values of 65 m^2 s^{-2} for WoFS and 130 m^2 s^{-2} for WoF-Hybrid. The higher UH threshold for WoF-Hybrid is due to better resolution of vertical vorticity and updraft speeds at 1.5 km grid spacing, which yields higher extrema for each. Since UH is the integral of the product of these two variables, higher UH values are expected in WoF-Hybrid than WoFS. Unlike reflectivity, there is no maximum threshold required, as these UH thresholds already filter out stratiform storms. Objects that are smaller than 144 km^2 are filtered out, and objects closer than 15 km are clustered together in the same way as reflectivity objects.

2.2.4 Matching Forecast Objects to Observations

Once the final set of reflectivity and UH/AWS objects have been determined whether or not a model object is considered a match to an observed object at the same time step is determined via a Total Interest (TI) score which is defined in Guerra et al. (2022) as the following equation:

$$TI_{nc} = 0.5 * \left[\frac{(D_{match} - D_{min})}{D_{match}} + \frac{(D_{match} - D_{cent})}{D_{match}}\right], \tag{3}$$

where TI_{nc} is the TI score for an object that was not clustered, D_{match} is a predefined distance threshold for object matching, D_{min} is the minimum distance between an object pair, and D_{cent} is the centroid distance between two objects. For this study, D_{match}= 40 km, and TI_{nc} must be greater than 0.2 for an object pair to be considered a match. These values are also used in Skinner et al. (2018) and Guerra et al. (2022) and were chosen to match two objects whose proximity to one another are within the typical scale of an NWS warning polygon. To determine if a clustered object is a match, a slight variation of the TI score is used:

$$TI_c = \frac{(D_{match} - D_{min})}{D_{match}}, \tag{4}$$

where TI_c is the TI score for clustered objects, and D_{match} and D_{min} are defined as in Equation 3. TI_c must also be greater than 0.2 to be considered an object match. A new centroid is not calculated for the newly clustered objects. Therefore, the centroid distance term is removed from the TI score calculation, and it is only dependent on the minimum distance between the two objects.

2.2.5 Verification Metrics

From this matching methodology, contingency table-based statistics, such as the probability of detection (POD; the fraction of the observed objects that are correctly forecast), success ratio (SR), false alarm ratio (FAR, 1-SR; the fraction of forecast objects that were not matched to an observed object), critical success index (CSI; how well the forecast objects correspond to the observed objects), and bias (how the average number of forecast objects relates to the average number of observed objects) can be assessed to determine model skill:

$$\begin{gathered} a)\ POD = \frac{MC}{MC+MS}, b)\ SR = 1 - FAR = \frac{MC}{FA+MC}, \\ c)\ CSI = \frac{MC}{MC+MS+FA} = \frac{1}{\frac{1}{SR}+\frac{1}{POD}-1}, d)\ bias = \frac{MC\ +\ FA}{MC\ +\ MS} \end{gathered} \tag{5}$$

where MC is the matched object count, MS is the missed object count, and FA is the number of false alarms. These variables are further depicted in Table 3 using a contingency table. When using POD, it is important to also show FAR/SR to compare results, as this provides a more comprehensive view of the data. In addition, the POD converges to one (a perfect forecast), as the number of missed objects tends to 0. The FAR converges to 0 when the number of false alarm objects approaches 0. As both the number of false alarm objects and observed objects not forecast, the CSI converges to one, a perfectly skillful forecast. It is important to note also that CSI is a function of both FAR/SR and POD.

Statistics of object diagnostic properties, such as average area ratio (AAR) and average minimum distance (AMD), can be produced for matched objects. To determine the AAR, first, the area covered by each respective forecast object is divided by the area associated with the observed object with which it was matched. Then, the average of all these ratios is taken, which yields the

AAR. For the AMD calculation, minimum distance refers to the shortest line that may be drawn between any one point along the boundary of a forecast object with any one point along the boundary of its matched observed object. The AMD is the average of these minimum distances across all matched sets.

2.2.6 Case Study Overview

Within the object-based framework, three cases that represent a variety of event types are examined more closely: 26 May, 27 May, and 10 December (Figure 3). In the late morning of 26 May, a weak upper-level trough moved eastward into the high plains in combination with a surface warm front moving northward into central Nebraska. Ample surface moisture, 1500-2500 J/kg of CAPE, steep lapse rates, and veering winds with height were present over most of Kansas and western Nebraska which provided an environment primed for supercells and tornadoes. Initial warm-sector convection at 15 UTC strengthened into a long-track supercell that tracked directly along the warm front producing several tornadoes and large hail over central Kansas through the afternoon hours. Numerous other storms initiated during peak afternoon heating from southeastern Wyoming through central Nebraska between 2100 and 2200 UTC. Initial supercells produced a few tornadoes before growing upscale into multiple MCSs and tracking eastward through Nebraska and northern Kansas into the early morning hours of 27 May.

For the second case, the combination of an unstable environment and various boundaries led to the development of severe thunderstorms on 27 May. Multiple MCS outflow boundaries in northern Missouri through western Oklahoma as well as an eastward tracking dryline in western

Texas, provided the formation mechanisms for thunderstorms. Daytime heating led to the depletion of the cap, and over 2000 J/kg of CAPE over western Texas, Oklahoma, Arkansas, southern Missouri, and southern Illinois. In addition, there were steep mid-level lapse rates present due to substantial lift ahead of the deep layer trough. Convection began to strengthen along the outflow boundaries between 1430 UTC and 1530 UTC near the KS/OK border and began to propagate eastward. By 1800 UTC, a large MCS stretching from western Oklahoma to central Illinois had formed and moved into an unstable environment, producing tornadoes, strong winds, and flash flooding. While the MCS continued to move southeastward in Oklahoma, hail-producing supercells were forming east of the dryline in western Texas near 2200 UTC. Cells near western north Texas eventually combined with the MCS at the TX/OK border, while supercells farther south dissipated by 0300 UTC. The MCS weakened in Texas and Louisiana by 10 UTC the next day.

The final case investigated in this study is the tornado outbreak of 10 December. By late afternoon on 10 December, a 500 mb jet with 70-90 kt winds was present over the mid-Mississippi Valley. There was an associated surface low in northern Missouri with a large, pronounced warm sector from Louisiana/Mississippi through Illinois/Indiana. Within the warm sector, there was a strong thermodynamic environment: northerly flow at the surface advecting 65-70 F dewpoints into the area, 7-8 C/km mid-level lapse rates, and 1000-2000 J/kg of MLCAPE. This, with effective storm relative helicity over 400 m^2/s^2, long clockwise turning hodographs, and 60-80 kts of 0-6 km vertical wind shear signaled an environment capable of producing a tornado outbreak. So much so that SPC had forecast an area of moderate risk and 15% tornado probability in the 1630 UTC Day 1 Convective Outlook. At 2000 UTC, weak convection that formed earlier in the day moved northward with the warm front over Missouri and Illinois. By 2300 UTC, convection began to initiate in the warm sector over central Arkansas and southwestern Missouri. Over the next few hours, several long track supercells and QLCSs would produce 66 tornadoes, large hail, and damaging winds from both the mid-Mississippi Valley through northern Kentucky and eastern Missouri into Indiana. The most destructive of these storms produced the "quad-state tornado," an EF-4 that was on the ground for about four hours leading to 60 deaths and over 600 injuries. By 09 UTC on 11 December, convection became a large squall line spanning from E. Texas through Ohio that slowly weakened while moving eastward.

2.2.7 Event Severity Definition

Another way to partition data in the object-based framework is by the severity of the impacts that occurred on the 31 dates studied. Events are divided into categories based on the number of filtered SPC storm reports in the WoFS domain. A date is considered a high-end event if it has at least 50 storm reports, a mid-severity event if the date has between 10 and 50 storm reports, and a low severity event if the date has 10 or fewer storm reports. There are a total of 8 high, 10 mid, and 13 low severity events in this study. Table 4 shows how the partitioning of events based on severity relates to the maximum SPC outlook category within the WoFS domain.

3. Results and Discussion

3.1 Fractions Skill Score

3.1.1 Results

Each of the 31 cases from WoF-Hybrid and WoFS are first analyzed by calculating the FSS for the three reflectivity thresholds. The reflectivity FSS for WoF-Hybrid is calculated at each ten-minute time step for every date and then averaged at each lead time. For WoFS, the FSS is calculated for each individual member first and then is averaged over the 18 members at every date for each individual ten-minute lead time, yielding what is referred to throughout this section as the WoFS member average. Figure 4 depicts the biases in model reflectivity for both systems compared to the observed reflectivity values averaged over all cases. As the dBZ threshold increases, the bias increases tremendously. It can also be seen in WoF-Hybrid at all thresholds that the biases temporarily spike 10 minutes after forecast initialization (second data point). Figure 5 shows one case, the 2200 UTC forecast initialization on 27 May, that spatially depicts this spike in reflectivity bias. Focusing first in western Nebraska there are a few convective cells present in WoF-Hybrid output at 2200 UTC (Figure 5a). By 2210 UTC there is a clear increase in reflectivity intensity represented by more contours over 45 dBZ (Figure 5b), and then a clear decrease in number by 2220 UTC (Figure 5c). In addition, 10 minutes into the forecast there is an increase in reflectivity coverage (at all dBZ thresholds) from the previous step, followed by a corresponding decrease by 2220 UTC. This behavior across the first few time steps is not seen in WoFS. Compared to WoFS member 2 in Figure 5d-f, we see the reflectivity contours above 45 dBZ remain fairly constant over the 20 minutes (Figure 5d-f), and there isn’t much change in spatial coverage over the same time frame. Similarly in central Kansas, WoF-Hybrid output shows there is an increase in spatial coverage of 30-45 dBZ contours at the ten-minute time step that diminishes by the 20-minute time step. This increase, and subsequent decrease, is not seen in the individual WoFS member. This may suggest that the analysis fields among different model variables are not well balanced, and some adjustments are needed in the first few minutes of WRF model integration for WoF-Hybrid.

Moving beyond the ten minute peak artifact in WoF-Hybrid, the difference in bias at the 50 dBZ threshold between the two systems is very notable within the first hour. Thereafter, biases are fairly similar, which was similarly seen in Miller et al. (2022) which compared WoFS with a 1.5 km version. WoF-Hybrid output also has between 6-10 times as many grid points with values above 50 dBZ as the observations, while the average WoFS member biases range from 2.5-7 times as many points at the same threshold. Even at the lower thresholds, there are between 1.5 and 3.5 times as many points as the observations in WoF-Hybrid. This result provides the motivation to use the bias correction method described earlier to calculate FSSs for both systems. Only bias-corrected FSSs are shown in this study, but the raw FSSs have quantitatively similar results.

Figure 6 depicts the reflectivity FSSs averaged over all 2021 cases at each threshold and radius for both WoF-Hybrid and the WoFS member average. First, the FSSs for both systems increase with decreasing dBZ thresholds. In addition, holding the dBZ threshold constant, FSS increases with increasing neighborhood radius. Finally, FSSs decrease with increasing lead time, and decrease much faster within the first hour of lead times. Figure 6 also shows a notable difference in FSS between the WoFS member average and WoF-Hybrid—as high as 0.17 within the first hour of the forecast period—at each radius and threshold, converging to zero near the end of the first hour.

As the WoF concept is in large part a response to the need for accurate short-term guidance in severe weather warning operations, understanding and improving upon forecast performance in the first hour is desirable. This result of poorer WoF-Hybrid performance at the initial time, converging toward WoFS performance at 1 hour, can be illustrated spatially using the 2200 UTC initialization on 27 May seen in Figure 7. In western Kansas it can be seen in Figures 7a and 7d that there is one convective object. At WoF-Hybrid model initialization, seen in Figure 7b, there is a large convective object present with many reflectivity contours over 50 dBZ. This object will cause a notable decrease in FSS for WoF-Hybrid, especially at the higher thresholds. Examining WoFS member 2 in Figure 7e, there is still an increase in reflectivity coverage relative to the verification, but no contours above 50 dBZ are present. This reflectivity representation over western Kansas is very similar in the output from most WoFS members at this time. While this is only one visualization of the difference between the two systems, similar reflectivity output, with WoF-Hybrid containing a greater number of spurious, high-intensity reflectivity cores, is seen over most cases and may be one of the reasons FSS scores are much lower for WoF-Hybrid during the first hour of the forecast period. The FSS results show that near hour one of the forecast period the difference in FSS between the two systems approaches zero. This can be illustrated by examining the 2300 UTC output from the 2200 UTC initialization on 27 May in Figures 7c and 7f. Focusing on the same convection in western Kansas, there is now a notable convective cell present in WoFS member 2 that is of similar intensity to the cell seen in WoF-Hybrid at this time. This depicts one example of how WoF-Hybrid intensifies storms earlier than WoFS in the first hour, which may lead to lower FSSs for WoF-Hybrid. Then, one hour into the forecast period, the cell is represented by both systems. After the first hour, at the majority of time steps at all radii and thresholds–especially at 30 and 40 dBZ thresholds–WoF-Hybrid scores higher than the WoFS member average (Figure 6).

To determine if these differences are significant, Figure 8 depicts the results of a paired t-test between the WoFS member average and WoF-Hybrid at each lead time. In general, the results at most lead times are statistically significant, especially at the largest radius owing to larger differences in FSS. The lead times between 20 and 100 minutes are where most of the differences are not statistically significant. This transition period where WoF-Hybrid performance trends toward and then eventually exceeds WoFS performance yields small differences between the two systems. At 30 and 40 dBZ thresholds, most lead times yield significant results. Of the statistically significant lead times for 30 and 40 dBZ, 80% of them show significant differences between all individual WoFS members and WoF-Hybrid. This indicates that when a difference between the two systems occurs (both WoFS greater than WoF-Hybrid and vice versa), there is a clear, systematic skill difference between the two systems at these lead times, not just between the member average and WoF-Hybrid. For the 50 dBZ threshold, less than half lead times show significant differences between the member average and WoF-Hybrid. In general, WoFS has notably better forecast skill in the first hour, and WoF-Hybrid has better forecast skill after hour two.

FSS was also calculated for both one-hour and six-hour precipitation averaged over 31 cases for WoFS and WoF-Hybrid. Results from the one-hour precipitation were nondescript; both systems scored quite similarly and showed the same bias trends to six-hour precipitation, and therefore are not shown. Results from the six-hour precipitation FSS/bias calculation, however, do produce interesting results**.** Like reflectivity, six-hour precipitation biases were calculated, and are shown in Figure 9. The WoF-Hybrid biases are lower than the average WoFS member bias at every threshold, which is the opposite of what is seen in the reflectivity biases in Figure 4. In addition, WoF-Hybrid biases at the first three thresholds and the average WoFS member biases at the two lowest thresholds are close to one, with both systems having fewer grid points above 0.1 inches than the observations. At the 2.0-inch threshold, biases are close to two for WoF-Hybrid and almost 3.5 for the average WoFS member. Too many large rainfall amounts are produced by the two systems over the 2021 cases, which is similar to reflectivity results.

Figure 10 displays results for the 2021 average bias-corrected FSSs with significance testing results using the paired t-test for WoF-Hybrid and the spread of the FSSs for individual members at each rainfall threshold and radius. Like reflectivity FSS results, at constant radius FSS decreases with increasing threshold and at constant threshold FSS increases with increasing radius of influence. At the 0.1-inch threshold for 12 km radius, the FSSs for all WoFS members and WoF-Hybrid are above 0.5 and are as high as 0.63 at the 24 km radius. The results at higher thresholds, however, are much lower, with FSSs between 0.25 and 0.4 for the 0.5-inch threshold, between 0.12 and 0.25 for the 1.0-inch threshold, and between 0.04 and 0.12 for the 2.0-inch threshold. Comparing the two systems at the three highest thresholds, the WoFS member average and at least 17 of the 18 WoFS members outperform WoF-Hybrid at each radius. At the lowest threshold, however, WoF-Hybrid outperforms 17 of the 18 WoFS members at all radii. For the 24 km radius, the differences are significant for WoF-Hybrid versus WoFS at all thresholds and for WoF-Hybrid versus 13 to 16 members, depending on the threshold (Figure 10c). At the 6 km and 12 km radii, the difference between the WoFS members and WoF-Hybrid is significant for both the WoFS member average and between 10 and 15 members at the three lowest rainfall thresholds, and only significant for 2-4 members for the highest threshold (Figures 10a, b).

3.1.2 Discussion

Determining the systematic reasons for the FSS results from both systems and understanding their implications for forecasters is important to better diagnose the potential operational utility of WoF-Hybrid alongside WoFS. It is seen in all three thresholds and spatially that the high reflectivity bias for WoF-Hybrid peaks 10 minutes into the forecast period and then decreases, eventually plateauing between the first and second hour. One possible reason for this may be the initial balance of microphysical parameters in the NSSL double moment scheme in models with different resolutions and data assimilation algorithms. With a double moment scheme, number concentrations and mixing ratios for rain and hail are both represented by model variables, whereas in a single moment scheme, only mixing ratios are represented by the model. Therefore, in the NSSL double moment scheme, the initial balancing of rain and hail number concentrations may cause these values to vary greatly, which could affect reflectivity calculations. To determine if a difference between the two system resolutions and assimilation algorithms could have an effect on the initial balance of microphysical parameters, future research could test the incremental analysis updating technique (Bloom et al. 1996; Lei and Whitaker 2016). The incremental analysis updating technique would integrate analysis increments into a model more gradually, potentially decreasing the peak in reflectivity seen in WoF-Hybrid throughout the initial balancing of microphysical properties. In addition, this would also lower the number of spurious reflectivity objects. By lowering the number of spurious cells, operational forecasters will more quickly be able to diagnose important threat areas and produce more confident forecasts.

Overall, FSSs decreasing with increasing *threshold* is attributed to increasing forecast difficulty for short-term severe weather forecasts. It is much easier for a model to determine where any convection is going to occur (30 dBZ threshold) rather than predicting exactly where heavier convection will occur (50 dBZ threshold). In addition, FSSs increase with increasing *radii* for both systems. This trend is seen because with an increase in radius of influence, forecast storms that are farther away from observed cells will be included in the FSS calculation, which increases the possibility for higher FSSs via the numerator in the second term of the FSS calculation. It is also shown that initially, the WoFS member average FSS is greater than that of WoF-Hybrid, converging about an hour into the forecast, and then lesser over the majority of remaining time steps. Within the first hour, plan-view case studies show areas where WoF-Hybrid produces convection that is not present in the observations at a noticeably greater rate than does WoFS. This, in combination with the highest biases seen at 10 minutes after forecast initialization, produces excess reflectivity contours at every threshold, which leads to lower FSSs for WoF-Hybrid. The notably higher bias and lower FSSs within the first hour is consistent with results seen in Miller et al. (2022) where a higher reflectivity bias was present for the first 30 minutes in a 1.5 km deterministic forecast using WoF-like member configuration, compared to the 3 km WoFS forecast. Thus, there are now multiple examples of higher resolution WRF-generated forecasts producing more spurious cells than similarly fashioned 3 km forecasts within the first 30-60 minutes of a forecast period. Another potential influence on the FSS and bias results within the first hour could be the resolution differences in MRMS data that is assimilated by each system. WoF-Hybrid assimilates native, 1 km resolution radar data while WoFS assimilates upscaled 5 km resolution radar data, which suggests that there may be differences in observed values being assimilated, which would have an impact on storms generated by the two systems.

After the first 1-1.5 hours, WoF-Hybrid forecasts have slightly higher FSSs than the WoFS member average and are statistically significant at all time steps for 30 and 40 dBZ versus at least 17 WoFS members, indicating greater model skill. This implies that the WoF-Hybrid system as a whole is slightly better at forecasting overall reflectivity coverage than the majority of WoFS members late in the six-hour forecast period. Previous results have also shown subjectively WoF-Hybrid adds value to WoFS in the prediction of storm structure (Clark et al. 2021b), which is at least partially attributed to higher model resolution. In addition, preliminary subjective analysis in this study indicates WoF-Hybrid reflectivity output propagates more similarly to the observations than WoFS members, especially with MCS storm modes. More research is needed as this assertion is only made looking at a small sample of the many real-time WoFS cases.

For verification against NCEP Stage IV rainfall observations, the six-hour precipitation bias values interestingly show average WoFS bias values larger than WoF-Hybrid at all four thresholds, which is the opposite of what is seen in the reflectivity bias. At first glance, this may seem counterintuitive, but previous research by Schwartz and Sobash (2019) has shown similar results, stating that a 1 km CAM produces less areal coverage on average for each precipitation object at every rain rate threshold analyzed compared to a 3 km CAM. This implies that there would be a higher rainfall amount on average in the 3 km CAM, which would yield higher precipitation biases. In addition, WoF-Hybrid has more than twice the number of grid points above 2.0 inches of rainfall, and the average WoFS member has almost 3.5 times as many grid points at this threshold compared to MRMS. This implies that both systems are notably over-forecasting heavy rainfall events, which for WoF-Hybrid is partially due to spurious convection discussed earlier that would produce excessive precipitation does not present in the observations. Other potential errors could be related to estimating rainfall rates from reflectivity values or the cloud microphysics scheme.

Like reflectivity, the lowest thresholds for six-hour precipitation yield the highest FSSs, while higher thresholds show the smallest FSSs. This implies that these two systems are notably more skillful at general coverage of rainfall events compared to pinpointing areas of heavy rainfall, which is expected. One interesting result from the six-hour precipitation FSS calculation is that WoF-Hybrid is more skillful than the WoFS member average at the 0.1-inch threshold and statistically significant with respect to the member average and the majority of individual members. At the 0.5-, 1.0-, and 2.0-inch thresholds, the opposite is true: WoF-Hybrid has lower FSSs than the majority of WoFS members, with statistical significance. For the 2021 cases, one could conclude that WoF-Hybrid provides a more skillful spatial representation of general rainfall amounts, while WoFS members have more skill in producing localized higher rainfall amounts. From this, if the two systems were to be used operationally, WoF-Hybrid may be more useful in the prediction of general storm tracks and locations while WoFS could be a helpful forecasting tool to increase confidence in heavy rainfall and flood-related watches and warnings.

3.2 Object-Based Reflectivity

<u>3.2.1 Results</u>

The total object count per event is shown in Figure 11a for all 6-hour data from WoFS, WoF-Hybrid, 1.5 km MRMS, and 3 km MRMS. Overall, there is a slight increase in total objects per event from the forecasts initialized at 1700UTC to the forecasts initialized at 2000 or 2100 UTC, then a relatively sharp decrease after 2100 UTC. The MRMS observed objects are only about half as many in number as the forecast objects per event at 1700UTC and are close to the number of the forecast objects at 2300 UTC and beyond. In Figure 11b, the number of matched objects from each system is relatively similar over the entire forecast initialization period, with WoF-Hybrid having slightly more at 1700UTC. In addition, there are clear drops of between 10-15 matched objects between 2200-2300 UTC and 0000-0100 UTC. The most common missing initialization times seen in the final column of Table 1 are 2300 and 0100 UTC. Of these, just over half are of high or mid-level event severity. These events tend to have a higher number of objects, which implies an increased likelihood of producing more matched objects. So, even though the objects are averaged per event, these two initialization times may be showing fewer matched objects owing to the omission of many high and mid-level events.

In Figure 11c, a performance diagram is depicted, which displays SR, POD, bias, and CSI averaged over all 31 six-hour forecasts. The PODs range between 0.365 and 0.415 for both WoF-Hybrid and the WoFS member average at all initialization times, except for the 17 UTC initialization for WoF-Hybrid, which has a POD just above 0.45. It is clear that initialization time does not have a large effect on POD difference between the two systems, which is consistent with earlier results as POD is the number of matched objects divided by the total number of MRMS objects, which are both very similar at each initialization for both systems. In addition, WoF-Hybrid POD is notably higher at 1700UTC than the WoFS member average but is similar to the WoFS member average POD for the rest of the period. With SR, however, there is a clear dependence on initialization time for both systems. The earlier initializations (1700-2300 UTC) yield SRs between 0.2 and 0.35, increasing with later lead times. From 0000-0300 UTC, SRs are the highest, with values between 0.35 and 0.45. There is also a difference in SR for both systems with initialization time. At early initialization times (1700-2200 UTC), the SRs are quite similar between the two systems. For 2300 UTC and later initializations, however, WoF-Hybrid has notably higher SRs than the WoFS member average. Finally, CSIs range from 0.17-0.27 with higher values seen at later initialization times. Since CSI is a function of SR and POD, WoF-Hybrid has a higher CSI at 1700UTC due to the notable difference in POD between the systems and is higher at the later initializations because of higher SRs.

Figures 11d-f depict the same information as 11a-c but averaged over all cases including only the first hour of lead times. In Figure 11d, the overall trend of total objects per event is similar between the models and the MRMS data, with object counts low but increasing at early initializations, peaking near 23/00, then decreasing through the rest of the period. While the general trend is similar, there is a high reflectivity object count bias present in both systems for all forecast initializations, with WoF-Hybrid having a higher object count per event than a majority of members through 2300 UTC. It is also encouraging to see, in Figures 11a and 11d, similarities between the object counts for MRMS data of different resolutions, indicating that the interpolation to different grid resolutions is not having an undue effect. Figure 11e depicts a similar matched object count per event for WoF-Hybrid and the WoFS members during the first hour of all forecasts from different initializations, with more matched objects occurring later in the forecast period.

In the final panel, Figure 11f, PODs are above 0.5 for most forecasts from different initialization times. This means that on average over half of the observed objects were correctly forecast in the first hour by both WoF-Hybrid and the WoFS member average. PODs seen in Figure 11f are also much higher than those seen in Figure 11c depicting the six-hour forecast PODs. This is the expected result, as forecasts are more skillful closest to the initialization times. Both systems have a slight improvement in SRs in the first hour compared to six-hour forecasts but are still below 0.5 for every forecast from different initialization times. Biases are on average higher within the one-hour forecasts compared to the six-hour forecasts but follow a similar general trend of decreasing high bias with the increase of initialization times. In addition, CSIs are much higher in the one-hour forecasts than in the six-hour forecast averages, with CSIs ranging between 0.05 and 0.1 higher, which is quite notable considering CSI values are below 0.4 in both forecast subsets in the six-hour forecast category. Comparing the two systems, WoF-Hybrid has notably higher PODs, paired with slightly lower SRs compared to the WoFS member average for the first four initialization times. This yields slightly higher CSIs for WoF-Hybrid at these times. For the remainder of the period, PODs are very similar between these two systems, but for the final three initialization times, WoF-Hybrid has slightly higher SRs. The combination of these two results leads to slightly higher CSIs for WoF-Hybrid.

In addition to the typical contingency table statistics generated by the object-based methodology, two metrics assessing matched spatial object quantities are analyzed. The first of these quantities is the AAR, which can be seen in Figure 12a. An AAR of one indicates, on average, all the matched forecast objects at that time cover the same circular area as the observed objects at the same grid resolution. While there is no direct dependence between AAR and initialization time, WoF-Hybrid has an AAR closer to one than a majority of individual WoFS members at all initializations except for 1700UTC. In addition, WoF-Hybrid matched objects have an AAR closer to one than all WoFS member forecasts from 8 of the 11 initialization times. WoF-Hybrid has an AAR between 0.95 and 1.25 for forecasts from all initialization times, meaning that matched objects are between 5% smaller and 25% larger on average than observed matched objects. On the other hand, the WoFS member average has an AAR between 1.15 and 1.35 for the same period. The largest differences in AAR between WoF-Hybrid and the WoFS member average are seen for the forecasts initialized at 22, 23, and 0000 UTC. Comparing this with the total matched object count per event in Figure 11b, while both systems have a very similar number of matched objects between 2200 and 0000 UTC, the sizes of objects in WoF-Hybrid are more like the sizes of objects seen in MRMS data.

Seen in Figure 12b, the second matched spatial object metric is the AMD. A general trend is that the AMD between the observed objects and the matched forecast objects decreases with the increase of initialization times. This could be attributed to the two systems better representing ongoing convection through rapid data assimilation, an effect that builds over time. In addition, a linear convective mode is more likely to occur in the final initialization times as individual cells congeal. This could be easier for models to resolve, as there tends to be more straightforward coverage and less variation in propagation than in a supercell. Comparing the two systems, the WoF-Hybrid AMD is lower than the WoFS member average at every initialization time, with the difference between the two systems ranging from 0.2 to 1 km. Therefore, WoF-Hybrid produces matched objects that are closer to the observed objects, which implies that these matched objects are better forecast in WoF-Hybrid.

To further analyze POD, FAR, and CSI, Figure 13 shows these statistics partitioned by event severity at each forecast initialization time. First, POD values are highest for high-end events and lowest for low end events for a majority of the period (Figure 13a). High-end events have POD scores ranging between 0.42 and 0.54 for both systems. These POD scores are also higher than the average PODs for each initialization time for both WoF-Hybrid and the WoFS member average seen in Figure 11a. In addition, for the high-end events, WoF-Hybrid has a higher POD for all initializations except 2100 UTC, by as much as 0.07. Interestingly, the opposite is true for the low and mid severity events, with most times yielding lower PODs for WoF-Hybrid, especially for initializations after 0000 UTC. Regarding FAR, which is 1-SR, seen in Figure 13b, the results are similar between the two systems for each event severity. In general, the high-end events produce the lowest FAR and the lowest severity events show the highest FAR for both systems. The FAR for low end events ranges from 0.74 to 0.85 and hovers in that range throughout the period. For the mid and high-end events, the FAR generally decreases over the course of the initialization period, with values for high (mid) severity events starting near 0.7 (0.8) at 1700UTC and ending between 0.5 and 0.57 (0.58 and 0.65). The FAR decreases as convection becomes more widespread and the systems can better understand object propagation, speed, and intensity in these high and mid severity events.

The final contingency table statistic that is analyzed by initialization time is CSI for all 2021 cases partitioned by event severity (Figure 13c). Similar to POD, it can be seen that high-end events have the highest CSI and low severity events have the lowest CSI at all initialization times for both systems. The CSIs of both systems are quite similar in each of the two lower severity categories throughout the majority of the time period with values ranging from 0.07 to 0.15 in the low-end events and 0.15 to 0.26 in mid-severity events. Values in the high-end category are similar through 2100 UTC, but diverge after, with WoF-Hybrid yielding higher values. Over the entire period, CSIs range from 0.23 to 0.29 for the WoFS member average and 0.23 to 0.34 for WoF-Hybrid, with the greatest separation of about 0.07 occurring at the 3 UTC initialization. Overall, in the two highest event severity categories, a generally consistent increase in CSI can be seen with later initialization times.

For the rest of the reflectivity analysis, object-based output is aggregated and analyzed with respect to lead time (Figure 14). This can provide another perspective on the average statistics over the course of a six-hour forecast at all initializations. POD decreases with increasing lead time, with overall values ranging from 0.65 to 0.3 for WoF-Hybrid and 0.58 to 0.25 for individual WoFS members (Figure 14a). There is a notable difference between WoF-Hybrid and all members in the first 20 minutes of the forecast period. Then, after one hour of lead time, the POD for all WoFS members is higher than WoF-Hybrid for the next two hours of lead time. In the final hour of lead times, WoF-Hybrid again has a higher POD than the majority of WoFS members. In Figure 14b, FAR generally increases with increasing lead time, with values starting near 0.4 and ending as high as 0.78 for individual WoFS members and ranging between 0.58 and 0.74 for WoF-Hybrid. This is the expected result, as the further model output is from forecast initialization time, the more likely the output is to deviate from observed values. Again, comparing the two systems, there is a large difference in FAR with WoF-Hybrid being much higher for the zero- to ten-minute lead times. During the rest of the six-hour forecast period, WoF-Hybrid FARs stay below the majority of the WoFS members between 30-120-, and 300-360-minute lead times.

CSIs show very similar scores throughout the first 4.5 hours of lead times between the two systems, decreasing from near 0.35 to 0.17 (Figure 14c). In the final 1.5 hours, WoF-Hybrid CSIs are between 0.01 and 0.03 higher than the WoFS member average, which is notable, given the CSIs are quite small, ranging from 0.15 to 0.19. CSI results show very similar scores between the two systems during the first 5 hours of lead time; however, when viewing the POD and FAR (of which CSI is a function) in Figures 14a and 14b, the scores between the systems are not the same. This highlights the importance of viewing various metrics when analyzing object-based methodology, as one statistic on its own may not tell the full story. The final contingency table statistic analyzed is bias, seen in Figure 14d. In the first 20 minutes, there is a large difference in reflectivity bias between the two systems. For the next 2.5 hours, WoF-Hybrid biases hover near 1.2 while the WoFS member average peaks near 1.6 and decreases to 1.3. After this, bias values for both systems are similar, slightly decreasing to 1.1 by the end of the period. Another important feature to note is the large spread of member bias, with a range of 0.2 to 0.4 for the entire period after the first 30 minutes.

Similar results of AAR and AMD for matched objects can be seen in Figures14e-f binned by lead time. Except for the first 30-40 minutes, WoF-Hybrid has an AAR closer to one than all individual WoFS members for the entirety of the six-hour forecast period averaged over all cases. The WoF-Hybrid AAR ranges from about 1.0 to just over 1.2 after the first hour of lead times while the WoFS members range between 1.15 and 1.7. The AMD increases with increasing lead time, meaning that matched forecast objects are closer to observed objects at earlier lead times. In addition, the AMD for WoF-Hybrid is lower than the WoFS member average by as much as 1 km for all lead times after 10 minutes.

The final reflectivity object-based method contingency table-based results can be seen in Figure 15, partitioning 2021 cases by event severity and binning by lead time. Figures 15a-c show a similar dependence of POD, FAR, and CSI on event severity as forecast initialization time. The higher severity events have the highest PODs and CSIs as well as the lowest FARs. The lowest severity events, on the other hand, yield the lowest PODs and CSIs and highest FARs. For high-end events, the PODs for both WoF-Hybrid and the WoFS member average are above 0.5 through the first 80 and 100 minutes of lead times, respectively. The overall shape of the average POD curves seen in Figure 14a is similar at each event severity. The main difference is that WoF-Hybrid becomes better than the WoFS member average at the end of the six-hour period. This occurs earliest in the high-end events at 170 minutes of lead time, with differences in POD depicted as high as 0.1. In the lowest severity events, however, the WoFS member average POD remains higher than WoF-Hybrid for all lead times after 20 minutes. Another difference can be seen between hours 1 and 3 of lead time, the section where WoF-Hybrid drops below the WoFS member average is the smallest in high-end events. Results in Figure 15b show that FARs at each event severity follow a similar trend in both systems after the first 30 minutes of lead time. In Figure 15c, CSI results show that WoF-Hybrid high-end event forecasts are more skillful than the WoFS member average in the final 3 hours of lead time, when considering hits, misses, and false alarms. Values start at 0.4 and taper to 0.2 for the WoFS member average and just above

0.4 to 0.25 for WoF-Hybrid (Figure 15c). For low-end events, both WoF-Hybrid and the WoFS member average follow a similar trend, starting between 0.22 and 0.25 and ending at 0.08.

3.2.2 Discussion

The object-based reflectivity model forecast skill results give insights into WoFS/WoF-Hybrid strengths and weaknesses regarding operational forecast potential. Figure 11a depicts MRMS data having the highest object counts between 2100 and 0000 UTC initialization times, indicating convection is typically fully developed and ongoing during these six-hour forecasts. Before these initialization times is when the highest forecast reflectivity bias can often be seen in the forecast systems. One thing seen in the 2021 cases is that WoFS and WoF-Hybrid are quite aggressive in the initiation of convection early in the period, perhaps owing to the propensity of ample-CAPE, ample-forcing environments for which most WoFS and WoF-Hybrid real-time experiments are run. In addition, seen in Figure 11c, the SRs in each system increase with later initializations. Previous research by Guerra et al. (2022) depicts one of the potential reasons for these results in this study: WoFS performs better when observed convective objects have been present in the WoFS domain for longer periods of time. This can also explain why high biases are present in the early initialization times compared to the later times. In addition, if convection is in the initial model analysis, then the model doesn’t need to spin up convection on its own, yielding more skillful forecast. By 21-2300 UTC, there are more convective objects to be assimilated allowing the two systems to produce more accurate object counts over the domain.

Comparing the two systems in Figure 11c, WoF-Hybrid yields higher SRs than the WoFS member average at the later initialization times. This may be due to the higher resolution of both

MRMS data and the WoF-Hybrid system as well as the combination of EnKF and 3DVAR data assimilation methods. This may be leading to a better prediction of storm structure and environment when more convection is present at later initialization times. In addition, the storm mode is most typically linear for 2021 cases at these later times. So, another possibility for the difference in SRs is that WoF-Hybrid produces a more accurate propagation of convection with linear storm modes. A final note regarding Figure 11c is that averaged scores seem quite low compared to previous research completed in Skinner et al. (2018). It is important to note that 13 of the 31 cases analyzed in this study are low severity events with 11 of those having marginal risk as the highest SPC Convective Outlook category (Table 1). In Skinner et al. (2018), however, the highest SPC Convective Outlook category was Slight or greater for the majority (31) of the 32 cases analyzed. In Figure 13, low end events produce much lower scores than high-end events. Therefore, this is one reason as to why this study yields lower overall contingency table statistics. Regarding the first hour of reflectivity forecasts seen in Figures 11d-f, over half of the observed events were correctly forecast in the first hour of lead times on average for most initializations. WoFS/WoF-Hybrid produce six-hour forecasts at the top of every hour, and therefore, high forecast object detection in the first hour of lead times is quite important, especially because new six-hour forecasts are generated every hour. There is more information in both systems than just the first hour and having six-hour forecasts are beneficial to see potential outcomes and assess probabilistic guidance but seeing high skill in the first hour is very promising for a CAM ensemble with a goal of predicting severe thunderstorms.

One area of improvement needed in both systems is the reduction of false alarms. SRs in the one-hour forecast, while higher than the entire forecast, are still lower than 0.5. This means over half of the predicted events were not present in the observations. There are a lot of reflectivity objects being produced by both systems that are either incorrectly placed or are in places where convection is not ongoing. Other effects of spurious convection can also be seen in the object-based methods within the first hour through Figures 11a and 11d. In 11a, the high object count is present through the first half of initializations and the bias converges to one by the last few initialization times. In Figure 11d, however, there is a high total object count in both systems within the first hour, being the highest in WoF-Hybrid through 0000 UTC. This again supports the claims of spurious convection being present in both systems prior to convective initiation, moreso in WoF-Hybrid within the first hour, as was seen in the FSS results. In the interim, one potential solution is to tell end users up front in training for operational use that there is a high false alarm rate. This way, forecasters are aware of this when using WoFS/WoF-Hybrid reflectivity guidance to produce warnings and forecasts.

Another way to assess spatial skill that is different from FSS is through spatial object attributes (Figure 12) from the object-based method. Even though both have a similar number of matches throughout the entire period (Figure 11b), objects in WoF-Hybrid are closer in size and distance to the observed objects than those seen in WoFS members. In the FSS calculation, WoF-Hybrid is penalized for false alarms in the beginning of each six-hour forecast, which yields lower FSSs compared to WoFS. But when focusing on only matched objects in the first hour using the object-based method, WoF-Hybrid depicts more skillful representations of reflectivity objects. Size differences and distance between matched reflectivity objects are very important features to diagnose when providing the public with severe weather guidance. A minor difference in either attribute could be the difference between a town experiencing heavy rainfall/severe weather and that same town seeing no rain at all. This is one strength of the higher resolution system, WoF-Hybrid. It has better spatial resolution, and ingests higher resolution MRMS data, while using the same domain as WoFS, which allows the system to better resolve storm structure. This could be one reason for this difference in these attributes between these two systems.

Breaking down forecast initialization times into event severity, it can be seen in Figure 13 that the high-end events have the highest (lowest) PODs and CSIs (FARs) at all but one initialization time. This is important because the high-end events are the cases these two systems are meant for, so it is encouraging that partitioning based on event severity produces these results. Higher scores for high-end events do make sense, as there are more objects to ingest, and better convective environments to understand. Mid/low-end events, on the other hand, may have low severe potential or conditional severe weather environments which are more difficult for models to diagnose. Both systems predicting between 42% and 54% of all observed objects in multiple six-hour forecasts over 8 high-end cases also shows promising skill. This provides more confidence for forecasters using WoFS/WoF-Hybrid guidance for these higher end events. In addition, the better skill scores in the latter half of initializations for WoF-Hybrid high-end events are important to note, as these forecasts are initialized after most of the convection has been initiated. This could imply that once storms have started, WoF-Hybrid better predicts small-scale intricacies of storm structure, which could lead to a more accurate propagation of convection, producing a higher POD/CSI, and lower FAR.

Examining object-based reflectivity binning by lead time offers a perspective on the same data through a different lens. Higher FAR and bias seen in Figure 14 during the first twenty minutes of lead times for WoF-Hybrid compared to WoFS is another result that could support the idea of more spurious convection being present in WoF-Hybrid early in the initialization period. In addition, there is a large spread in the biases for the individual WoFS members. Model spread is one of the strengths of an ensemble of forecasts. When there is a large spread in the number of reflectivity objects, it is most beneficial to identify where these objects overlap between the multiple members, indicating high confidence for the presence of those objects. This way, forecasters can determine which objects are more likely to be false alarms in the members with high biases.

Figure 14 strengthens the case for a systematically more accurate propagation of convection in WoF-Hybrid, first seen when forecasts were binned by initialization time. Figure 14 depicts a higher POD and lower FAR for WoF-Hybrid in the final hour of lead times for each initialization. Looking at the object spatial statistics in Figure 14, AAR for WoF-Hybrid is closer to one than the WoFS member average for all lead times after 30 minutes within the forecast period, like the result seen at each forecast initialization. This result is interesting, as it shows WoF-Hybrid better predicts storm structure and size compared to the WoFS members throughout the majority of the 2021 cases. The only time this is not true is within the first 30 minutes of the six-hour period.

When partitioning by event severity in Figure 15, the high (low) severity events produce the best (worst) PODs, CSIs, and FARs at all lead times after 30 minutes for both systems. This, again, supports the idea that WoFS/WoF-Hybrid have a better handle on events with convective environments marked by ample CAPE and ample forcing. Looking closer at hours one to three in both systems, the WoF-Hybrid POD drops below the WoFS member average POD (Figure 14). When examining by event severity in Figure 15, this difference between the systems is very small for high-end events, between 0 and 0.04 with some lead times showing higher scores for WoF-Hybrid. For the other event severities, the WoFS member average is higher than WoF-Hybrid by between 0 and 0.1. This indicates that, while overall WoF-Hybrid forecasts for reflectivity are less skillful over these time periods, high-end event forecasts by WoF-Hybrid are much closer in skill to the WoFS members than at lower event severities in the middle hours of lead time. Therefore, the large differences in POD between hours one and three seen in Figure 14a at these times can be attributed mostly to the low- and mid-severity events. This, again, supports the claim that WoFS/WoF-Hybrid produce better forecasts for high-end severe events.

3.3 Object-Based UH/AWS

3.3.1 Results

Total objects, matched objects, and contingency table-based results from all cases for six-hour UH WoF-Hybrid and WoFS forecasts of UH binned by initialization time are analyzed in Figures 16a-c. Overall, the number of MRMS AWS objects is highest between 2200 and 0000 UTC with about 25-30 per event, indicating that these six-hour periods contain the most mesocyclone development. Both WoF-Hybrid and the WoFS member average have a similar number of UH objects throughout, with high biases through 2300 UTC and low biases afterwards (Figures 16a). There is quite a large ensemble spread early in the initializations which decreases after 0000 UTC. The matched object counts of both systems show the highest number of matches between 2000 and 0000 UTC initializations, with WoF-Hybrid having on average 1-2 more matches than the WoFS member average (Figure 16b). These times also yield most of the high POD values between 0.25 and 0.35 over the forecast period, seen in the performance diagram (Figure 16c). Overall, PODs between the two systems at the same initialization time are within 0.05 of each other. The SRs are similar at the beginning of the initialization period but are higher for WoF-Hybrid than the WoFS member average from 2000-0300 UTC initializations (Figure 16c). Overall CSIs are similar, with larger scores for WoF-Hybrid compared to the WoFS member average mainly attributed to the difference in FAR between the systems.

Next, the same statistics are analyzed for the first 90 minutes of lead times (which is one hour's worth of 30-minute UH swaths, since there are no UH objects before 30 minutes) from each six-hour WoFS/WoF-Hybrid forecast (Figures 16d-f). The number of objects per event in the first hour at each initialization time is similar for both systems, throughout the period. The largest increase in observed UH object number occurs between 2200 and 0100 UTC, while the largest increase in objects for the models is between 2000 and 2200 UTC. Depicted in Figure 16e, object matches are again similar between the two systems, increasing to nearly three matches on average at 2300 UTC in the first 90 minutes of all forecasts. Looking at the performance diagram for the first 90 minutes of UH swaths, statistics are notably higher than over the six-hour period, with PODs ranging between 0.25 and 0.55 after the first two initialization times. SRs are in a similar range as the six-hour forecast and are higher for WoF-Hybrid compared to the WoFS member average at most initialization times. These statistics need to be viewed with caution, however, as there are between 0.5 and 11 objects per initialization, so one more or one fewer objects matched has a large effect on these statistics.

In addition to the general contingency table-based statistics, the spatial object properties of AAR and AMD are analyzed in Figure 17 for AWS/UH matched object pairs averaged over all 31 cases and binned by initialization time. WoF-Hybrid AARs start near one at the 1700UTC initialization and then increase, peaking near two by the 2100 UTC run. For the rest of the initializations after 2100 UTC, the AAR ranges from just under one to 1.5. The WoFS members follow a slightly different general pattern. The initial increase from the 1700UTC to 2200 UTC initialization times is very similar to WoF-Hybrid, with values for the WoFS member average between 0.1 and 0.4 lower than WoF-Hybrid. After the 2200 UTC run, however, the AAR for WoFS member average hovers between just under 1.5 and 1.8. At these final 6 initializations, WoF-Hybrid is much closer to 1, with AARs between 0.2 and 0.6 lower than the WoFS member average. In addition, WoF-Hybrid AAR is closer to one than all WoFS members during the 2300-0100 UTC forecast runs. For AMD, there is a large WoFS ensemble spread over the first and last few initialization times (Figure 17b). This can be attributed to the small number of matched objects present at these times, as outliers have a much larger effect. At 8 of the 11 forecast runs, the AMD of WoF-Hybrid objects is within the WoFS ensemble spread, and lower than the WoFS member average.

The average POD, FAR, and CSI over all cases for each initialization partitioned by event severity are depicted in Figure 18. Unlike for reflectivity event severity (Figure 13), the low-end events are omitted from each panel in Figure 18 as the POD and CSI are zero and the FAR is one for each initialization time. This occurs because these low-end events have very few, if any, developed mesocyclones, which results in the worst score for each statistic. Focusing on the mid and high-end events, high-end events have higher PODs and CSIs and lowest FARs than the mid severity events at every initialization for both systems. Differences between WoF-Hybrid and WoFS statistics for the two severity categories range between 0.01 to 0.25. Comparing the two systems, WoF-Hybrid has a higher POD and CSI (lower FAR) for high-end events than the WoFS member average from 2000-0100 UTC. The difference between the statistics of the two systems at these times is as large as 0.1, which is quite large as a percentage of the absolute score.

Like object-based reflectivity, UH/AWS average results over all cases can be viewed with respect to lead time, displayed in Figure 19. PODs start near 0.3 for both systems, increase over the first hour while mesocyclones develop from initial cells, and then peak near 0.4 at between 1.0 and 1.5 hours. Throughout the rest of the period, PODs generally decrease nearing between 0.1 and 0.2. During the first three hours of lead time, WoF-Hybrid PODs are most often at the bottom edge of the WoFS member spread. After the third hour, however, WoF-Hybrid PODs are above the majority of members, higher than the WoFS member average by as much as 0.09. Looking at FAR against lead time in Figure 19b, WoF-Hybrid has a lower FAR compared to the WoFS member average at all but one lead time by as much as 0.08 (Figure 19b). Moreover, it is lower than all individual WoFS members at just under half of the lead times. These two metrics are combined through the CSI, which is seen in Figure 19c. At all but 4 of the 37 lead times, WoF-Hybrid yields a higher CSI, with a larger difference seen in the final three hours of lead times. In general, CSIs drop from near 0.5 to about 0.2 over the six-hour lead time, which is due to both a decreasing POD and increasing FAR as forecasts deviate further from the observed values.

Bias values seen in Figure 19d increase rapidly from 0.8 to above 1.2 in both systems by the end of the first hour of lead times. This occurs because of mesocyclone development as storms strengthen in both model environments just after initialization. After one hour, the bias values in the individual WoFS members continue to increase to as high as 1.9, while they remain under 1.4 for WoF-Hybrid. From this point onward in the six-hour forecast, the individual WoFS members depict a large spread of 0.4 and members with low (high) biases stay near the bottom (top) of the spread throughout. Overall, the biases of WoFS members decrease throughout the period after 1.5 hours of lead time from between 1.5-1.9 to about 0.8-1.2. WoF-Hybrid, on the other hand, produces more consistent biases between 1.1 and 1.4 for all lead times after one hour.

To further analyze UH results, the matched object spatial statistics of AAR and AMD are depicted in Figures 19e-f. As mesocyclones develop in the first hour, the sizes of the matched objects increase with respect to observed objects for both systems. Between hour one and hour two, the AAR is between 0.95 and 1.25 for WoF-Hybrid while it hovers just above 1.5 for the WoFS member average. After the second hour of lead times, the AAR of WoF-Hybrid is within the WoFS member spread and AARs for both systems range from about 1.0 to 3.0. The AMD for WoF-Hybrid and the WoFS member average matched objects generally increases with increasing lead time, starting between 1-3 km and ending near 6.5 km. This result is expected as matched objects should be closer together at the beginning of the forecast since the initial forecasts contain more information from the observations. The AMD between both systems is similar throughout, with most times for WoF-Hybrid yielding AMD values toward the bottom of WoFS member spread (with lower values being desirable).

The final analysis for UH/AWS contingency table-based statistics is detailed by partitioning cases by event severity, seen in Figure 20. Low-end events are omitted as they produce constant values of zero (one) for POD and CSI (FAR). While there was a clear dependence of POD on event severity within each system in the object-based reflectivity results, the UH/AWS POD results are not as straightforward (Figure 20a). Within the first two hours, the WoFS member average PODs for mid-severity events are greater than for high-end events for over half of the lead times. For the rest of the time period, however, the high-end events yield larger PODs than the mid-severity events by at least 0.05 and as high as 0.12. For WoF-Hybrid, the high-end events produce notably higher PODs than mid-severity events at all lead times by between 0.08 and 0.17. Comparing the two systems, the high-end events yield similar scores through the first 3.5 hours, and then WoF-Hybrid produces higher PODs than the WoFS member average. For mid-severity events, the WoFS member average outperforms WoF-Hybrid for the first 2-3 hours of lead time, with the opposite occurring to a lesser extent in the final three hours of lead time.

For both systems, there is a very high FAR ranging between 0.56 and 0.95 that increases with increasing lead time (Figure 20b). These values are much higher than the POD values at the same lead times, indicating a very large ratio of false alarms to correctly forecast objects. These values are higher than the object-based reflectivity FAR seen in Figure 15b by around 0.1. High-end events have the lowest FARs at all lead times, with WoF-Hybrid during high-end events yielding a smaller FAR than the WoFS member average. The CSI values, representing bulk skills, are decreasing with increasing lead time at each event severity (Figure 20c). Again, the highest CSIs are with the high-end events, with WoF-Hybrid high-end event PODs greater than the WoFS member average at all but two lead times. The difference between the two system

PODs at the mid-severity events in the first two hours can be attributed to both lower POD and higher FAR in WoF-Hybrid at these times. CSIs then transition to WoF-Hybrid having higher PODs than the WoFS member average, which is owing to generally lower FARs and higher PODs in the final three hours for WoF-Hybrid.

3.3.2 Discussion

Overall, UH forecasts are high in false alarms, and low in CSI and POD. In general, CSIs both binned by initialization (Figure 16c) and lead time (Figures 19c) for six-hour forecasts, are about 0.1-.25 lower on average than object-based reflectivity results (Figures 11c and 14c). This difference between UH and reflectivity is similar to previous research in Skinner et al. (2018). It can be at least partially attributed to UH objects being less frequent events than composite reflectivity objects. Since CSI calculations are based on the frequency of an event, rarer events are more likely to produce lower CSIs.

When binning by initialization time, both systems yield a high total UH object count during the first few initialization times (Figure 16a). These times are mainly earlier than or at the beginning of convective initiation, , and thus prior to a ramp up of AWS objects in the verification dataset. This further supports for the claim made in the object-based reflectivity results sections that WoFS and WoF-Hybrid produce better forecasts after convective objects have been ingested by the data assimilation techniques. In addition, the highest number of matches for both systems occur between 2000-0000 UTC (Figure 16b), which aligns well with the times containing the most observed objects seen in Figure 16a.

For periods covering the first 90 minutes of these forecasts, at the hourly initialization times studies here, an increase in observed UH objects is most notable for forecasts initialized between 2200 and 0100 UTC (Figure 16d). This indicates that most mesocyclone development is occurring between these hours. It is shifted one-two hours later than the sharpest increase in observed reflectivity objects seen in Figure 11d. This could provide evidence of the time lag between the presence of a reflectivity object and a developed mesocyclone. In addition, the quickest increase in UH object count occurs in both systems between 2000 and 2200 UTC. One possible explanation for this is that WoF-Hybrid and WoFS are producing rotating updrafts too early in the forecasts. Another explanation is that the comparison of a UH threshold to an AWS threshold will not be exact. Different UH/AWS threshold pairings should be tested to determine if this trend is consistently depicted. Like what was seen in the first 60 minutes of the object-based reflectivity forecasts, the number of UH matches reaches its highest level for both systems after most of the convection has been initiated. This further demonstrates the results seen in Guerra et al. (2022) that WoFS produces better forecasts with an increasingly large number of data assimilation cycles completed for a given storm object.

When analyzing the spatial object properties for the average of all cases binned by lead time, an interesting result is depicted in the AAR data. The AAR is slightly higher for WoF-Hybrid than the majority of WoFS members (AAR for both systems is near or above 1) through the 2100 UTC initialization as seen in Figure 17a. After this time, however, WoF-Hybrid AAR is notably lower than the WoFS member average–and lower than all WoFS members between 2300-0100 UTC. In addition, Figure 16d shows that the largest increase in observed AWS objects occurs between 2200 and 0000 UTC. Once these observed AWS objects are reasonably forecast by the two systems, WoF-Hybrid UH matched objects are much closer in size, supporting the idea that the higher resolution WoF-Hybrid system has a better grasp on storm size and structure compared to WoFS at the most important mesocyclone development stages.

Like reflectivity object lead time analysis, it can be seen early in each six-hour forecast that WoF-Hybrid has a lower POD than WoFS, and the opposite occurs later in the forecast period. In Figures 19b-c, the final three hours also show the biggest differences in favor of WoF-Hybrid in FAR and CSI. These metrics again support the idea stated earlier that higher resolution MRMS data and a higher resolution data assimilation system could allow for a better prediction of storm structure and storm propagation. Another potential reason for this difference in skill at later initialization times may be WoF-Hybrid accurately maintaining intense updrafts longer than WoFS members. More evidence of this is present in the biases seen in Figure 19d. After 1.5 hours of lead time, the bias values decrease consistently throughout the forecast period in the WoFS members, with values falling from near 1.7 to just under 1.0 in the WoFS member average. WoF-Hybrid biases, however, hover between 1.1 and 1.4 throughout this period. A more constant bias indicates that, while a high bias is occurring, the WoF-Hybrid UH objects are either increasing or decreasing in size at a proportional rate to the observed objects. This implies that WoFS members are either not maintaining updrafts long enough or are not producing new updrafts at a proportional rate to the observed values. A more comprehensive dive into each real-time case is necessary to determine if WoF-Hybrid better maintains intense updrafts throughout the forecast period.

Partitioning by event severity, high-end events produce the most skillful UH forecasts in both systems, yielding the best POD, CSI, and FAR values. This is seen for forecasts at all times when binning by initialization time (Figure 18), and at a majority of times when binning by ten-minute lead times (Figure 20), again supporting what was seen in reflectivity results. These statistics indicate that high-end events are more easily predicted by WoF-Hybrid and WoFS, most likely attributable to stronger, more easily trackable storms in environments very supportive of robust convection. Per initialization time, WoF-Hybrid has the biggest positive difference in POD and CSI generally from 2000 to 0100 UTC in general. These are also the five six-hour forecasts in which convection is typically mature and ongoing, indicating that WoF-Hybrid is producing more skillful forecasts during the most impactful hours of severe weather events. When binning by lead time, WoF-Hybrid high-end events produce higher (lower) PODs and CSIs (FARs) than the WoFS member average in the final 2-3 hours. This could also provide evidence of WoF-Hybrid more accurately propagating convective cells throughout the six-hour forecast period. In addition, this could indicate that WoF-Hybrid is maintaining updraft strength more accurately than WoFS through better handling of storm structure at higher resolution.

3.4 Case Studies

3.4.1 26 May

The first case study date, 26 May, is a typical great plains severe weather event with initial supercells that grow upscale into multiple MCSs during the latter half of the model initialization period. Figure 21 depicts reflectivity and UH/AWS object-based statistics for forecasts from each initialization time on 26 May. In general, the total object counts for reflectivity between both systems and MRMS observed objects are similar through 2200 UTC. Throughout this time period, the number of objects over each initialization time increases from nearly 200 to about 500 objects. After this initialization time, WoFS members depict a notable low bias by as much as 100 objects, while WoF-Hybrid stays closer to the observed number of total objects (Figure 21a). The matched object count follows the same general pattern of the total object count (Figure 21b). Looking at the reflectivity performance diagram, PODs and SRs generally increase with later initialization times with SRs for both systems being near 0.2/0.3 at the beginning of the period and ending near 0.6. PODs, on the other hand, are different between the two systems (Figure 21c). WoF-Hybrid has higher PODs than the WoFS member average for all but one initialization time, and by as much as 0.17 by 0300 UTC. This difference in POD is the driving force behind higher CSIs for WoF-Hybrid at all six-hour reflectivity forecasts for 26 May. The higher PODs and similar SRs for WoF-Hybrid compared to the WoFS member average shows that a larger percentage of reflectivity objects are being correctly matched in WoF-Hybrid over the course of this event, with a similar percentage of false alarms.

Overall, WoF-Hybrid produces higher CSIs and therefore more skillful forecasts than the WoFS member average for this case study. In addition, after 2200 UTC, PODs for the entire six-hour forecast are above 0.5 for WoF-Hybrid, while the WoFS member average does not get above 0.5 at any initialization time. SRs are above 0.5 after 2300 UTC for the WoFS member average and after 0100 UTC for WoF-Hybrid (Figure 21c). Both scores are much higher than the average six-hour reflectivity forecasts seen in Figure 11c, which yield PODs/SRs only as high as 0.45. In addition, having both PODs above 0.5 at the later initialization times indicates that over half of observed reflectivity objects are matched in the forecast. Having these values present over an entire six-hour forecast indicates very good skill in both the systems, depicting promising results for operational value, particularly after convective initiation and during upscale growth. In the 26 May event, the storm mode during the later six-hour forecasts was mainly linear, supporting earlier claims that both systems produce more skillful forecasts with linear storm modes.

UH/AWS object-based results for 26 May (Figures 21d-f) however, tell a different story. Unlike reflectivity results, there is a very high bias in UH objects compared to observed AWS objects in both systems. For the first seven initialization times, observed mesocyclone development slowly increases with later initialization times from near 25 AWS objects to 150. In both model systems, however, the number of UH objects starts near 90-135 and peaks near 250 objects (Figures 21d). This may be indicating that the two systems are intensifying updrafts too quickly for this event. After the 2300 UTC initialization, the number of UH objects for both systems is very close to the number of observed AWS objects. Like reflectivity results, the number of UH matched objects increases over the first seven initialization times. During the first four initialization times, WoF-Hybrid has double the matched UH objects as the WoFS member average, while having about 1.5 times as many objects (Figure 21e). This yields the notably higher WoF-Hybrid PODs from 1700-2000 UTC seen in the performance diagram (Figure 21f). High PODs paired with similar SRs indicate WoF-Hybrid has a better handle on mesocyclone development at the early forecast hours of this event. During these early forecasts, WoF-Hybrid may be benefiting from assimilating higher resolution of radar data, allowing the system to better depict storm structure and small-scale intricacies that may not be present in coarser resolution radar data used in WoFS leading to higher PODs/CSIs. After 0000 UTC, however, the WoFS member average yields higher PODs and SRs by as much as 0.2 compared to WoF-Hybrid (Figures 21f). While both systems have a similar number of objects at these initialization times, the WoFS member average has a higher number of matches, leading to higher PODs and SRs. Like reflectivity, most initializations for this case yield higher PODs and SRs than are seen on average over all 2021 cases as seen in Figure 16c.

To take a deeper dive into these two variables on 26 May, the times with the most impactful convection are analyzed. There are two observed supercells that would produce severe weather reports near the Kansas/Nebraska border around 2230 UTC: one that produces multiple tornado reports (producing an EF-2 tornado) from 2130 until 2300 UTC near 40°N 100.9°W and another supercell near 38.6°N 98.4°W that produces severe winds just before 2300 UTC (Figure 22). First focusing on the supercell in the upper left corner of each panel in Figure 22, it can be seen in the first five forecasts initialized at 1700, 1800, 1900, 2000, and 2100 UTC, the WoFS member with the highest CSI across the full domain (referred to as the “best member”) does not produce reflectivity objects at 2230 UTC for this supercell. WoF-Hybrid on the other hand, produces smaller cells near the observed contour in most time steps, but nothing of similar size to the observed object. In the forecast initialized at 2100 UTC, WoF-Hybrid places a strong reflectivity object at the far northern edge of the object (Figure 22e). This is a large success for WoF-Hybrid, as at the forecast initialized at 2100 UTC, there was only weak convection under 40 dBZ in the MRMS data, but the cell quickly grew larger within the next half hour. WoF-Hybrid locked on to this small area of weak convection and predicted its development. The cell, however, propagated too quickly to the northeast in WoF-Hybrid, which leads to the slight difference in exact placement, but overall, the forecast and observed objects are still considered a match. While the best member did not have this cell in the forecasts initialized at all up through

2100 UTC, it is important to note that six of the 18 WoFS members did have a supercell present at the 2100 UTC initialization.

At the half hour forecast initialized at 2200 UTC, both the best member and WoF-Hybrid have reflectivity contours present for the supercell at the Kansas/Nebraska border (Figure 22f). The best member has one larger area of convection while WoF-Hybrid depicts two areas of convection above the object threshold. The two cells in WoF-hybrid are connected by a small area of reflectivity below 45 dBZ (not shown). This is more correct compared to the observed value as it better captures the structure seen. Over the previous two initializations (2000 and 2100 UTC), the supercell was beginning to split in WoF-Hybrid, which is what occurs in the observed data just after 2230 UTC, with the right supercell going on to produce several tornado reports over the next hour. The majority of WoFS members, however, predict a non-splitting supercell that propagates eastward. In general, WoF-Hybrid better captures the structure and propagation in the forecasts than a majority of WoFS members, and better than the best member. For this EF-2 tornado-producing supercell, the higher resolution WoF-Hybrid provided a large benefit, correctly forecasting a strong reflectivity object before a majority of the members. The dBZ object may have been too small for the coarser WoFS to capture.

Looking at the second supercell in the lower right-hand corner of each panel in Figure 22, both systems have a better handle on this supercell over the course of forecasts from six initialization times. For the forecast initialized at 1700UTC, WoF-Hybrid produces intense convection with a slight northwestward bias, while the best WoFS member has nothing present. It is important to note that just under half of the 18 members did place convection here, with a northwestward bias similar to WoF-Hybrid. For the forecast initialized at 1800 UTC, both systems placed convection in the correct area. Unlike the first supercell analyzed, this cell initiated at 1740 UTC, and therefore, is forecast from the first available model runs, those initialized at 1700UTC. With each forecast, both systems produce a supercell that is closer in object size and placement to the observed reflectivity object. For the forecast initialized at 2100 UTC, WoF-Hybrid produces multiple reflectivity objects that are spot on with the observed objects, while the best member has an eastward bias. In addition, all WoFS members have convection present in this area, but most of them yield an eastward bias, although five members produce accurately placed objects (Figure 22e). It is important to note, however, all WoFS members and WoF-Hybrid will produce object matches for this supercell in the object-based methodology.

The analysis of these two supercells on 26 May provide further evidence that WoFS/WoF-Hybrid produces better forecasts the longer an object is present in the observations, which was depicted in Guerra et al. (2022). With the first supercell, some WoFS members and WoF-hybrid detect the potential for convection around this supercell, but nothing accurate in shape or size until the cell was forecast in the 2200 UTC initialization, only 30 minutes prior to tornadogenesis. The second supercell was predicted with 4.5 hours of lead time in the forecast initialized at 1800 UTC (after storm initiation), and convection is present in both systems near the observed cell. With each forecast initialized at later times, the propagation of the supercell and object area are closer to the observed reflectivity object. It is encouraging that both systems accurately assimilate and then maintain this supercell. That the forecasts should improve with decreasing lead time is not surprising and is again consistent with the analysis of Figure 14.

For UH, swaths are depicted from 2300-0000 UTC on 26-27 May, to analyze WoFS and WoF-Hybrid mesocyclone development (Figure 23). Overall, the skill of each system, depicted through CSI over the entire domain, increases with decreasing lead time. Forecasts from the first two initializations yield very low CSIs, all under 0.2. By 2100 UTC, WoF-Hybrid and the best scoring WoFS member earned a CSI near 0.3, while the average scoring member was close to zero, indicating a large WoFS member spread in the 2100 UTC run for forecasts valid at 2300-0000 UTC. Finally, at 2300 UTC, WoF-Hybrid is over 0.1 higher than the best scoring member and almost double the average scoring member. Low CSIs throughout the period, in general, can be attributed to the many UH false alarms. There are many strong updraft objects being predicted in western Kansas, northeastern Colorado, and Nebraska Panhandle, especially by the two WoFS members depicted, that are not associated with severe storm reports.

Examining more closely the forecast from 1800 UTC, there is an observed supercell over central Kansas that produced many hail and tornado reports over the hour depicted. The best WoFS member and WoF-Hybrid UH forecasts exhibit great skill in the propagation of the supercell over the course of the hour, with both systems producing UH swaths over almost the entire area of storm reports associated with this object (Figure 23a). The average scoring member only yields a small UH object over one storm report (Figure 23a). By the 2100 UTC run, both WoFS members and WoF-Hybrid produce UH swaths near the storm reports displayed and do so throughout the remaining initializations. This convective object had been assimilated for many hours prior to the final three initializations, and the forecasts produced by both systems further shows what is discussed in Guerra et al. (2022) that the more assimilation cycles completed on an observed object, the higher the skill of the two systems.

While WoFS/WoF-Hybrid produced excellent forecasts of the supercell in central Kansas, the UH forecasts of the tornado and hail reports on the Kansas/Nebraska border tell a different story. None of the depicted UH forecasts detect swaths at the border at 1800 and 1900 UTC initializations (Figures 23a and 23b). By 2000 UTC, WoF-Hybrid has a UH swath to the east of the storm reports, while the average scoring member produces a small swath to the southwest of the reports (Figure 23c). At the forecast from 2100 UTC, all three forecasts depict strong UH swaths to the north and west of the storm reports (Figure 23d). This is because, near 2230 UTC the supercell splits and the rightward moving cell produces severe weather at the Kansas/Nebraska border. All three systems did not forecast the splitting of this supercell at this time and therefore yield a northward bias. By 2200 UTC, however, the two WoFS members predict the correct propagation direction of the split supercell, but still contain a northward bias (Figure 23e). In this initialization, WoF-Hybrid does not predict a strong supercell in this area. At 2300 UTC, all systems accurately predict the motion of this supercell, correctly creating UH swaths over the storm reports (Figure 23f). Overall, the two tornado clusters near the Nebraska/Kansas border are represented slightly earlier in the two WoFS members than in WoF-Hybrid.

3.4.2 27 May

The next case study analyzed in this study is 27 May, which produced multiple MCSs that grew upscale into a large MCS tracking southeastward over Oklahoma. The MCS produced flash flooding in the Oklahoma City metropolitan area. Because the focus of this event is flash flooding, reflectivity and six-hour precipitation are analyzed. The object-based methodology provides many metrics with which to analyze this case, seen in Figure 24. In the first panel, the total object count versus initialization time for 27 May is displayed. In general, there is a very high bias in reflectivity objects in both systems, especially from 1700-2100 UTC. The bias decreases over the initialization period, with the total number of objects being similar between the observations and both systems by 0200 UTC. The high biases in the beginning of the forecast period can be attributed to intense convection being produced by both systems near the main line of convection, and in west central Texas. During the first few forecasts, convection was not organized and many outflow boundaries were present, making it difficult for the model to lock onto where exactly convection would initiate. In addition, reflectivity objects initiate in west central Texas much earlier in both systems than what was observed.

The number of matches, seen in Figure 24b, are near 200 at 1700UTC (Figure 24b). This number drops to near 80 by 0300 UTC as the number of total objects decreases to near 80, with both WoF-Hybrid and the WoFS member average yielding similar matched object counts throughout. The forecast objects match the observed objects with a large percentage, leading to very high POD values for both systems (Figure 24c). PODs are above 0.6 through 0200 UTC, and lower to near 0.5 by 0300 UTC, indicating that over half of the observed events are correctly forecast by both systems. This is between 0.1 and 0.4 higher than the 2021 average PODs seen in Figure 11c and most forecasts in the 26 May case study. This provides evidence to the idea that CAM systems can more skillfully predict MCS events compared to supercell events, as there is less uncertainty involved in forecasting where convection will occur.

A weakness of both systems for this event is the number of false alarms over the domain. At 1700UTC, there are 650 (590) objects present in WoF-Hybrid (WoFS member average), with 195 (197) matched objects. This indicates that there are 455 (393) false alarms in WoF-Hybrid (WoFS member average), which produces very low SRs near 0.3 for both systems. Throughout much of the period, SRs are similar for the WoFS member average and WoF-Hybrid, but by the end of the period there is a notable difference between the two. At 0200 and 0300 UTC, WoF-Hybrid SRs are between 0.15 and 0.2 higher than the WoFS member average, with similar PODs. This, along with higher CSIs, implies that WoF-Hybrid forecasts are more skillful at these final initialization times. This difference can be explained by the fact that WoF-Hybrid produces a more accurate propagation of the main line of convection, while most WoFS members produce an MCS that propagates too slowly. In addition, some WoFS members maintain convection in west central Texas longer than what was observed, leading to a higher number of false alarms.

One spatial example for the very high PODs, SRs, and CSIs for this event overall can be seen in Figure 25, which depicts the main MCS's peak severity at 0030 UTC over six forecasts started from at the top of hour from 1800 UTC 27 May to 0000 UTC 28 May. Both systems performed exceptionally well in the placement and coverage of the main heavy rainfall threat at 0030 UTC, even with 5.5 hours of lead time at the forecast from 1900 UTC. This level of skill was also seen in 17 of the 18 members at the same period of the forecast. Additionally, this reflectivity being accurately forecast by both systems is impressive, because at the 1900 UTC initialization, convection was overall less organized, especially in central Oklahoma where several storm clusters were developing with differing propagation directions. Despite this, both systems had a good representation of the environment to produce upscale growth and one MCS in the correct location 5.5 hours prior. In general, throughout the six forecasts from different initialization times, every panel in Figure 25 shows great forecast accuracy with the MCS at peak strength.

The main severe threat for 27 May was flash flooding, so it is important to spatially analyze six-hour rainfall through 0000 UTC. In Figure 26a, the areas with over one inch of rainfall are depicted for WoF-Hybrid, the highest scoring WoFS member, lowest scoring WoFS member, and stage IV gauge corrected rainfall. Overall, FSSs using a 24 km radius and one inch threshold are very high, with the ensemble ranging from 0.7216 to 0.8344 and WoF-Hybrid at 0.7484, which is notably higher than the average of between 0.2 and 0.26 for all cases seen in Figure 10c. Spatially, the observed rainfall above one inch is well covered by six-hour rainfall products from each system. The difference in skill between WoF-Hybrid, the best member, and the worst member is due to areas where there are forecast rainfall amounts over one inch that are not present in the observations. All systems produce higher rainfall amounts in west central Texas that are not present in the stage IV data. Behind the main line of convection that produced heavy rainfall, however, is where the differences between two systems are present. WoF-Hybrid produces several areas of heavy rainfall that are not observed, while the two WoFS members shown do not. Additionally, WoF-Hybrid and the worst scoring WoFS member show rainfall greater than one inch out in front of the main line of convection in eastern Oklahoma and northern Arkansas. Besides small areas of incorrect heavy rainfall placement, the main convection in the event was well predicted in coverage.

Looking at the location of actual rainfall values over one inch in Figures 26b-d, a high bias can be seen in the forecast systems. The maximum rainfall amount is just over five inches north of

Oklahoma City, while the two forecasts show highest rainfall amounts of 6.029 inches for the best WoFS member and 9.645 inches for WoF-Hybrid to the east of Oklahoma City. WoF-Hybrid also notably produces a high bias in rainfall over the entire domain, with widespread rainfall over three inches, while only isolated three-inch amounts appear in the stage IV data. The best WoFS member also has a high bias that shows up mostly in central Oklahoma, as opposed to the entire domain. Another shortcoming of WoF-Hybrid is that it produces backbuilding convection over northern Arkansas that yields rainfall amounts between three and nine inches in northern Arkansas while no rainfall above one inch is observed there. Both systems produce great coverage of the event; however, both systems have a high rainfall bias, more severely seen in WoF-Hybrid. This supports the overall idea taken from FSS results seen in Figure 10c that WoFS produces more skillful forecasts for higher rainfall thresholds, which can be mainly attributed to a higher number of false alarms at high rainfall thresholds in WoF-Hybrid.

Overall, WoFS and WoF-Hybrid produced very skillful forecasts for 27 May. In fact, WoFS output was used by NWS Norman and the Weather Prediction Center (WPC) regarding flash flooding in the Oklahoma City metropolitan area. NWS Norman issued a Flash Flood Warning for several counties in central Oklahoma due to high probabilities of 2-3 inches of rainfall spreading north of Oklahoma City (Lindley 2021). WoFS provided enough guidance to allow NWS forecasters to create Flash Flood Warnings in one county where rain had not yet occurred. One hour later, there was one foot of water present on roadways in the warned counties (Lindley 2021). WPC also referenced WoFS guidance in a Mesoscale Precipitation Discussion at 2239 UTC regarding the continuation of flash flooding in central Oklahoma (NOAA WPC 2021). It was stated that several WoFS forecasts in a row displayed probabilities over 90% of seeing rainfall above five inches, with some members depicting isolated totals over eight inches. Using WoFS probabilistic guidance helped increase confidence in WPC forecaster decisions to better articulate the severity of the flash flooding risk in central Oklahoma.

3.4.3 10 December

The final case analyzed in this study is the tornado outbreak over the mid-Mississippi Valley region on 10 December. Overall reflectivity object-based results are depicted in Figures 27a-c. The total number of reflectivity objects are similar between the WoFS member average and WoF-Hybrid, with both systems maintaining a low bias throughout the period. Most objects occur later in the initialization period with the highest number of objects occurring in the final three 0-6 hour forecasts. For the first five 0-6 hour forecasts there are very few, if any, matched objects in both systems as most of the convection through 2100 UTC is lower than the reflectivity object threshold. Since very few observed objects appear at a few early forecasts, neither system has much information on where convection is most likely to occur, which leads to a low matched object count. In the forecasts initialized from the 2200 UTC through the 0300 UTC, the object number increases very quickly, with WoF-Hybrid having more matched objects than all WoFS members at all but one initialization time. This difference in matched object counts paired with the similar total object counts yields differences in contingency table statistics after 2200 UTC. PODs/SRs/CSIs are between 0.05 and 0.22 higher for WoF-Hybrid compared to the WoFS member average for the final six 0-6 hour forecasts (Figure 27c). For the first five 0-6 hour forecasts, the PODs/CSIs are under 0.05 for both systems, with the first two forecasts yielding scores of zero. WoF-Hybrid produces more skillful reflectivity forecasts at 2200 UTC and after, when most of the convection and severe weather occurred on 10 December.

The UH/AWS object-based results depicted in Figures 27d-f also show a stark difference between WoF-hybrid and WoFS for 10 December. The total number of AWS objects, seen in Figure 27d, increases over the different forecast initialization times. For the forecasts initialized at 2100 UTC, both WoF-Hybrid and the WoFS member average have fewer than 20 UH objects over the six-hour forecast period, while the number of AWS objects is about 82. This, like the reflectivity results, can be attributed to slow convective initiation in the model systems at the early forecast periods. From 2200 to the 0000 UTC (when convection has been initiated), however, a large difference occurs between the two systems. WoF-Hybrid yields between 40 and 100 UH objects which is much closer to the observed AWS object count than the WoFS member average which produces between 20 and 50 UH objects (Figure 27d). For these forecasts, there is a similar large difference seen in the number of matched objects between the two systems. WoF-Hybrid has between 30 and 40 matches during the 2300 and 0000 UTC initializations while the WoFS member average has fewer than 10 (Figure 27e). It is clear as well that after the 2100 UTC forecast run WoF-Hybrid has notably more matched objects than the WoFS member average. In addition, there is a large spread in total and matched objects in the WoFS members after 2100 UTC, indicating that there may be less confidence in areas of strong mesocyclone development in WoFS.

The effects of large differences in UH matched object count between the two systems can be clearly seen in the contingency table-based statistics in Figure 27f. All initialization times before

2100 UTC have statistics of zero for both systems and are omitted from the plot. After 2300 UTC, there is a notable difference between WoF-Hybrid and the WoFS member average when most of the severe weather was reported. PODs (SRs) range from 0.16 to 0.3 (0.38 to 0.57) for WoF-Hybrid while these values range from 0.01 to 0.14 (0.08 to 0.36) for the WoFS member average (Figure 27f). Both maximum WoFS member average PODs and SRs do not surpass the lowest PODs and SRs shown for the forecasts from 2300 UTC and after. The largest difference between the two systems can be seen for the forecast at 2300 UTC: WoF-Hybrid POD is 0.26 higher, SR is 0.31 higher, and CSI is 0.18 higher than the WoFS member average. It is clear that WoF-Hybrid produces more skillful UH forecasts overall and earlier skilled UH forecasts than the majority of WoFS members.

The differences between the two systems for 10 December are quite large for both reflectivity and UH/AWS results. As this performance difference does not appear to be systematic across all types of events, it is important to determine if the higher WoF-Hybrid PODs/CSIs at most of the later lead times compared to the WoFS member average for high-end events seen in Figures 15 and 20 still holds true when omitting 10 December from the analysis. The difference in PODs, FARs, and CSIs between all high-end events with and without 10 December for both reflectivity and UH can be seen in Figure 28. Looking at the PODs in the top row of panels, the inclusion of 10 December lowers overall PODs for both systems at most lead times for both UH and reflectivity. The WoFS member average is affected slightly more than WoF-Hybrid by the inclusion of 10 December and therefore, increases the gap between WoF-Hybrid and WoFS. The overall trend between the two systems is similar, but more exaggerated when including 10 December.

In general, overall reflectivity FARs (Figures 28b) in both systems do not change much, but they do change in opposite directions: WoF-Hybrid FARs increase, while the WoFS member average FARs decrease without December 10th. For UH, however, the WoFS member average yields almost no change, while WoF-Hybrid FARs are slightly higher without 10 December. In general, WoF-Hybrid still yields lower FARs without 10 December than the WoFS member average at most lead times. Both reflectivity and UH CSIs in Figures 28c and 28f, respectively, undergo similar changes as PODs in Figure 28a and 28d: both systems see an increase in CSIs when not including 10 December in high-end events, but the WoFS member average increases more. This leads to an overall similar trend in both groupings of high-end events; however, the difference is again exaggerated with the inclusion of December 10th. In the end, the differences between the two systems seen when binned by lead time in the high-end events are still present without the inclusion of December 10th. Therefore, the conclusion stated earlier in this paper that WoF-Hybrid produces higher reflectivity and UH PODs and CSIs in the final hours of lead time implying that WoF-Hybrid may be propagating storms more accurately, is robust. Data from 10 December also upholds the other conclusion that WoF-Hybrid FAR is lower than WoFS member average FAR throughout each forecast period, on average, for high-end events may indicate that the higher resolution radar data usage and model grid may allow WoF-Hybrid to better predict mesocyclone development.

The most impactful supercell on 10 December produced the quad-state tornado over Arkansas, Missouri, Tennessee from 0107 UTC to 0236 UTC and over Tennessee and Kentucky from 0249 UTC to 0547 UTC (NOAA NWS 2021). To spatially analyze reflectivity forecasts of this supercell, two different stages of the supercell are presented: 0130 UTC, which depicts the tornado near the Missouri/Arkansas border and 0330 UTC, when the tornado produced the most structural damage and fatalities, directly hitting downtown Mayfield, Kentucky. Figure 29 depicts MRMS composite reflectivity at 0130 UTC and WoF-Hybrid and best WoFS member forecast composite reflectivity from 2300, 0000, and 0100 UTC respectively. The quad-state supercell depicted in MRMS data is located at 36°N 90.3°W. For the forecast started at 2300 UTC, WoF-Hybrid sustains the supercell present at 2300 UTC in the observed data, but, with a slight northeastward bias (Figure 29a). WoF-Hybrid also produces convection over central Arkansas, slightly to the west of where reflectivity objects are present at 0130 UTC. The best overall scoring WoFS member produces zero reflectivity objects near these two areas of convection. In fact, only one WoFS member produces convection near the observed supercell, but the forecast cell is not associated with strong UH.

Moving to the forecast initialized at 0000 UTC, WoF-Hybrid places a supercell correctly over the quad-state supercell at 0130 UTC (Figure 29b). In addition, the convection trailing behind the supercell is more accurately propagated by WoF-Hybrid at the 0000 UTC run compared to 2300 UTC. The best scoring WoFS member does produce convection near the observed supercell; however, the initial observed supercell in the 0000 UTC run was not sustained by this member, and the forecast reflectivity objects present are initiated 30 minutes into the forecast period by the model run. Additionally, the observed convection over central Arkansas is not forecast by the best scoring WoFS member. At the 0100 UTC run, both WoF-Hybrid and the best scoring WoFS member forecast convection over central Arkansas and correctly place the quad-state supercell (Figure 29c). In general, in this forecast run, WoF-Hybrid more accurately forecasts the shape and size of both the central Arkansas convection and the quad-state supercell. WoF-Hybrid produces much better forecasts for the quad-state supercell at 0130 UTC. Not only does it produce more spatially accurate depictions of convection, it also correctly sustains and strengthens the quad-state supercell at 0130 UTC with 2.5 hours of lead time. The best scoring WoFS member, however, only predicts a sustained supercell with 30 minutes of lead time, and only one member strengthens this cell in the 0000 UTC run. Finally, in addition to better spatial representations, WoF-Hybrid also has higher CSIs by as much as 0.082 over the entire domain from initialization time until 0130 UTC compared to the highest scoring WoFS member at each forecast represented.

By 0330 UTC, the quad-state supercell is entering downtown Mayfield, Kentucky, and can be seen in each panel of Figure 30 in the observed MRMS data at 36.8°N 88.5°W. At the 2200 UTC run, neither system predicts reflectivity objects near the quad-state supercell. WoF-Hybrid does correctly sustain convection in southeastern Missouri at 0330 UTC, but with a westward bias, while this is not present in the member with the highest CSI over the domain. For the forecast from 2300 UTC, WoF-Hybrid sustains the quad-state supercell, but with a slight northeastward bias (Figure 30b). A similar directional bias is seen in Figure 29b at the early stage of the supercell, meaning the storm propagated at a similar speed to the observed object from 0130 UTC to 0330 UTC. While the location of the supercell is not entirely correct, WoF-Hybrid still produced a similarly sized reflectivity object in the correct location for the most impactful supercell location with an impressive 4.5 hours of lead time. The best scoring WoFS member, however, has zero reflectivity contours in the area. In fact, only one WoFS member has weak convection just west of Mayfield, Kentucky from the 2300 UTC run. In Figures 30c and 30d depicting the 0000 and 0100 UTC runs, WoF-Hybrid produces a reflectivity object in the correct location with respect to the quad-state supercell, while the best scoring member still does not sustain this convection in both forecasts. By 0200 UTC, the best scoring member and WoF-Hybrid have objects overlapping with the observed quad-state supercell. By this point, almost all the WoFS members also place convection of various intensities at the same location. And finally, at 0300 UTC run, both systems produce a similarly sized supercell at the correct location.

The main takeaway from this spatial analysis is the difference in lead time provided by each system regarding the quad-state supercell. WoF-Hybrid sustained convection as early as the 2300 UTC run, providing 4.5 hours of lead time and correctly forecast the location of the quad-state supercell. The majority of WoFS members, however, did not place any convection near Mayfield, Kentucky until the 0200 UTC run, providing 1.5 hours of lead time. One benefit of producing forecasts every hour is the ability to look back at previous runs to determine if reflectivity objects are being consistently forecast in the same place at the same verification time. For this case, WoF-Hybrid consistently produces strong convection near Mayfield, Kentucky for five straight initializations, which could help increase forecaster confidence in strong convection occurring in this area around the predicted time.

Operationally, more skillful spatial and temporal representations are very important in increasing severe thunderstorm warning lead times and potentially saving lives. This spatial reflectivity analysis provides an example that shows WoF-Hybrid has great forecasting potential for high-end events. In addition, WoF-Hybrid may provide an added benefit of one more model output for operational forecasters to use when other systems may not have as good of a grasp on a specific event. Finally, the difference between the two similar systems in both CSI and spatial representations seen in Figures 29 and 30 of impactful supercells is similarly seen in a study by Kong et al. (2020) over the 10 May 2010 Oklahoma tornado outbreak. Kong et al. (2020) determined that 3DEnVAR outperformed the EnKF system with the lowest root-mean-square innovations for reflectivity and produced better storm structure and intensities. More research on the 10 December WoF-Hybrid and WoFS reflectivity forecasts needs to occur to determine how the data assimilation techniques are affecting output.

In addition to reflectivity, it is important to analyze UH swaths during peak event severity to determine how well the two systems forecast mesocyclone development from 0200-0400 UTC over the entire domain on 10-11 December starting from 2300-0200 UTC initializations(Figure 31). At the 2300 UTC run, WoF-Hybrid CSI scores are 0.115 higher than the average scoring member (Figure 31a). This is mainly due to WoF-Hybrid maintaining both a mesocyclone over the path of the quad-state tornado from far northeastern Arkansas through western Kentucky and secondary UH swaths over severe weather reports in northeastern Arkansas. The best scoring member produces a swath to the northeast of the quad-state supercell without any convection over northeastern Arkansas. The average member depicted shows a very low CSI at this time, because while the UH swath looks to be over the storm reports in northeastern Arkansas, the UH swath occurs over 30 minutes earlier than was observed. By the 0000 UTC run, WoF-Hybrid produces two UH tracks, and again the quad-state UH track which matches the storm reports well. There are many false alarm objects over southern Illinois and Indiana by the WoFS members depicted, that are not present in WoF-Hybrid, leading to lower CSIs.

At the 0100 UTC forecast, WoF-Hybrid shows a swath directly over all storm reports associated with the quad state supercell, while the highest scoring member shows a similar swath with a northeastward bias (Figure 31c). The average scoring member, however, does not maintain a strong updraft over these reports, a main reason for a much lower CSI at this time. At the 0200 UTC run, the WoF-Hybrid and best WoFS member forecasts are quite similar for the quad-state supercell (Figure 31d). The average scoring WoFS member, however, has a large westward bias for the updraft associated with the quad-state supercell. All two later forecasts match the storm reports behind the quad-state supercell in northeastern Arkansas, with WoF-Hybrid producing the best UH swath (Figure 31c-d). In addition, storm reports in eastern Missouri through central Illinois are decently captured by all systems.

Overall, like reflectivity forecasts, WoF-Hybrid better depicted the quad-state supercell and general severe convection than the best performing WoFS member at three of the four 0-6 hour forecasts. WoF-Hybrid stands out because it produces fewer false alarm UH swaths and more accurately depicts the propagation and location of the quad-state supercell hours before an average performing member. Again, operationally, it is very important for forecasters to receive model output that produces consistent UH depictions over multiple model initializations to increase confidence in severe hazard warnings. WoF-Hybrid successfully produces consistent UH swaths of the most impactful supercells in this event throughout the forecast period, with only slight variations in each ensuing forecast run. The case of 10 December provides strong motivation for further research to investigate the use of WoF-Hybrid as a companion piece to WoFS, especially for high-impact weather events, due to the system’s potential to better resolve and propagate strong updrafts.

4. Conclusion

Two rapidly updating, high resolution data assimilation systems are verified against observations to find the strengths and weaknesses of the systems to forecast life-threatening severe weather hazards. WoFS is one such system that aims to be operational within the next 5-10 years. The operational future of WoF-Hybrid, a deterministic forecast of higher resolution than WoFS and developed also by the WoF group is still unknown. This research aims to identify whether WoF-Hybrid provides a more accurate depiction of thunderstorms than WoFS, or if it provides enough additional useful information for operational forecasters, specifically for high-impact events, to make its continued development and potential operationalization worthwhile. Based on a qualitative and quantitative analysis of reflectivity, precipitation accumulation, and UH, there is promise for the operational use of the higher resolution, deterministic WoF-Hybrid alongside WoFS.

Some results are derived from both systems' 2021 runs based on verification metrics and individual case study analyses. Within the first hour of lead times from the FSS analysis, there is a very high reflectivity bias in WoF-Hybrid that leads to less skillful forecasts, while after this first hour WoF-Hybrid produces more skillful forecasts than WoFS. This implies better forecast coverage possibly because of the better propagation of storms from WoF-Hybrid later in the forecast period. This is also seen in the object-based reflectivity results where WoF-Hybrid matched objects have a lower AMD (i.e., are closer in space to the observed objects) than WoFS at all lead times after one hour. This may be attributed to higher resolution radar data and model grids used in WoF-Hybrid, allowing the system to better resolve storm structure. In the six-hour precipitation analysis, WoF-Hybrid is significantly more skillful at the 0.1-inch threshold, while WoFS is more skillful at the three higher thresholds. This may mean that WoF-Hybrid rainfall forecasts better predict spatial coverage of an event, while WoFS members are better used for localized rainfall forecasts of higher intensity, indicating each system may have a unique use for operational forecasters.

Within the object-based verification framework, further reason for the inclusion of WoF-Hybrid alongside WoFS is presented. One of the main conclusions is that both systems yield fewer reflectivity and UH false alarms after most of the convection is initiated, indicating forecasts from both models produce more skillful forecasts when radar data which contain convective information are assimilated in multiple cycles. This is further supported by a higher reflectivity and UH bias in each system in the several early-hour forecasts; the bias converges to one after convection is fully developed. Comparing both systems, WoF-Hybrid in general produces higher SRs in six-hour forecasts of reflectivity and UH than the WoFS member average at later forecast initialization times. This may again be a benefit of a higher resolution system that better predicts storm structure. It could also be that WoF-Hybrid better propagates convection in MCSs, as this is the dominant convective mode at the later forecast initialization times. In addition, WoF-Hybrid yields an AAR (average area ratio) closer to one at all initialization times for reflectivity forecasts and after the 2100 UTC initialization for UH forecasts. This implies that WoF-Hybrid matched objects are more accurate in size to the observed objects than WoFS. This is a notable strength of WoF-Hybrid, as small differences in object size may be the difference between severe weather or no impact in a specific location.

Viewing by lead time, WoF-Hybrid also produces lower FARs and higher PODs/CSIs in the final 1-2 hours of each forecast. This could also support the claim that WoF-Hybrid provides more accurate storm propagation. In addition, WoF-Hybrid could be producing updrafts that are better sustained and more accurate to observations than in WoFS, owing to the higher resolution. Also, WoF-Hybrid UH bias values are consistently between 1.1 and 1.4 while WoFS member average starts near 1.7 and ends under 1.0, which also implies that WoF-Hybrid is creating/maintaining/diminishing updrafts at a similar rate to the observations. One necessary improvement for each system is in the reduction of false alarms. In general, over half of the predicted objects by each system in reflectivity and UH forecasts are not present in the observations, which may lead to incorrect operational forecaster interpretations. A large amount of spurious convection is also present within the first hour in both systems, but especially WoF-Hybrid. One reason for this may be the initial balancing of microphysical parameters in the cloud physics scheme. Another reason can be a consequence of using a higher resolution model, as similar results regarding spurious convection within the first hour of a higher resolution system can be seen in Miller et al. (2022). To create more accurate model output and produce more confident operational forecasts, the false alarm objects must be diminished.

Partitioning the events by severity and investigating individual case studies provided further insight into the potential operational benefits of WoF-Hybrid. Both systems overall perform the best during the high-end events (those with the greatest number of local storm reports), but results show slightly higher skill in WoF-Hybrid. The higher PODs and lower FARs for WoF-Hybrid seen in the final 1-2 hours of lead times in both reflectivity and UH forecasts are driven by the high-impact events. A larger difference between WoF-Hybrid and WoFS can be seen in the high-end events, while the low-end and mid-severity events show similar scores, or an opposite relationship. In addition, when binning event severity by initialization time, WoF-

Hybrid yields higher PODs/CSIs at most lead times for both variables and lower FARs for UH compared to the WoFS member average.

In the three case studies analyzed, WoF-Hybrid in general produces similar or better spatial depictions of the most impactful convection than the best scoring WoFS member (the member that scores best over the full domain). For 27 May, the reflectivity forecasts are quite similar, and of very high skill from both systems, while WoF-Hybrid precipitation accumulation forecasts are toward the bottom of member FSS spread. In general, both predict the propagation of the MCS similarly, however, WoF-Hybrid yields a much higher rainfall bias and addition of rainfall that is not present in the observations. One of the largest successes of WoF-Hybrid is that the most impactful reflectivity objects from 26 May and 10 December were predicted from 1-3 initialization times before the most skillful WoFS member for each case. Early hints of strong convection within WoF-Hybrid 2-4 hours before the storm occurred is very promising and should be verified via more cases when considering the operational inclusion of WoF-Hybrid.

WoF-Hybrid produces fewer false alarms can be seen in UH forecasts on 10 December in WoF-Hybrid during the most impactful event times compared to the most skillful WoFS member. For 26 May, UH forecasts are better for the best scoring WoFS member compared to WoF-Hybrid. The WoFS member better predicts both high impact supercells earlier than WoF-Hybrid, with later initializations producing forecasts of similar skill. The 10 December case, however, depicts UH swaths of the quad-state tornado two initialization times (i.e., two hours) earlier than the best scoring member and 3-4 initialization times earlier than most members. Operationally, more skillful spatial and temporal representations are very important in increasing severe thunderstorm warning lead times and potentially saving lives. These case studies show that WoF-Hybrid has great forecasting potential for high-end events. In addition, WoF-Hybrid may provide an added benefit of one more model output for operational forecasters to use when other systems may not have as good of a grasp on a specific event.

It is important to note that there are several limitations of this study. Firstly, only 31 cases from 2021 are examined. While this provides a sufficient sample size for this study, additional years of WoFS runs must be analyzed in a similar manner to determine if these results consistently hold. There are also several initialization times missing from various cases due to missing MRMS, WoF-Hybrid, and WoFS data. In addition, only two verification methods are analyzed, and while these depict different perspectives of the data, more methods should be applied to bolster results. The analysis also only includes three variables (reflectivity, UH and precipitation). There are a plethora of variables (temperature, dew point, CAPE, etc.) from these systems that could be analyzed to determine the differences between the two systems more comprehensively. The thresholds chosen for this study were determined by a clear separation in clusters of cases when binned by the number of storm reports. There are many ways to classify the severity of the events, and therefore studies may partition event severity differently.

This study also brings up many ideas for future research. Firstly, a potential solution to a high bias peak at the first ten-minute time step is the incremental analysis technique (Bloom et al. 1996; Lei and Whitaker 2016). This would smooth out the analysis increments from the cloud microphysics scheme and potentially decrease the large peak in reflectivity. Future code development and sensitivity tests should be conducted to determine if this technique provides a solution. Another important result discussed in this study is the potential effect on forecast skill of assimilating MRMS data of varying resolution. Various MRMS resolutions should be tested in each system to weigh the effect of increased skill versus computational power. This would also help determine how much of an effect higher resolution MRMS data has on increasing forecast skill in both WoFS and WoF-Hybrid.

More research is also needed to determine ways to decrease model false alarms and reflectivity biases. These are major weaknesses of both systems, more so in WoF-Hybrid. One potential short-term solution is to tell end users in training for operational use that a high false alarm rate and reflectivity bias are present, particularly at earlier initialization times prior to much convective initiation within the domain. This way, forecasters are aware of this shortcoming and will take this into consideration when producing forecasts. In addition, this study analyzes each ensemble member as a deterministic forecast; going forward, using the Probability Matched Means technique–which corrects for the low biases of peak values seen in the ensemble mean–to analyze WoFS as a system compared to WoF-Hybrid may provide useful results. Finally, more individual cases from 2021 should be analyzed, especially those of lower event severity, to diagnose system differences, and find ways to improve forecasts in these more conditional severe environments.

In general, this study outlines the importance of verification techniques to analyze model strengths and weaknesses in the prediction of severe hazards. Over the last few real-time experiments, WoFS has shown great success in the prediction of such events and strong promise for operational use. This study shows the benefit of including a deterministic, higher resolution system, WoF-Hybrid, in forecast operations alongside WoFS. While overall statistics are quite similar between the two systems, the fewer number of false alarms, more accurately sized objects that are also closer in distance, and additional forecast skill in high impact events in WoF-Hybrid are strong motivators for the continuation of research to improve this system and its potential for use in an operational setting.

References

Bloom, S. C., L. L. Takacs, A. M. da Silva, and D. Ledvina, 1996: Data Assimilation Using Incremental Analysis Updates. Mon. Weather Rev., 124, 1256–1271, https://doi.org/10.1175/1520-0493(1996)124<1256:DAUIAU>2.0.CO;2.

Brooks, H., C. Doswell III, and L. Wicker, 1993: STORMTIPE: a forecasting experiment using a three-dimensional cloud model. Weather Forecast., 8, 352–362, https://doi.org/10.1175/1520-0434(1993)008<0352:SAFEUA>2.0.CO;2.

Bryan, G.H., J.C. Wyngaard, and M.J. Fritsch, 2003: Resolution requirements for the simulation of deep moist convection. Mon. Wea. Rev., 131, 2394-2416.

Burke, P., 2021. NSSL laboratory review 2021. NOAA National Severe Storms Laboratory. Accessed June 10, 2022. https://www.nssl.noaa.gov/about/events/review2021/.

Charney, J. G., and Y. Ogura, 1960: A Numerical Model for Thermal Convection in the Atmosphere. J. Meteorol. Soc. Jpn. Ser II, 38, 19a–19a, https://doi.org/10.2151/jmsj1923.38.6_19a.

Chen, L., C. Liu, Y. Jung, P. Skinner, M. Xue, and R. Kong, 2022: Object-Based Verification of GSI EnKF and Hybrid En3DVar Radar Data Assimilation and Convection-Allowing Forecasts within a Warn-on-Forecast Framework. Weather Forecast., 37, 639–658, https://doi.org/10.1175/WAF-D-20-0180.1.

Choate, J. J., A. J. Clark, B.T. Gallo, E. Grimes, P. L. Heinselman, P. S. Skinner, and K. A.Wilson, 2018: Examining the use of the NSSL experimental warn-on-forecast system for ensembles for the prediction of severe storms through short-term forecast outlooks during the 2018 spring forecasting experiment. 29th Conf. on Severe Local Storms, Stowe, VT, Amer. Meteor. Soc., 3A.5, https://ams.confex.com/ams/29SLS/webprogram/Paper348346.html

Clark, A. J., and Coauthors, 2020: A Real-Time, Simulated Forecasting Experiment for Advancing the Prediction of Hazardous Convective Weather. Bull. Am. Meteorol. Soc., 101, E2022–E2024, https://doi.org/10.1175/BAMS-D-19-0298.1.

Clark, A. J., and Coauthors, 2021a: A Real-Time, Virtual Spring Forecasting Experiment to Advance Severe Weather Prediction. Bull. Am. Meteorol. Soc., 102, E814–E816, https://doi.org/10.1175/BAMS-D-20-0268.1.

Clark, A. J., and Coauthors, 2021b: Spring forecasting experiment 2021 - preliminary findings and results. NOAA Hazardous Weather Testbed. Accessed Oct 1, 2022. https://hwt.nssl.noaa.gov/sfe/2021/docs/HWT_SFE_2021_Prelim_Findings_FINAL.pdf

Clark, A. J., and Coauthors, 2022: Spring forecasting experiment 2022 - preliminary findings and results. NOAA Hazardous Weather Testbed. Accessed Sept 9, 2022. https://hwt.nssl.noaa.gov/sfe/2022/docs/HWT_SFE_2022_Prelim_Findings_FINAL.pdf

Colle, B. A., J. B. Olson, and J. S. Tongue, 2003: Multiseason Verification of the MM5. Part I: Comparison with the Eta Model over the Central and Eastern United States and Impact of MM5 Resolution. Weather Forecast., 18, 431–457, https://doi.org/10.1175/1520-0434(2003)18<431:MVOTMP>2.0.CO;2.

Demuth, J. L., and Coauthors, 2020: Recommendations for developing useful and usable convection-allowing model ensemble information for NWS Forecasters. Weather and Forecasting, 35, 1381–1406, doi:10.1175/waf-d-19-0108.1.

Dowell, D. C., L. J. Wicker, and C. Snyder, 2011: Ensemble Kalman Filter Assimilation of radar observations of the 8 May 2003 Oklahoma City Supercell: Influences of reflectivity observations on storm-scale analyses. Monthly Weather Review, 139, 272–294, doi:10.1175/2010mwr3438.1.

Dowell, D. C., and Coauthors, 2021: The High-Resolution Rapid Refresh (HRRR): An hourly updating convection permitting forecast model. Part 1: Motivation and system description. Wea. Forecasting, in preparation.

Droegemeier, K.K., M. Xue, A. Sathye, K. Brewster, G. Bassett, J. Zhang, Y. Liu, M. Zou, A. Crook, V. Wong, and R. Carpenter, 1996: Realtime numerical prediction of storm-scale weather during VORTEX '95, Part I: Goals and methodology. Preprints, 18th Conf. on Severe Local Storms, 15-20 Jan., Amer. Meteor. Soc., San Francisco, CA, 6-10/

Droegemeier, K.K., 1997: The numerical prediction of thunderstorms: Challenges, potential benefits, and results from realtime operational tests. WMO Bulletin, 46, 324-336.

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. Weather Forecast., 34, 1721–1739, https://doi.org/10.1175/WAF-D-19-0094.1.

Gallo, B. T., and Coauthors, 2022: Exploring the Watch-to-Warning Space: Experimental Outlook Performance during the 2019 Spring Forecasting Experiment in NOAA’s Hazardous Weather Testbed. Weather Forecast., 37, 617–637, https://doi.org/10.1175/WAF-D-21-0171.1.

Guerra, J. E., P. S. Skinner, A. Clark, M. Flora, B. Matilla, K. Knopfmeier, and A. E. Reinhart, 2022: Quantification of NSSL Warn-On-Forecast System Accuracy by Storm Age using Object-based Verification. Weather Forecast., 1, https://doi.org/10.1175/WAF-D-22-0043.1.

Gao, J., M. Xue, A. Shapiro, and K. K. Droegemeier, 1999: A Variational Method for the Analysis of Three-Dimensional Wind Fields from Two Doppler Radars. Mon. Weather Rev., 127, 2128–2142, https://doi.org/10.1175/1520-0493(1999)127<2128:AVMFTA>2.0.CO;2.

Gao, J., M. Xue, K. Brewster, and K. K. Droegemeier, 2004: A Three-Dimensional Variational Data Analysis Method with Recursive Filter for Doppler Radars. J. Atmospheric Ocean.

Technol., 21, 457–469, https://doi.org/10.1175/1520-0426(2004)021<0457:ATVDAM>2.0.CO;2.

Gao, J., and Coauthors, 2013a: A Real-Time Weather-Adaptive 3DVAR Analysis System for Severe Weather Detections and Warnings. Weather Forecast., 28, 727–745, https://doi.org/10.1175/WAF-D-12-00093.1.

Gao, J., M. Xue, and D. Stensrud, 2013b: The Development of a Hybrid EnKF-3DVAR Algorithm for Storm-Scale Data Assimilation. Adv. Meteorol., 2013, https://doi.org/10.1155/2013/512656.

Gao, J., and D. J. Stensrud, 2014: Some Observing System Simulation Experiments with a Hybrid 3DEnVAR System for Storm-Scale Radar Data Assimilation. Mon. Weather Rev., 142, 3326–3346, https://doi.org/10.1175/MWR-D-14-00025.1.

Gao, J., Y. Wang, D. M. Wheatley, K. H. Knopfmeier, T. A. Jones, and G. Creager, 2017: Test of a Weather-Adaptive Hybrid 3DEnVAR and WRF-DART Analysis and Forecast System During the HWT Spring Experiments in 2017. 38th Conference on Radar Meteorology, AMS https://ams.confex.com/ams/38RADAR/meetingapp.cgi/Paper/321145.

Heinselman, P., and Coauthors, 2022: Rapid prediction of high-impact weather: The Warn-on-Forecast System. Wea. Forecasting, in preparation.

Hou, D., E. Kalnay, and K.K. Droegemeier, 2001: Objective verification of the SAMEX '98 ensemble forecasts. Mon. Wea. Rev., 129, 73-91.

Hu, J., and Coauthors, 2021: Evaluation of an experimental Warn-on-Forecast 3DVAR analysis and forecast system on quasi-real-time short-term forecasts of high-impact weather events. Q. J. R. Meteorol. Soc., 147, 4063–4082, https://doi.org/10.1002/qj.4168.

Janish, P.R., K.K. Droegemeier, M. Xue, K. Brewster, and J. Levit, 1995: Evaluation of the advanced regional prediction system (ARPS) for storm-scale modeling applications in operational forecasting. Proc.,14th Conf. on Wea. and Forecasting, 15-20 Jan., Amer. Meteor. Soc., Dallas, TX., 224-229.

Jones, B., 2014: How does the skill of global model precipitation forecasts over Europe depend on spatial scale? University of Reading, http://www.met.reading.ac.uk/~sws00rsp/teaching/postgrad/jones.pdf.

Jones, T. A., K. Knopfmeier, D. Wheatley, G. Creager, P. Minnis, and R. Palikonda, 2016: Storm-Scale Data Assimilation and Ensemble Forecasting with the NSSL Experimental Warn-on-Forecast System. Part II: Combined Radar and Satellite Data Experiments. Weather Forecast., 31, 297–327, https://doi.org/10.1175/WAF-D-15-0107.1.

Jones, T. A., P. Skinner, K. Knopfmeier, E. Mansell, P. Minnis, R. Palikonda, and W. Smith, 2018: Comparison of Cloud Microphysics Schemes in a Warn-on-Forecast System Using Synthetic Satellite Objects. Weather Forecast., 33, 1681–1708, https://doi.org/10.1175/WAF-D-18-0112.1.

Kong, R., M. Xue, C. Liu, and Y. Jung, 2020: Comparisons of Hybrid En3DVar with 3DVar and EnKF for Radar Data Assimilation: Tests with the 10 May 2010 Oklahoma Tornado Outbreak. Mon. Weather Rev., 149, 21–40, https://doi.org/10.1175/MWR-D-20-0053.1.

Lawson, J. R., J. S. Kain, N. Yussouf, D. C. Dowell, D. M. Wheatley, K. H. Knopfmeier, and T. A. Jones, 2018: Advancing from Convection-Allowing NWP to Warn-on-Forecast: Evidence of Progress. Weather Forecast., 33, 599–607, https://doi.org/10.1175/WAF-D-17-0145.1.

Lei, L., and J. S. Whitaker, 2016: A Four-Dimensional Incremental Analysis Update for the Ensemble Kalman Filter. Mon. Weather Rev., 144, 2605–2621, https://doi.org/10.1175/MWR-D-15-0246.1.

Lilly, D.K., 1991: Numerical prediction of thunderstorms – has its time come? Quart. J. Royal Met. Soc., 116, 779-798.

Lindley, T., 2021: The 27 May 2021 Flash Floods: A spectrum of Effective Messaging. NOAA National Weather Service. Accessed October 26, 2022.

Liu, H., M. Hu, G. Ge, D. Stark, H. Shao, K. Newman, and J. Whitaker, 2018: Ensemble Kalman Filter (EnKF) User's Guide Version 1.3. Developmental Testbed Center. Available at https://dtcenter.org/community-code/ensemble-kalman-filter-system-enkf/documentation, 80 pp.

Mansell, E. R., C. L. Ziegler, and E. C. Bruning, 2010: Simulated electrification of a small thunderstorm with two-moment bulk microphysics. Journal of the Atmospheric Sciences, 67, 171–194, doi:10.1175/2009jas2965.1.

Miller, W. J. S., and Coauthors, 2022: Exploring the Usefulness of Downscaling Free Forecasts from the Warn-on-Forecast System. Weather Forecast., 37, 181–203, https://doi.org/10.1175/WAF-D-21-0079.1.

NOAA NCEI, 2022: Billion-dollar weather and climate disasters. Accessed June 3, 2022. https://www.ncei.noaa.gov/access/billions/.

NOAA NSSL, 2022a: Research Tools: Simulation. NOAA National Severe Storms Laboratory. Accessed August 12, 2022. https://www.nssl.noaa.gov/tools/simulation/.

NOAA NSSL, 2022b: WoFS - Realtime Viewer. NOAA National Severe Storms Laboratory. Accessed October 6, 2022. https://wof.nssl.noaa.gov/realtime/?model=wofs&product=member&sector=wofs.

NOAA NWS, 2021: The Violent Tornado Outbreak of December 10-11, 2021. NOAA National Weather Service. Accessed September 1, 2022, https://www.weather.gov/pah/December-10th-11th-2021-Tornado.

NOAA WPC, 2021: Mesoscale Precipitation Discussion: #0246. NOAA Weather Prediction Center. Accessed October 26, 2022, https://www.wpc.ncep.noaa.gov/metwatch/metwatch_mpd_multi.php?md=246&yr=2021

Pan, S., J. Gao, T. A. Jones, Y. Wang, X. Wang, and J. Li, 2021: The Impact of Assimilating Satellite-Derived Layered Precipitable Water, Cloud Water Path, and Radar Data on Short-Range Thunderstorm Forecasts. Mon. Weather Rev., 149, 1359–1380, https://doi.org/10.1175/MWR-D-20-0040.1.

Piani, C., G. P. Weedon, M. Best, S. M. Gomes, P. Viterbo, S. Hagemann, and J. O. Haerter, 2010: Statistical bias correction of global simulated daily precipitation and temperature for the application of hydrological models. J. Hydrol., 395, 199–215, https://doi.org/10.1016/j.jhydrol.2010.10.024.

Rife, D. L., and C. A. Davis, 2005: Verification of Temporal Variations in Mesoscale Numerical Wind Forecasts. Mon. Weather Rev., 133, 3368–3381, https://doi.org/10.1175/MWR3052.1.

Roberts, N. M., and H. W. Lean, 2008: Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events. Mon. Weather Rev., 136, 78–97, https://doi.org/10.1175/2007MWR2123.1.

Roebber, P. J., D. M. Schultz, B. A. Colle, and D. J. Stensrud, 2004: Toward Improved Prediction: High-Resolution and Ensemble Modeling Systems in Operations. Weather Forecast., 19, 936–949, https://doi.org/10.1175/1520-0434(2004)019<0936:TIPHAE>2.0.CO;2.

Schwartz, C. S., and R. A. Sobash, 2019: Revisiting Sensitivity to Horizontal Grid Spacing in Convection-Allowing Models over the Central and Eastern United States. Mon. Weather Rev., 147, 4411–4435, https://doi.org/10.1175/MWR-D-19-0115.1.

Skamarock, C., B. Klemp, J. Dudhia, O. Gill, M. Barker, W. Wang, and G. Powers, 2005: A Description of the Advanced Research WRF Version 2. https://doi.org/10.5065/D6DZ069T.

Skamarock, W., and Coauthors, 2008: A description of the Advanced Research WRF Version 3. No. NCAR/TN-475+STR. https://doi.org/10.5065/D68S4MVH.

Skinner, P. S., and Coauthors, 2018: Object-based verification of a prototype warn-on-forecast system. Weather and Forecasting, 33, 1225–1250, doi:10.1175/waf-d-18-0020.1.

Smith, A. B., 2022: 2021 U.S. billion-dollar weather and climate disasters in historical context. Accessed June 2, 2022. https://www.climate.gov/news-features/blogs/beyond-data/2021-us-billion-dollar-weather-and-climate-disasters-historical.

Steiner, J. T., 1973: A Three-Dimensional Model of Cumulus Cloud Development. J. Atmospheric Sci., 30, 414–435, https://doi.org/10.1175/1520-0469(1973)030<0414:ATDMOC>2.0.CO;2.

Stensrud, D. J., and Coauthors, 2009: Convective-Scale Warn-on-Forecast System: A Vision for 2020. Bull. Am. Meteorol. Soc., 90, 1487–1500, https://doi.org/10.1175/2009BAMS2795.1.

Stensrud, D. J., and J. Gao, 2010: Importance of Horizontally Inhomogeneous Environmental Initial Conditions to Ensemble Storm-Scale Radar Data Assimilation and Very Short-Range Forecasts. Mon. Weather Rev., 138, 1250–1272, https://doi.org/10.1175/2009MWR3027.1.

Stensrud, D. J., and Coauthors, 2013: Progress and challenges with Warn-on-Forecast. Atmospheric Res., 123, 2–16, https://doi.org/10.1016/j.atmosres.2012.04.004.

United Nations Office for Disaster Risk Reduction (UNDRR), 2020: Human cost of disasters: An overview of the last 20 years. UNDRR. Accessed June 6, 2022. https://www.undrr.org/publication/human-cost-disasters-overview-last-20-years-2000-2019

Wang, X., C. Snyder, and T. M. Hamill, 2007: On the theoretical equivalence of differently proposed ensemble/3D-Var hybrid analysis schemes. Mon. Wea. Rev. 135, 222-227.

Wang, Y., J. Gao, P. S. Skinner, D. M. Wheatley, J. J. Choate, T. A. Jones, and G. Creager, 2018: Test of a Hybrid 3DEnVAR and WRF-DART Analysis and Forecast System during the HWT Spring Experiments in 2017. 98th American Meteorological Society Annual Meeting, AMS https://ams.confex.com/ams/98Annual/meetingapp.cgi/Paper/330920 (Accessed September 8, 2022).

Wang, Y., J. Gao, P. S. Skinner, K. Knopfmeier, T. Jones, G. Creager, P. L. Heiselman, and L. J. Wicker, 2019: Test of a weather-adaptive dual-resolution hybrid warn-on-forecast analysis and forecast system for several severe weather events. Weather and Forecasting, 34, 1807–1827, doi:10.1175/waf-d-19-0071.1.

Weisman, M. L., and J. B. Klemp, 1982: The Dependence of Numerically Simulated Convective Storms on Vertical Wind Shear and Buoyancy. Mon. Weather Rev., 110, 504–520, https://doi.org/10.1175/1520-0493(1982)110<0504:TDONSC>2.0.CO;2.

Weisman, M. L., and J. B. Klemp, 1984: The Structure and Classification of Numerically Simulated Convective Storms in Directionally Varying Wind Shears. Mon. Weather Rev., 112, 2479–2498, https://doi.org/10.1175/1520-0493(1984)112<2479:TSACON>2.0.CO;2.

Weisman, M. L., W. C. Skamarock, and J. B. Klemp, 1997: The Resolution Dependence of Explicitly Modeled Convective Systems. Mon. Weather Rev., 125, 527–548, https://doi.org/10.1175/1520-0493(1997)125<0527:TRDOEM>2.0.CO;2.

Wheatley, D. M., K. H. Knopfmeier, T. A. Jones, and G. J. Creager, 2015: Storm-Scale Data Assimilation and Ensemble Forecasting with the NSSL Experimental Warn-on-Forecast System. Part I: Radar Data Experiments. Weather Forecast., 30, 1795–1817, https://doi.org/10.1175/WAF-D-15-0043.1.

Wicker, L. J., M. P. Kay, and M. P. Foster, 1997: STORMTIPE-95: Results from a Convective Storm Forecast Experiment. Weather Forecast., 12, 388–398, https://doi.org/10.1175/1520-0434(1997)012<0388:SRFACS>2.0.CO;2.

Wilhelmson, R. B., 1974: The Life Cycle of a Thunderstorm in Three Dimensions. J. Atmospheric Sci., 31, 1629–1651, https://doi.org/10.1175/1520-0469(1974)031<1629:TLCOAT>2.0.CO;2.

Wilhelmson, R. B., and L. J. Wicker, 2001: Numerical Modeling of Severe Local Storms. Severe Convective Storms, C.A. Doswell, Ed., Meteorological Monographs, American Meteorological Society, 123–166.

Wilson, K. A., and Coauthors, 2019: Exploring Applications of Storm-Scale Probabilistic Warn-on-Forecast Guidance in Weather Forecasting. 557–572.

Wilson, K. A., B. T. Gallo, P. Skinner, A. J. Clark, P. Heinselman, and J. J. Choate, 2021: Analysis of End User Access of Warn-on-Forecast Guidance Products during an Experimental Forecasting Task. Weather Clim. Soc., 13, 859–874, https://doi.org/10.1175/WCAS-D-20-0175.1.

Xue, M., K. Brewster, K. Droegemeier, F. Carr, V. Wong, Y. Liu, A. Sathye, G. Bassett, P. Janish, J. Levit and P. Bothwell, 1996: Realtime numerical prediction of storm-scale weather during VORTEX '95, Part II: Operations summary and example predictions. Preprints, 18th Conf. on Severe Local Storms, 19-23 Feb., Amer. Meteor. Soc., San Francisco, CA., 178-182.

Yussouf, N., 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. Weather Rev., 143, 3044–3066, https://doi.org/10.1175/MWR-D-14-00268.1.

Yussouf, N., J. S. Kain, and A. J. Clark, 2016: Short-Term Probabilistic Forecasts of the 31 May 2013 Oklahoma Tornado and Flash Flood Event Using a Continuous-Update-Cycle Storm-Scale Ensemble System. Weather Forecast., 31, 957–983, https://doi.org/10.1175/WAF-D-15-0160.1.

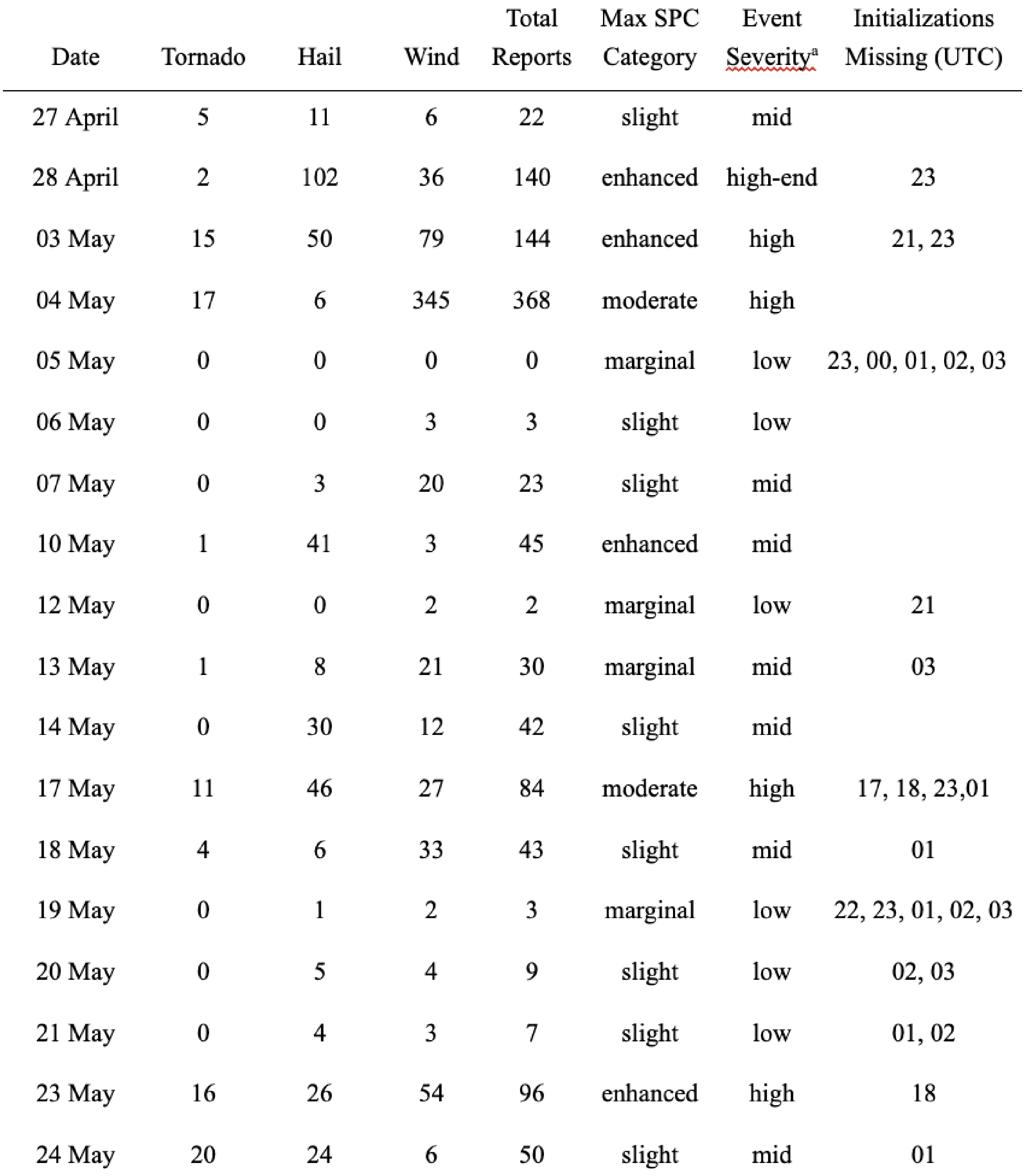
Zhang, F., C. Snyder, and J. Sun, 2004: Impacts of Initial Estimate and Observation Availability on Convective-Scale Data Assimilation with an Ensemble Kalman Filter. Mon. Weather Rev., 132, 1238–1253, https://doi.org/10.1175/1520-0493(2004)132<1238:IOIEAO>2.0.CO;2.

Zhang, F., A. M. Odins, and J. W. Nielsen-Gammon, 2006: Mesoscale Predictability of an Extreme Warm-Season Precipitation Event. Weather Forecast., 21, 149–166, https://doi.org/10.1175/WAF909.1.

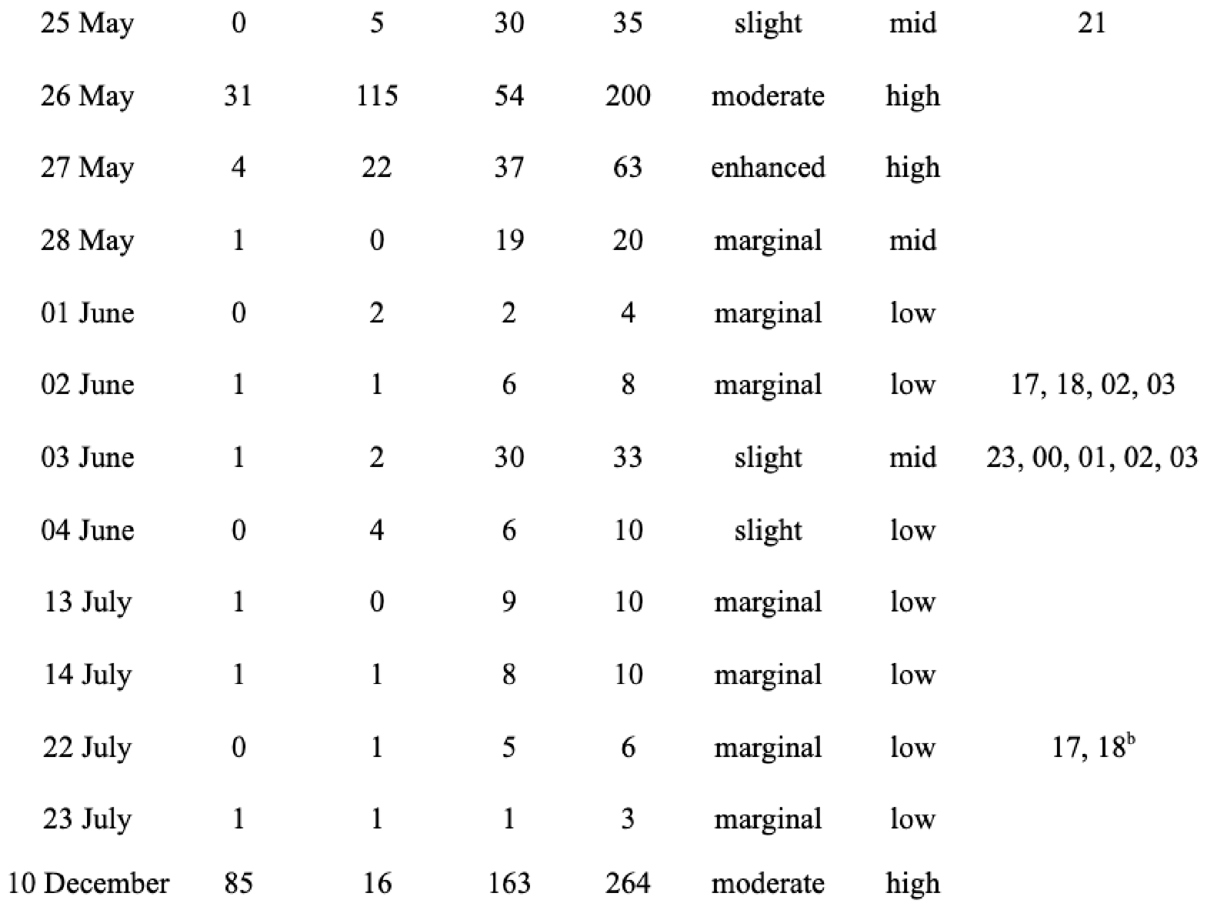
Appendix

Table 1

Summary of 2021 WoFS/WoF-Hybrid real time cases. For each case, the number of tornado, hail, wind, and total reports within the WoFS domain, maximum SPC risk category for the 1630 UTC outlook, event severity, and missing initialization times are provided.



Date	Tornado	Hail	Wind	Total Reports	Max SPC Category	Event Severity[a]	Initializations Missing (UTC)
27 April	5	11	6	22	slight	mid	
28 April	2	102	36	140	enhanced	high-end	23
03 May	15	50	79	144	enhanced	high	21, 23
04 May	17	6	345	368	moderate	high	
05 May	0	0	0	0	marginal	low	23, 00, 01, 02, 03
06 May	0	0	3	3	slight	low	
07 May	0	3	20	23	slight	mid	
10 May	1	41	3	45	enhanced	mid	
12 May	0	0	2	2	marginal	low	21
13 May	1	8	21	30	marginal	mid	03
14 May	0	30	12	42	slight	mid	
17 May	11	46	27	84	moderate	high	17, 18, 23,01
18 May	4	6	33	43	slight	mid	01
19 May	0	1	2	3	marginal	low	22, 23, 01, 02, 03
20 May	0	5	4	9	slight	low	02, 03
21 May	0	4	3	7	slight	low	01, 02
23 May	16	26	54	96	enhanced	high	18
24 May	20	24	6	50	slight	mid	01



25 May	0	5	30	35	slight	mid	21
26 May	31	115	54	200	moderate	high	
27 May	4	22	37	63	enhanced	high	
28 May	1	0	19	20	marginal	mid	
01 June	0	2	2	4	marginal	low	
02 June	1	1	6	8	marginal	low	17, 18, 02, 03
03 June	1	2	30	33	slight	mid	23, 00, 01, 02, 03
04 June	0	4	6	10	slight	low	
13 July	1	0	9	10	marginal	low	
14 July	1	1	8	10	marginal	low	
22 July	0	1	5	6	marginal	low	17, 18[b]
23 July	1	1	1	3	marginal	low	
10 December	85	16	163	264	moderate	high	

Note: [a]Event Severity is defined in the following way: "low" severity events have ≤ 10 total storm reports in the WoFS domain; "mid" severity events have >10 and ≤ 50 total storm reports in the WoFS domain; "high" severity events have > 50 total storm reports in the WoFS domain. [b]Both initializations are present in the FSS method but are not present in the object-based method.

Table 2

Summary of thresholds and radii used in the FSS methods for one-hour precipitation, six-hour precipitation, and reflectivity variables.

	One-hour Precipitation	Six-hour Precipitation	Reflectivity
Thresholds	0.1, 0.25, 0.5, 1.0 (in.)	0.1, 0.5, 1.0, 2.0 (in.)	30, 40, 50 (dBZ)
Radii	6, 12, 24 (km)	6, 12, 24 (km)	6, 12, 24 (km)

Table 3

Contingency table for the object-based methodology for matches, misses, and false alarms.

		Forecast	
		yes	no
Observed	yes	matches	misses
	no	false alarms	N/A

Table 4

Summary of all 2021 cases with respect to maximum SPC risk category and event severity.

SPC Category	Event Severity		
	High	Mid	Low
Marginal	0	2	9
Slight	0	7	4
Enhanced	4	1	0
Moderate	4	0	0
Total	8	10	13

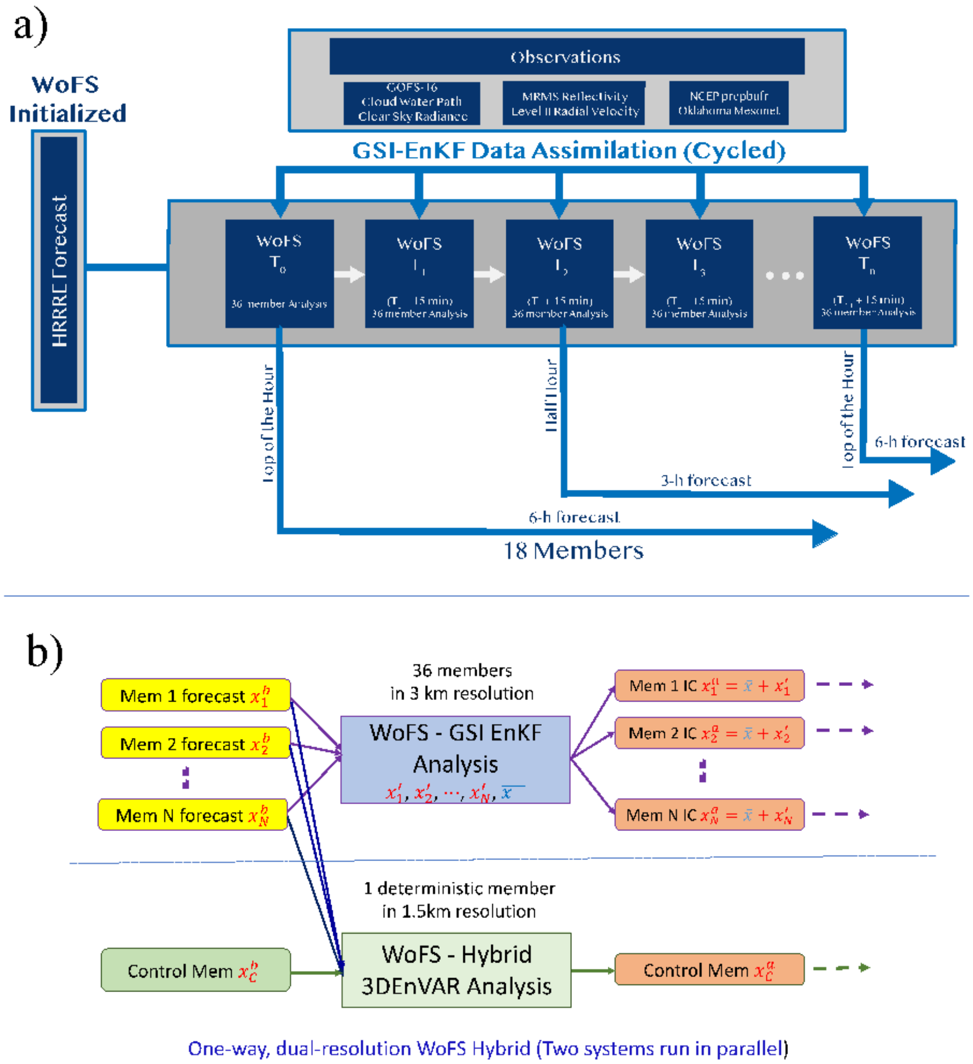

a)
WoFS Initialized
HRRRE Forecast
Observations
GOES-16 Cloud Water Path Clear Sky Radiance
MRMS Reflectivity Level II Radial Velocity
NCEP prepbufr Oklahoma Mesonet
GSI-EnKF Data Assimilation (Cycled)
WoFS T_0 36-member Analysis
WoFS T_1 (T_0 + 15 min) 36-member Analysis
WoFS T_2 (T_1 + 15 min) 36-member Analysis
WoFS T_3 (T_2 + 15 min) 36-member Analysis
WoFS T_n (T_{n-1} + 15 min) 36-member Analysis
Top of the Hour
Half Hour
Top of the Hour
6-h forecast
3-h forecast
6-h forecast
18 Members
b)
Mem 1 forecast x_1^b
Mem 2 forecast x_2^b
Mem N forecast x_N^b
36 members in 3 km resolution
WoFS - GSI EnKF Analysis
$x_1', x_2', \cdots, x_N', \overline{x}$
Mem 1 IC $x_1^a = \bar{x} + x_1'$
Mem 2 IC $x_2^a = \bar{x} + x_2'$
Mem N IC $x_N^a = \bar{x} + x_N'$
1 deterministic member in 1.5km resolution
Control Mem x_C^b
WoFS - Hybrid 3DEnVAR Analysis
Control Mem x_C^a
One-way, dual-resolution WoFS Hybrid (Two systems run in parallel)

Figure 1: Flowchart of the workflow for a) 36-member WoFS and b) deterministic WoF-Hybrid. Both systems are continuously cycled every 15 minutes. Forecasts from both systems, launched at the top of each hour, are projected 6-h. All WoFS forecasts are derived from 18 ensemble members. HRRRE members provide boundary conditions for WoFS, and the HRRRE control member provides boundary conditions for WoF-Hybrid. HRRRE also provides initial background fields for both systems (Heinselman et al. 2022).

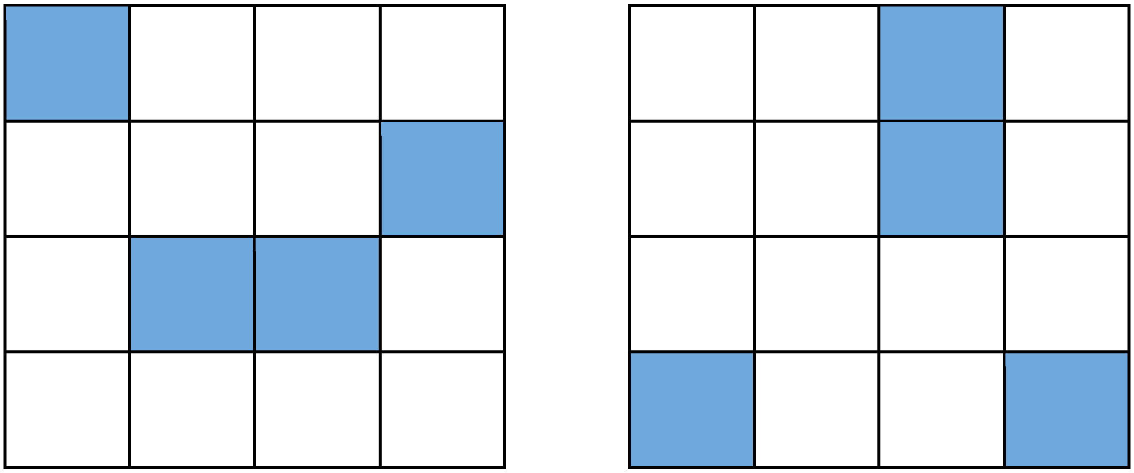


Figure 2: A hypothetical schematic of a forecast (a) and corresponding observed data (b) on a 4 x 4 grid, where each rectangle represents a grid point. A blue rectangle means the meteorological parameter was forecast and/or observed to have exceeded a prescribed threshold at that grid point; a blank rectangle means the parameter was not forecast and/or observed to have exceeded the threshold.

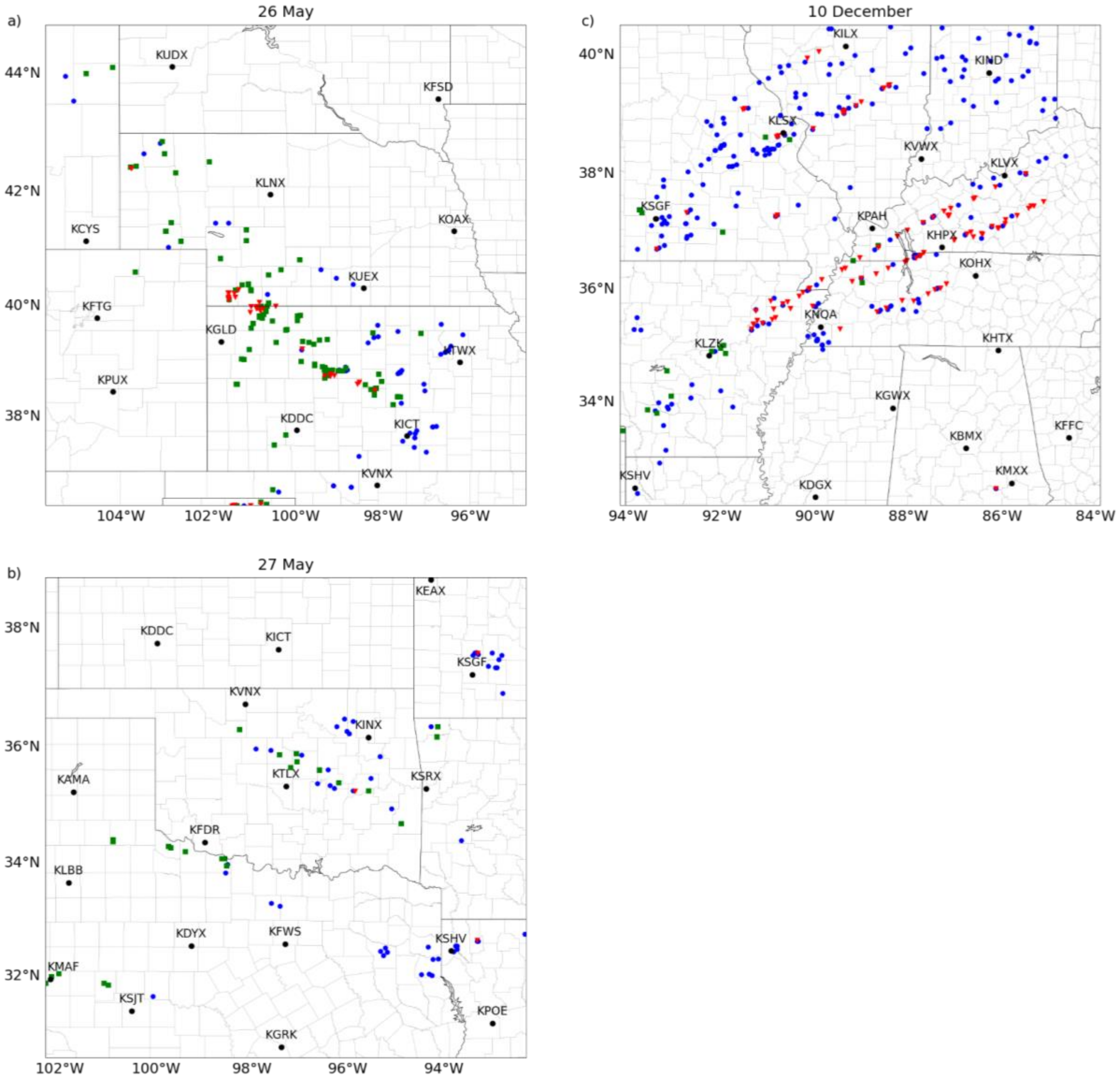


Figure 3: WoFS/WoF-Hybrid domain for a) 26 May, b) 27 May, and c) 10 December cases in 2021. NEXRAD radar locations are labeled with their locations represented with black circles. The red triangles, green squares, and blue circles are tornado, hail, and wind reports from NWS/SPC offices, respectively. The storm reports from between 1700-0900 UTC are depicted, which are the time period covered by the HWT Spring Experiments analyzed in this study.

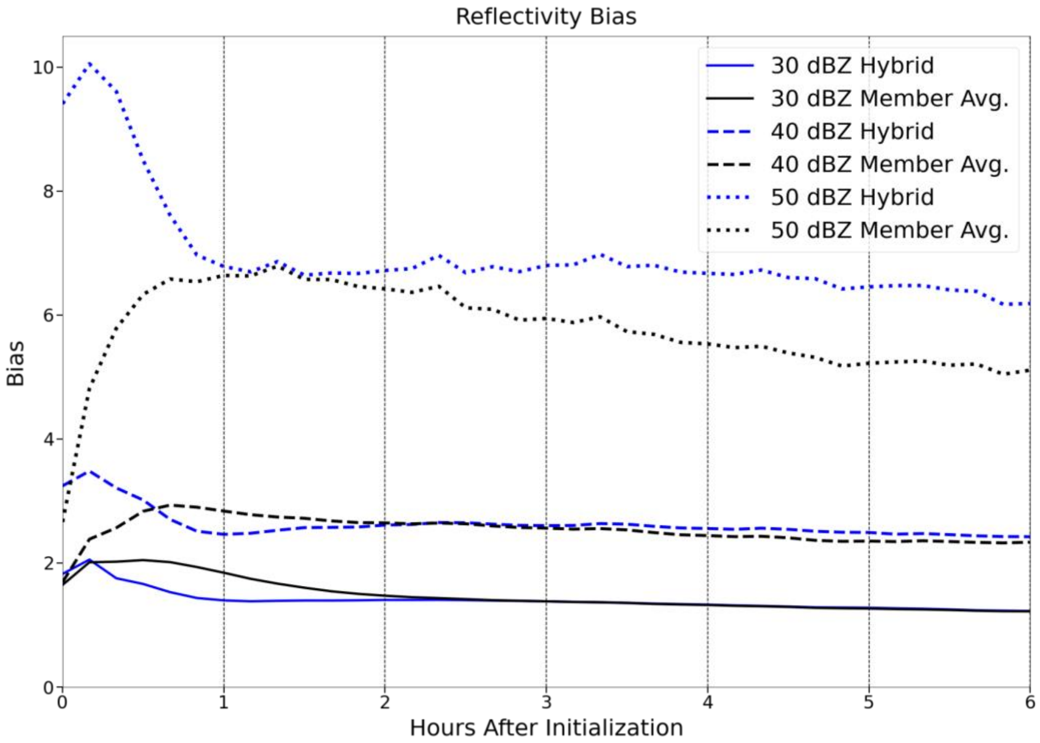


Figure 4: 2021 average WoF-Hybrid and WoFS member reflectivity bias at each ten-minute time step. Three thresholds, 30 dBZ (solid), 40 dBZ (dashed), and 50 dBZ (dotted), are shown for both WoF-Hybrid (blue) and WoFS member average (black).

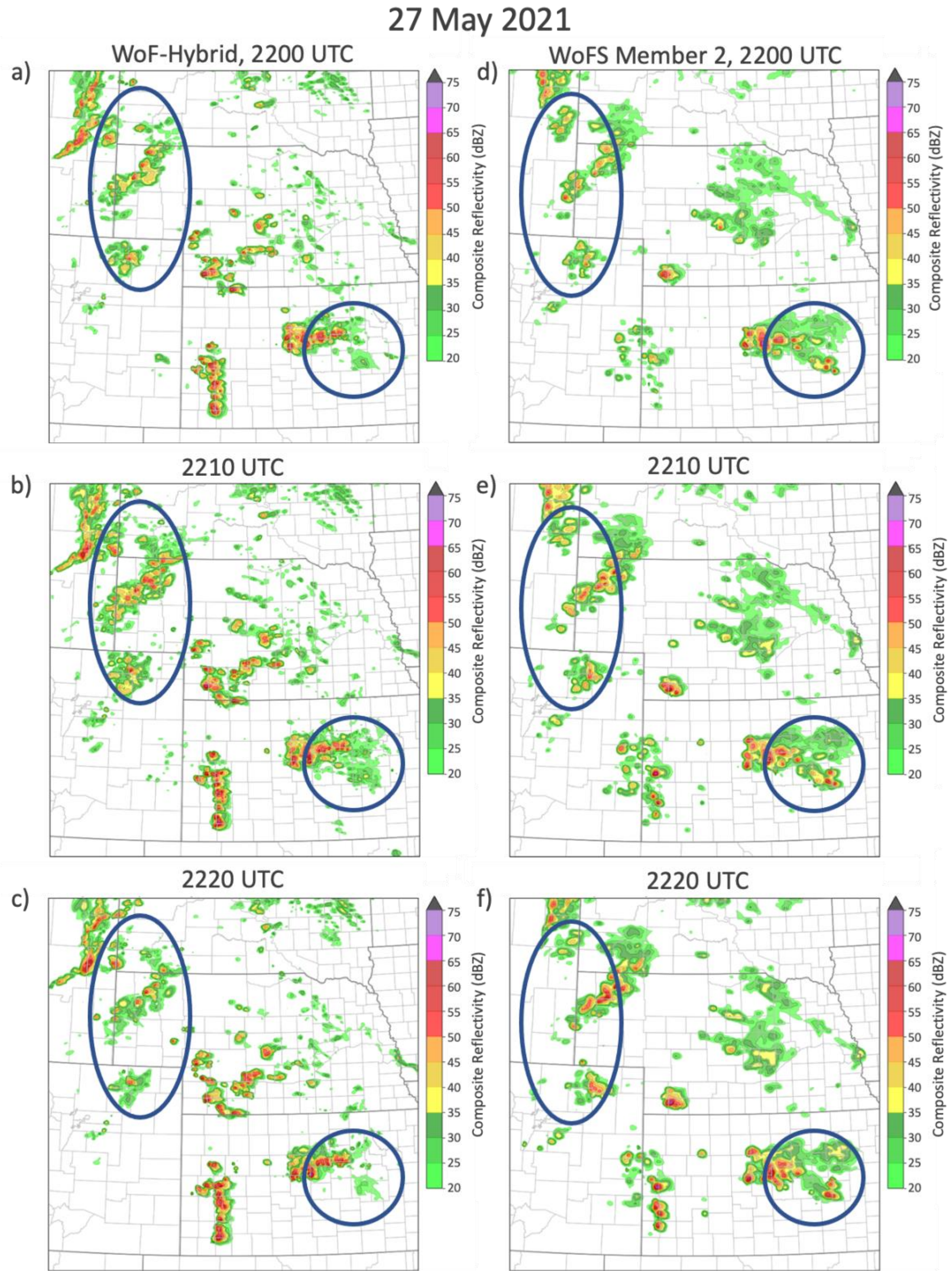


Figure 5: A spatial representation of composite reflectivity on 27 May 2021 from a-c) WoF-Hybrid and d-f) WoFS member 2 2200 UTC initialization at a,d) 2200 UTC, b,e) 2210 UTC, and c,f) 2220 UTC (NOAA NSSL 2022b).

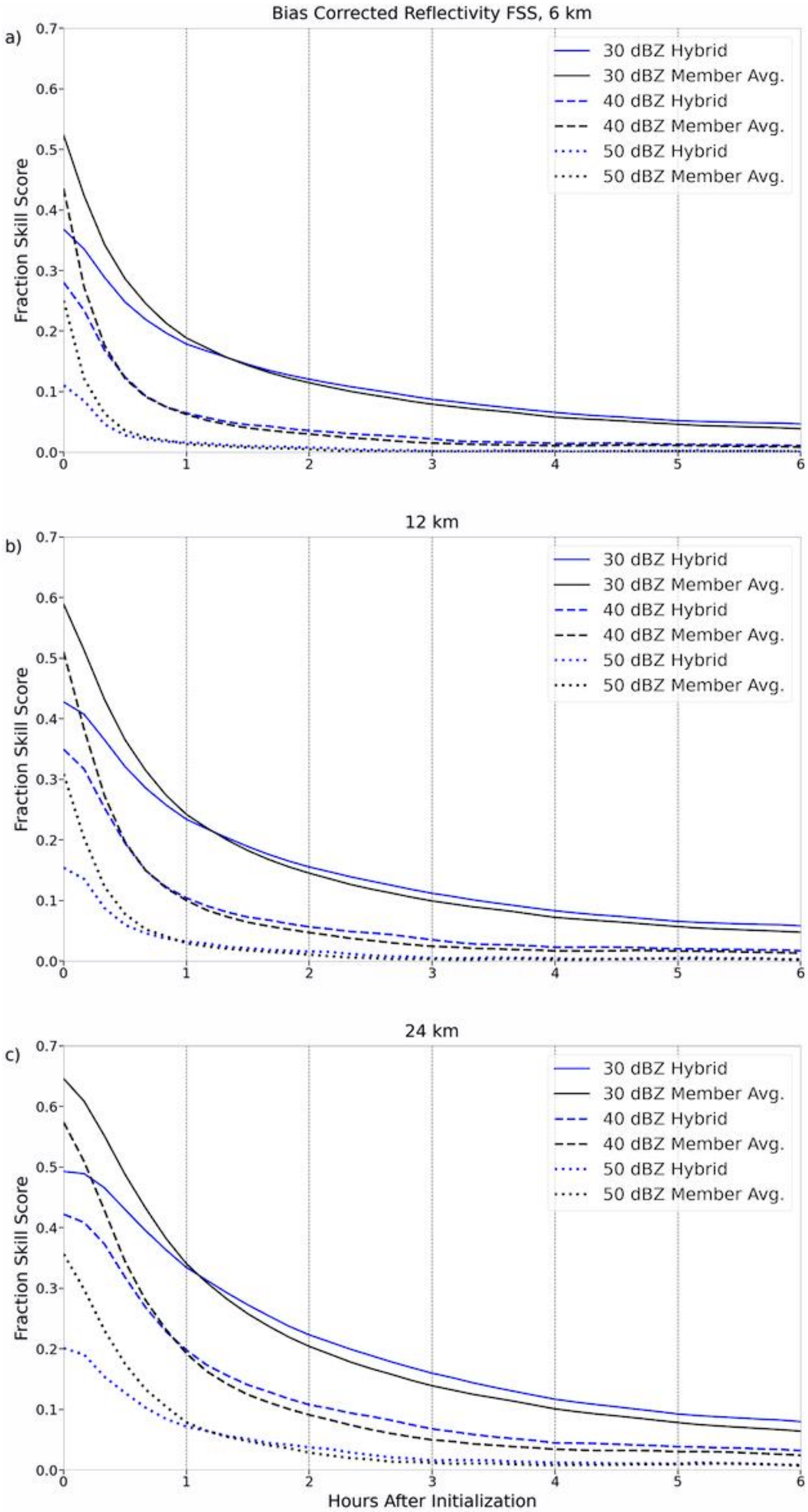


Figure 6: 2021 average WoF-Hybrid and WoFS member bias-corrected FSS for reflectivity forecasts at each ten-minute time step. Three thresholds, 30 dBZ (solid), 40 dBZ (dashed), and 50 dBZ (dotted), are shown for both WoF-Hybrid (blue) and WoFS member average (black). FSSs are calculated with a radius of influence of a) 6km b) 12 km c) 24 km.

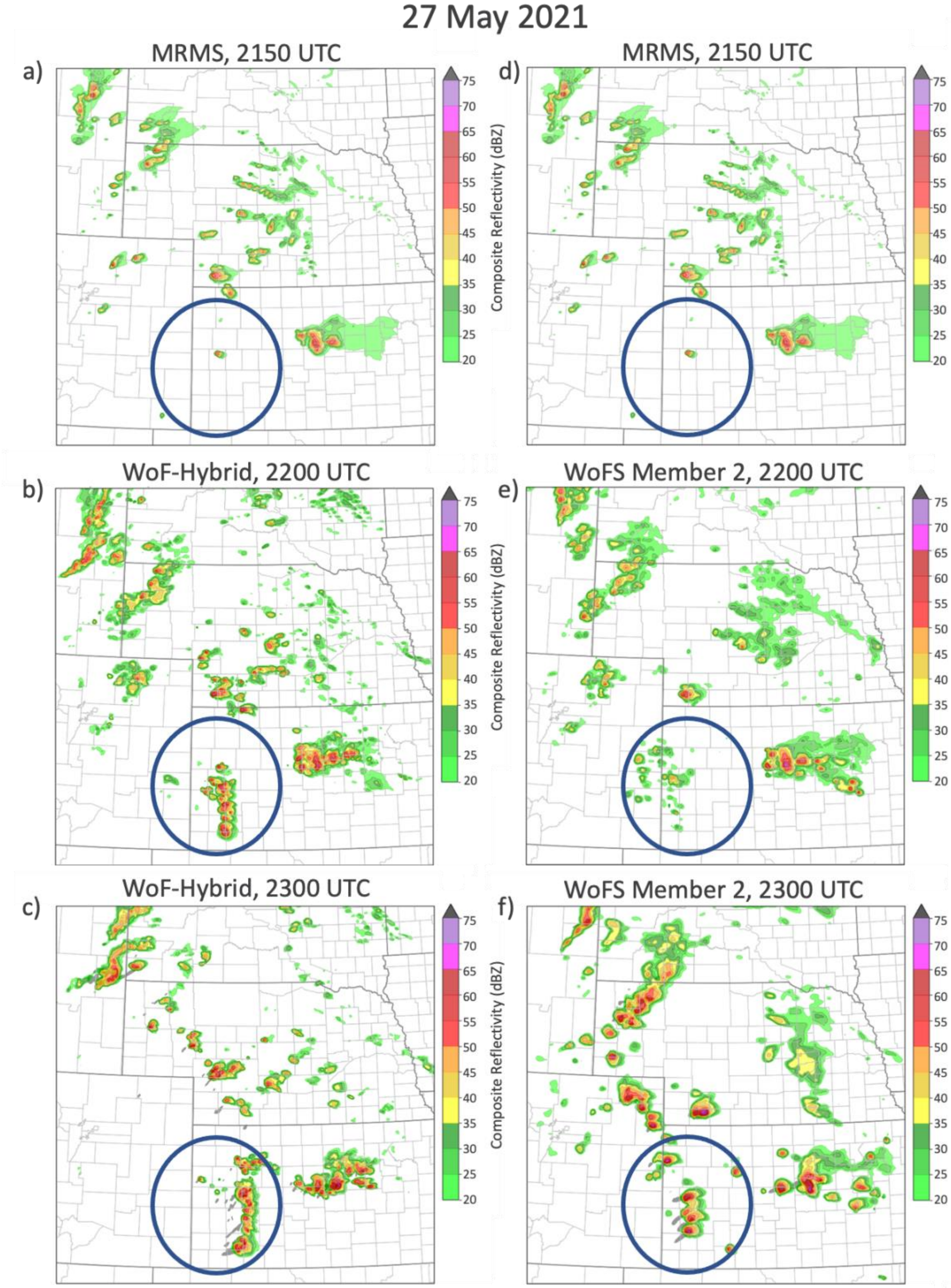
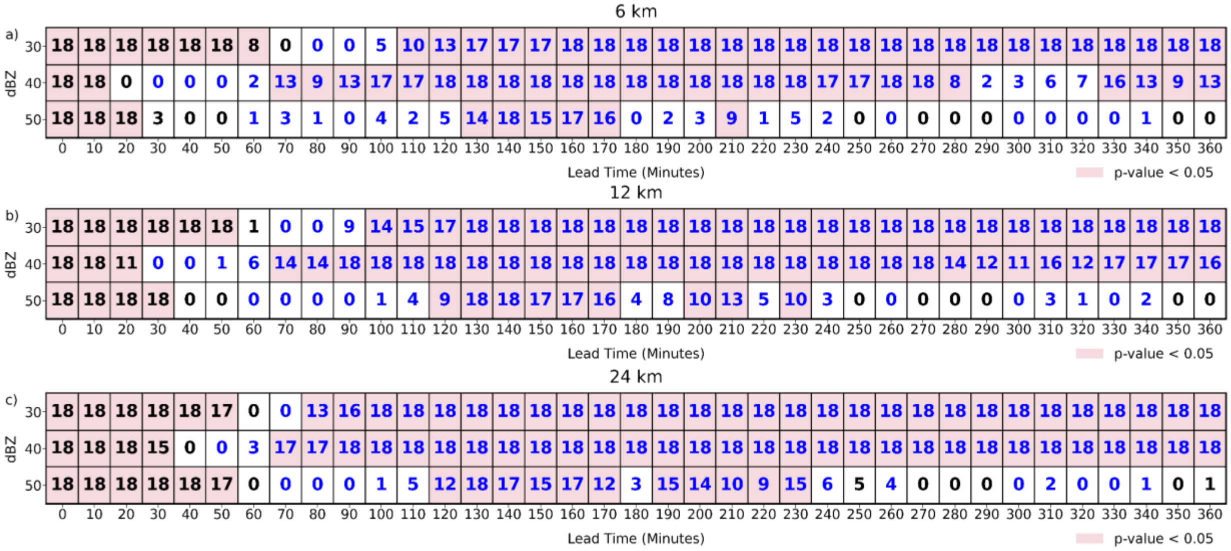


Figure 7: A spatial representation of composite reflectivity on 27 May 2021 from a,d) MRMS, b-c) WoF-Hybrid, and e-f) WoFS member 2 with the 220o UTC model initialization at a,d) 2150 UTC, b,e) 2200 UTC, and c,f) 2300 UTC (NOAA NSSL 2022b).



dBZ	0	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200	210	220	230	240	250	260	270	280	290	300	310	320	330	340	350	360
30	18	18	18	18	18	18	8	0	0	0	5	10	13	17	17	17	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18
40	18	18	0	0	0	0	2	13	9	13	17	17	18	18	18	18	18	18	18	18	18	18	18	18	17	17	18	18	8	2	3	6	7	16	13	9	13
50	18	18	18	3	0	0	1	3	1	0	4	2	5	14	18	15	17	16	0	2	3	9	1	5	2	0	0	0	0	0	0	0	0	0	1	0	0

dBZ	0	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200	210	220	230	240	250	260	270	280	290	300	310	320	330	340	350	360
30	18	18	18	18	18	18	1	0	0	9	14	15	17	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18
40	18	18	11	0	0	1	6	14	14	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	14	12	11	16	12	17	17	17	16
50	18	18	18	18	0	0	0	0	0	0	1	4	9	18	18	17	17	16	4	8	10	13	5	10	3	0	0	0	0	0	0	3	1	0	2	0	0

dBZ	0	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200	210	220	230	240	250	260	270	280	290	300	310	320	330	340	350	360
30	18	18	18	18	18	17	0	0	13	16	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18
40	18	18	18	15	0	0	3	17	17	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18
50	18	18	18	18	18	17	0	0	0	0	1	5	12	18	17	15	17	12	3	15	14	10	9	15	6	5	4	0	0	0	0	2	0	0	1	0	1

Figure 8: Results from a paired t-test of differences between WoF-Hybrid and WoFS reflectivity average FSS scores seen in Figure 4 at each threshold and radius of influence. A box is shaded pink if the difference between WoF-Hybrid and the WoFS member average FSSs is statistically significant. The number within a box represents the number of individual member FSSs that are significantly different from the WoF-Hybrid FSS. A black (blue) number indicates the WoFS member average FSS is larger (smaller) in value than the WoF-Hybrid FSS.

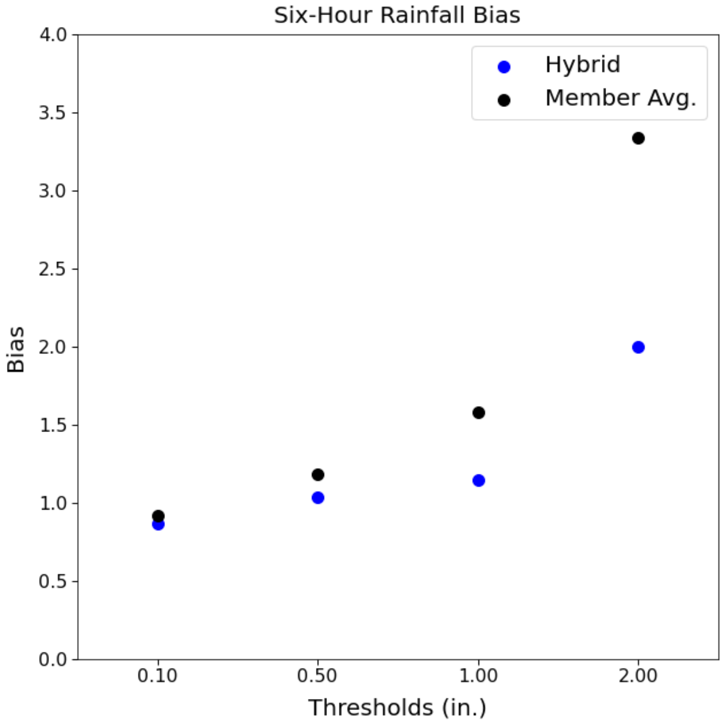


Figure 9: 2021 average WoF-Hybrid and WoFS member six-hour precipitation bias with respect to each rainfall threshold. Blue points correspond to WoF-Hybrid, while black points correspond to the WoFS member average.

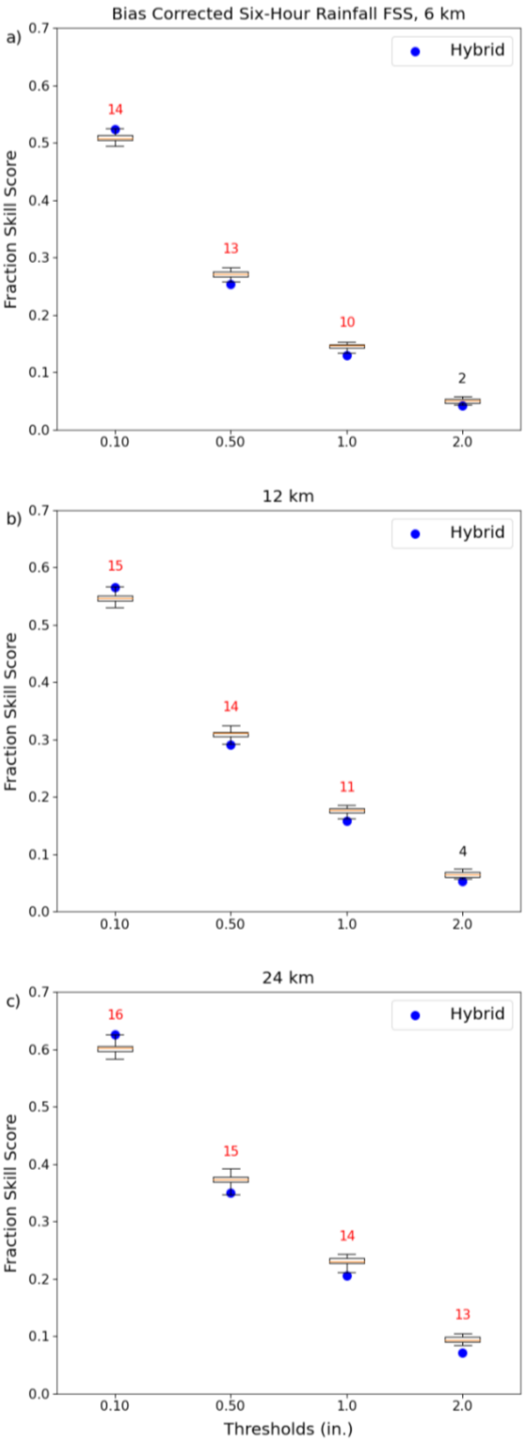


Figure 10: 2021 average WoF-Hybrid and WoFS member bias-corrected FSS for 6-hr precipitation forecasts at four rainfall thresholds. The boxplots represent the FSSs of the 18 WoFS members, and blue points represent WoF-Hybrid FSSs. FSSs are calculated with a radius of influence of a) 6km b) 12 km c) 24 km. Numbers above boxplots represent the number of individual member FSSs that are significantly different from the WoF-Hybrid FSS. The number is red (black) if the difference between WoF-Hybrid and the WoFS member average FSSs is (isn't) statistically significant (p-value < 0.05).

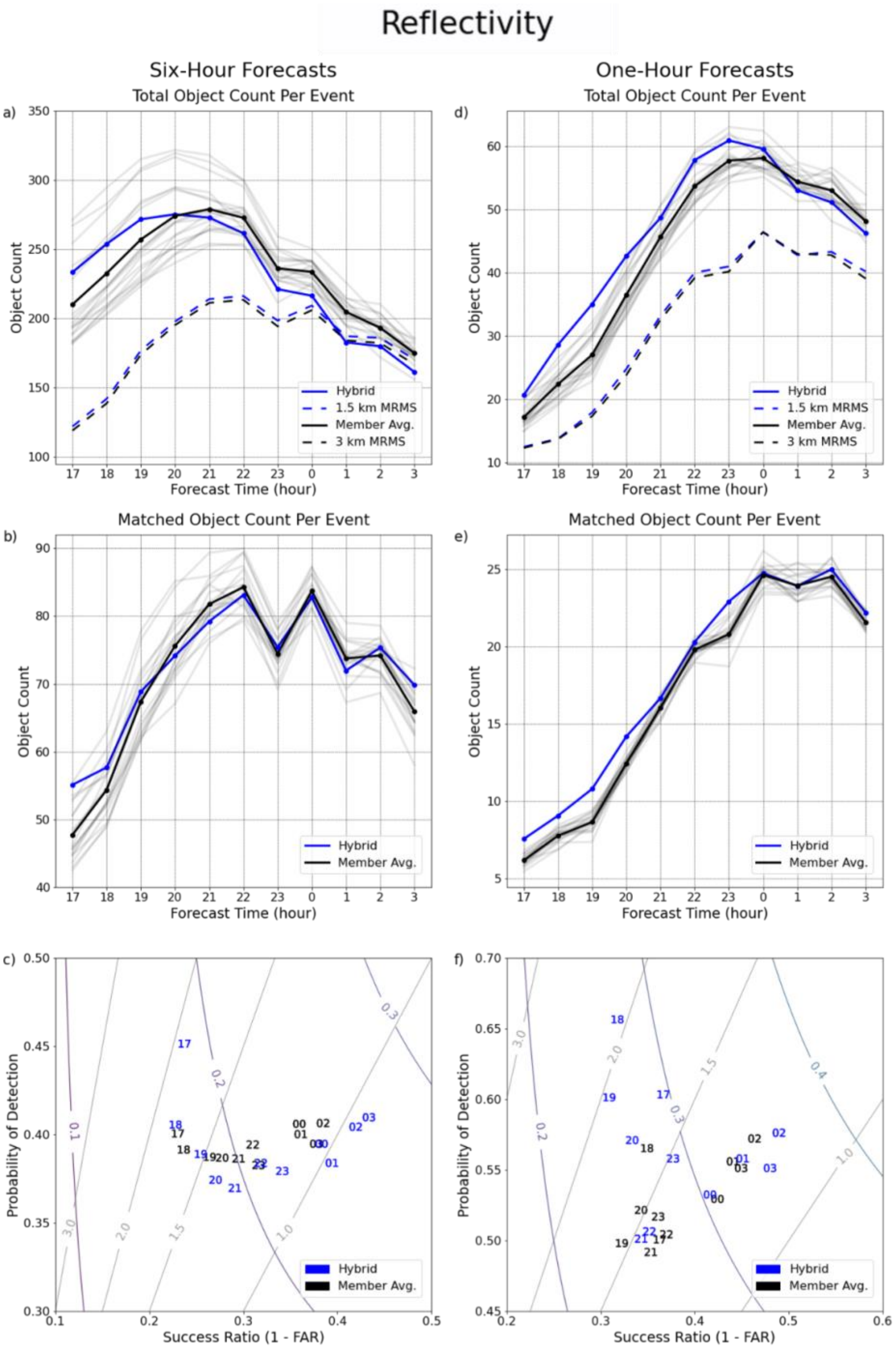


Figure 11: 2021 average a-c) six-hour forecast and d-f) first hour of lead time results for the object-based reflectivity methodology. a,d) Depict the total object count per event binned by initialization time, b,e) Show the matched object count per event by initialization time, and c,f) display a performance diagram with each number representing an initialization time for WoF-Hybrid (solid blue), WoFS member average (solid black), and individual WoFS members (gray). a,d) Show 1.5 km MRMS (dashed blue), and 3 km MRMS (dashed black) total object counts for each six-hour forecast. c,f) Labeled gray lines represent bias contours and labeled curved lines represent CSI.

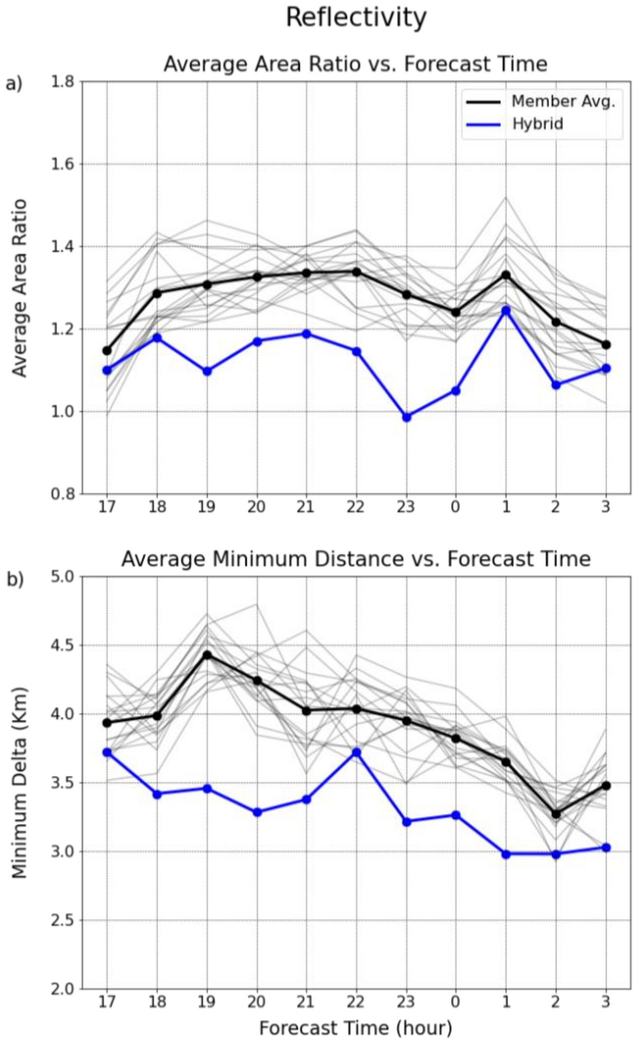


Figure 12: a) Average area ratio and b) Average minimum distance of matched reflectivity objects binned by forecast initialization time for all 31 cases. Blue lines represent WoF-Hybrid. Black (gray) lines represent the WoFS member average (individual WoFS members).

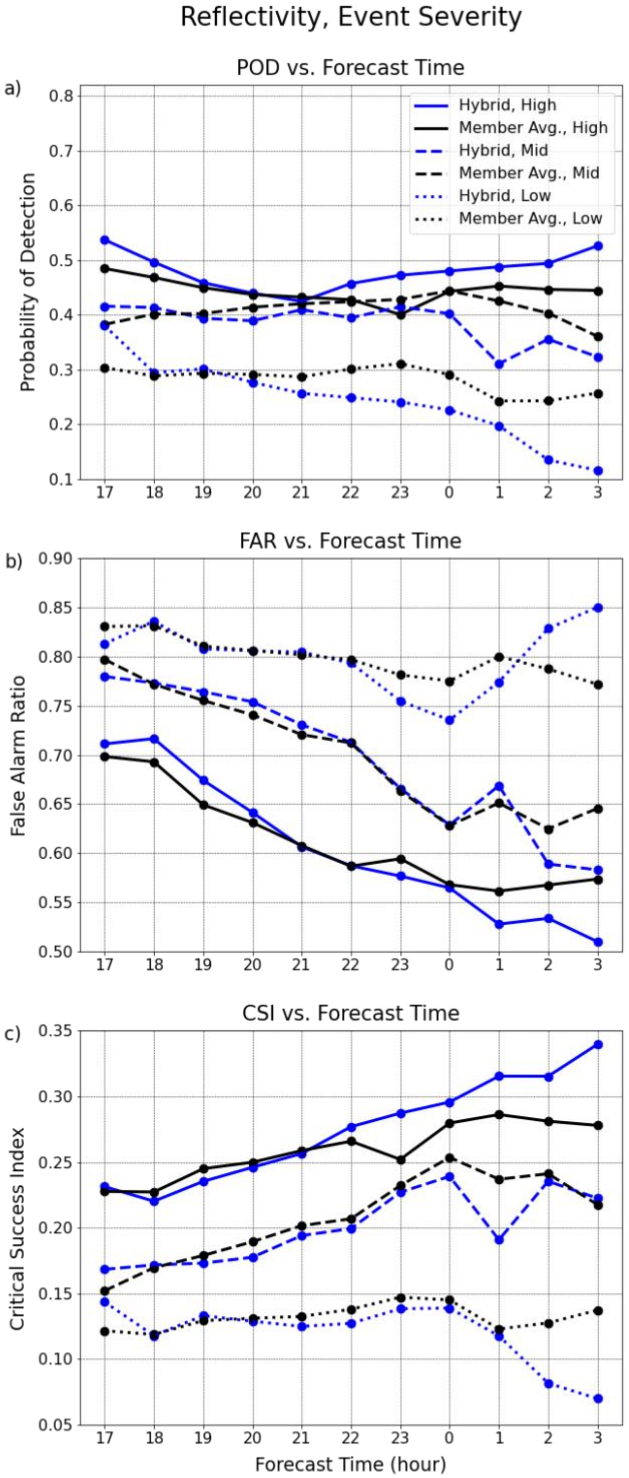


Figure 13: Reflectivity a) POD, b) FAR, c) CSI binned by forecast initialization time and analyzed via event severity type for all 2021 cases. Solid blue (black) lines represent high-end cases for WoF-Hybrid (WoFS member average). Dashed blue (black) lines represent mid severity cases for WoF-Hybrid (WoFS member average). Dotted blue (black) lines represent low severity cases for WoF-Hybrid (WoFS member average). Each point represents a ten-minute time step.

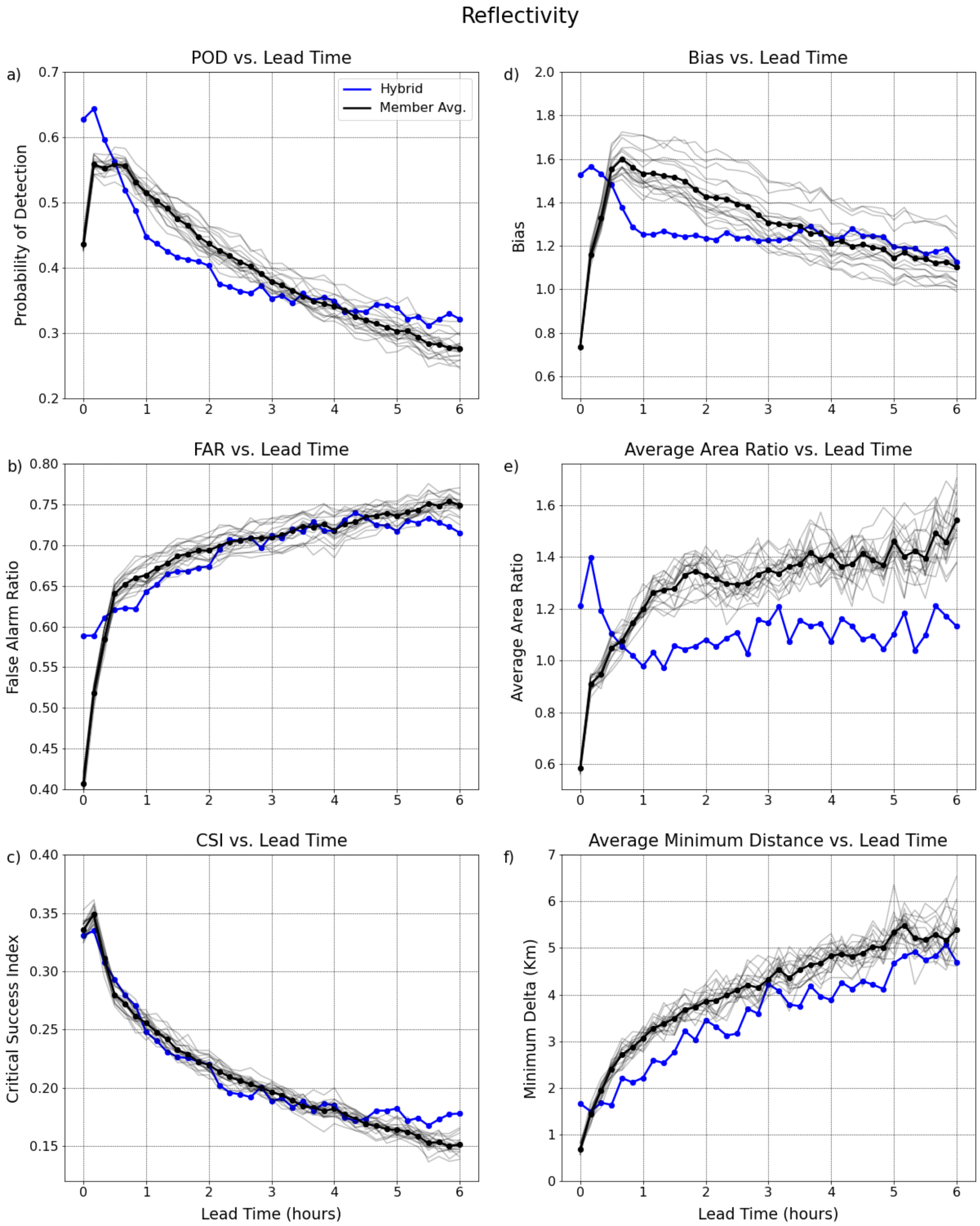


Figure 14: Reflectivity a) POD, b) FAR, c) CSI, d) Bias, e) AAR, f) AMD binned by forecast lead time for all 2021 cases. Blue lines represent WoF-Hybrid. Black (gray) lines represent the WoFS member average (individual WoFS members). Each point represents a ten-minute time step.

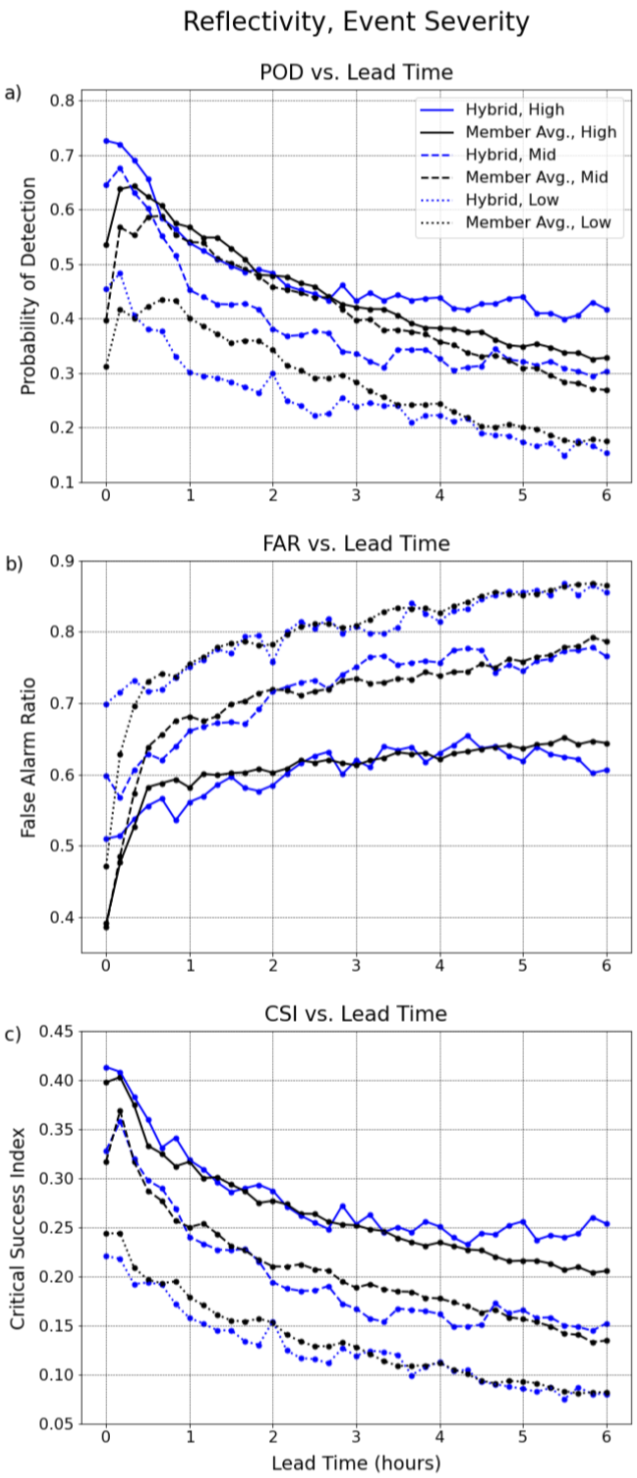


Figure 15: Reflectivity a) POD, b) FAR, c) CSI binned by forecast lead time and analyzed via event severity type for all 2021 cases. Solid blue (black) lines represent high-end cases for WoF-Hybrid (WoFS member average). Dashed blue (black) lines represent mid severity cases for WoF-Hybrid (WoFS member average). Dotted blue (black) lines represent low severity cases for WoF-Hybrid (WoFS member average). Each point represents a ten-minute time step.

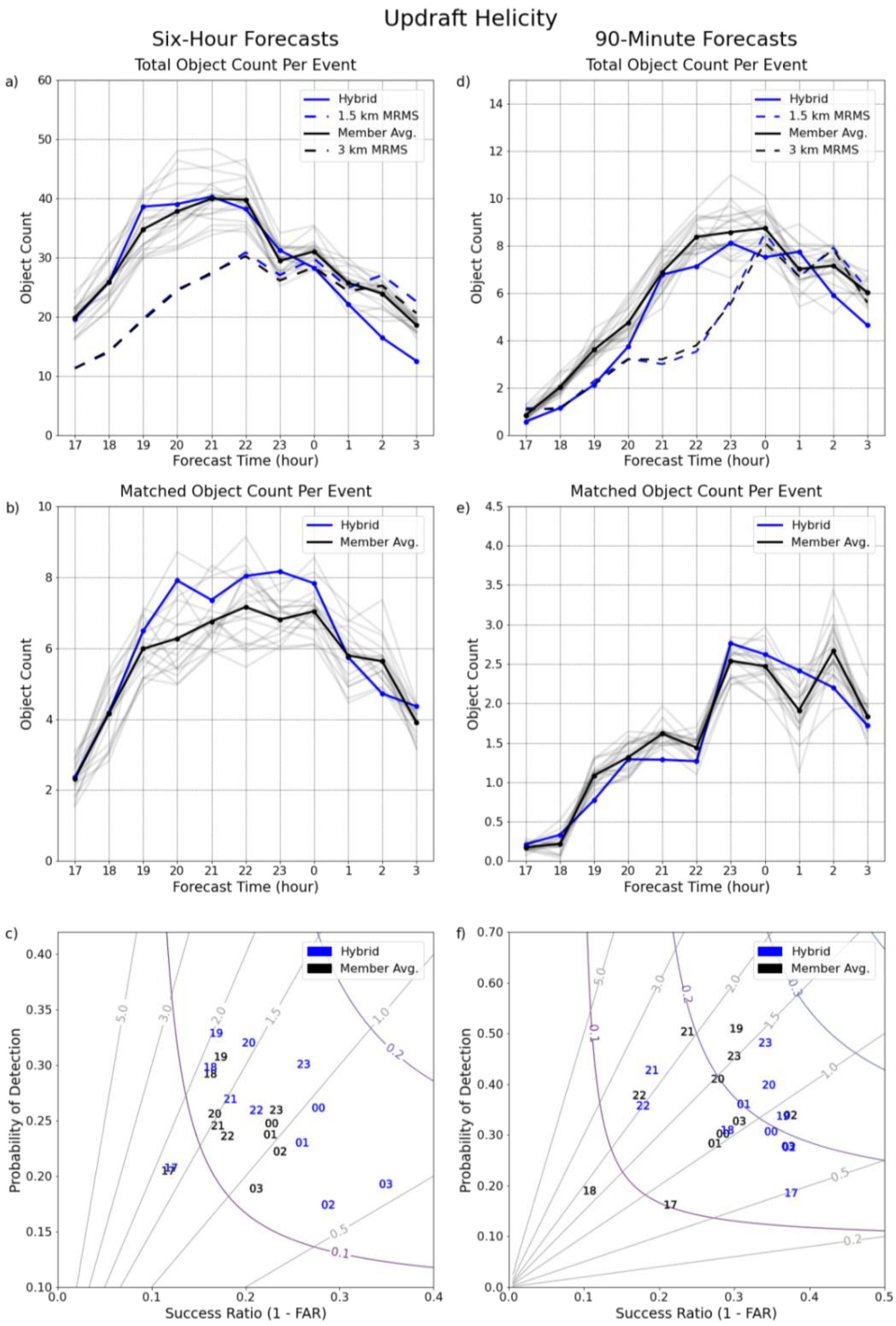


Figure 16: As in Figure 11, but for updraft helicity and 90-minute forecasts instead of 60-minute forecasts for d-f.

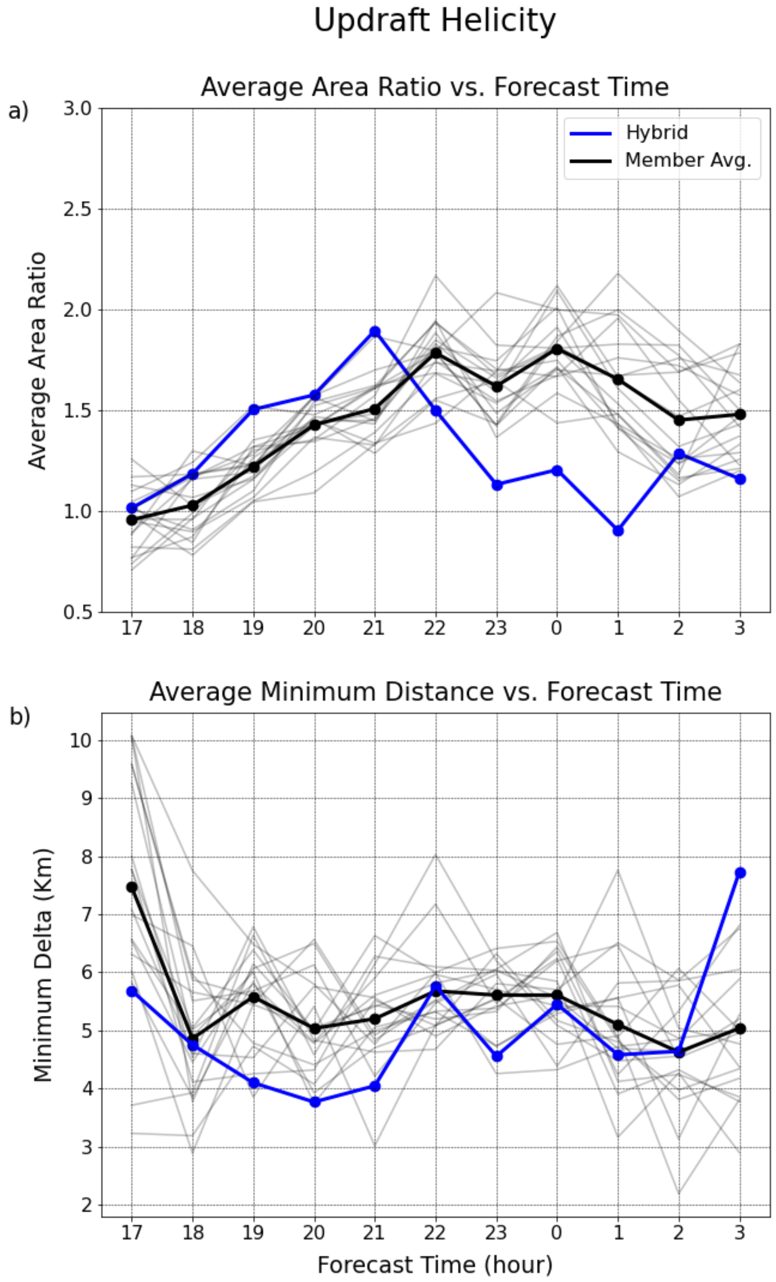


Figure 17: As in Figure 12, but for updraft helicity.

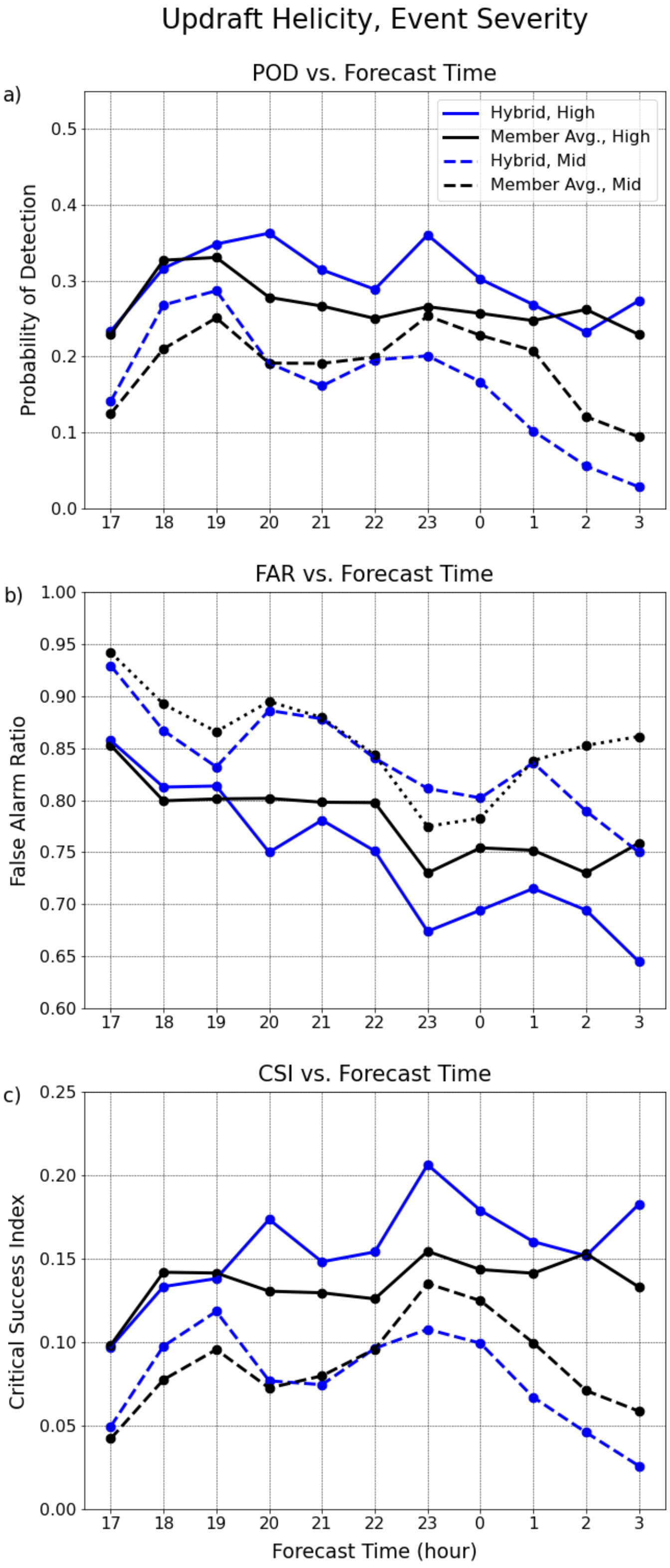


Figure 18: As in Figure 13, but for updraft helicity and without low severity events omitted. Low severity events are omitted from these plots as the value is 0 for a,c and 1 for b at all lead times.

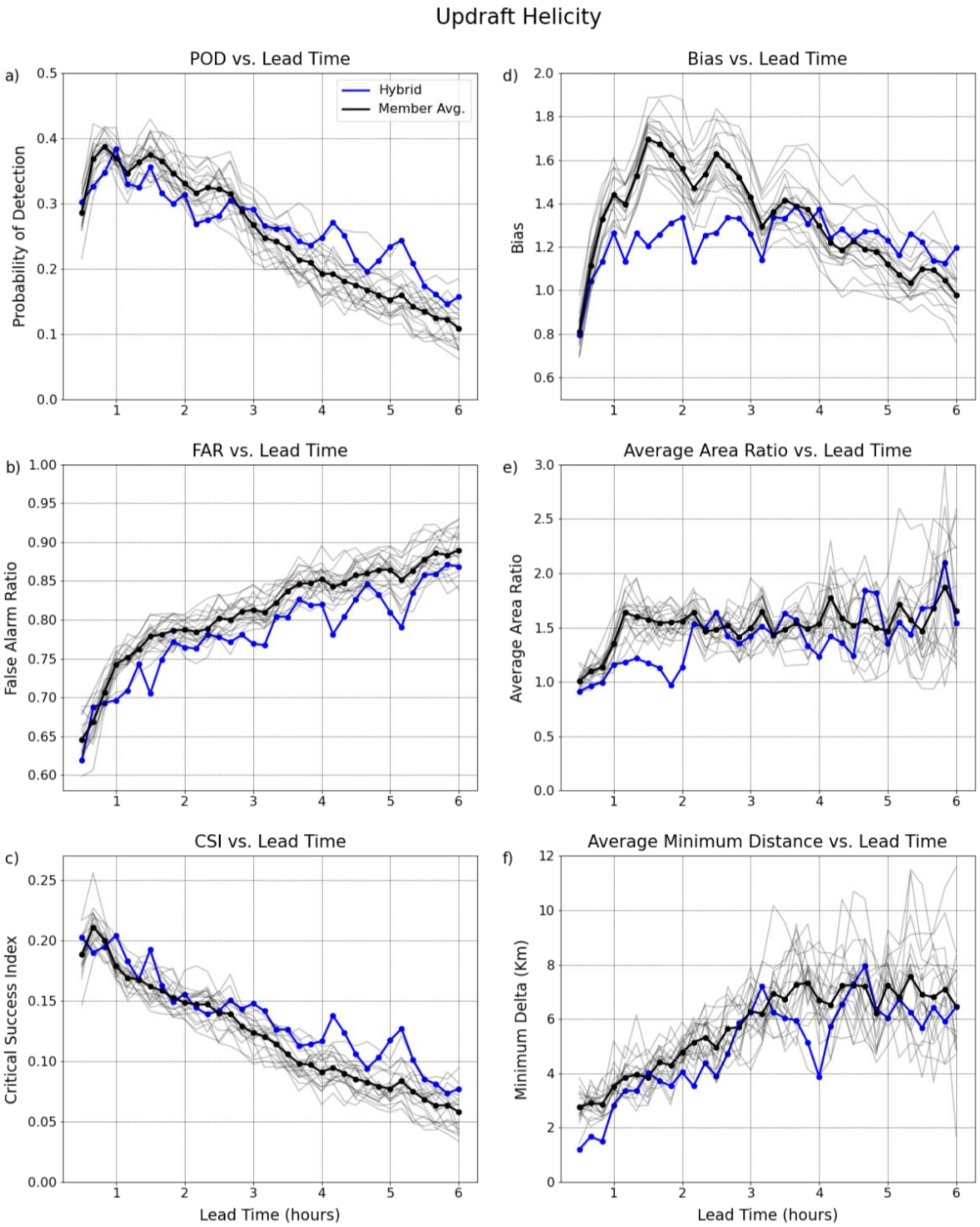


Figure 19: As in Figure 14, but for updraft helicity.

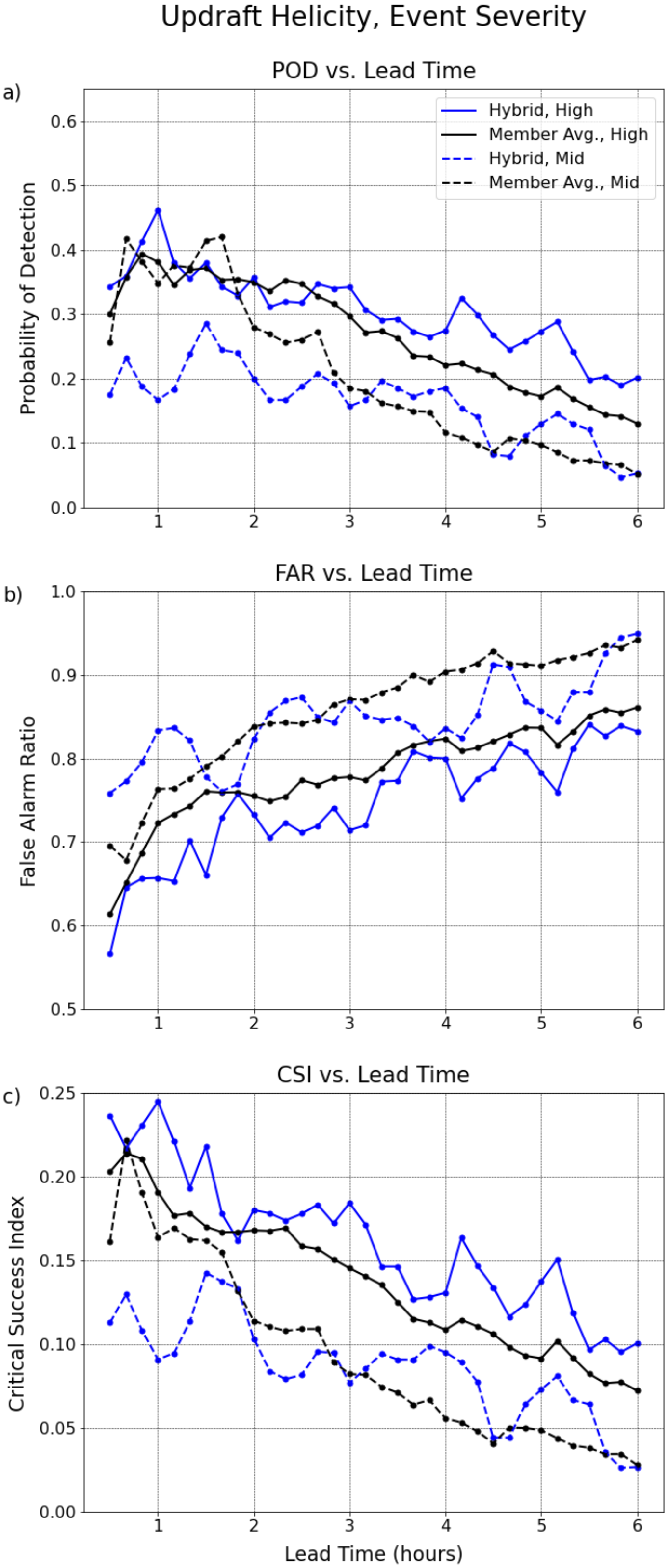


Figure 20: As in Figure 15, but for updraft helicity and without low severity events. Low severity events are omitted from these plots as the value is 0 for a,c and 1 for b at all initialization times.

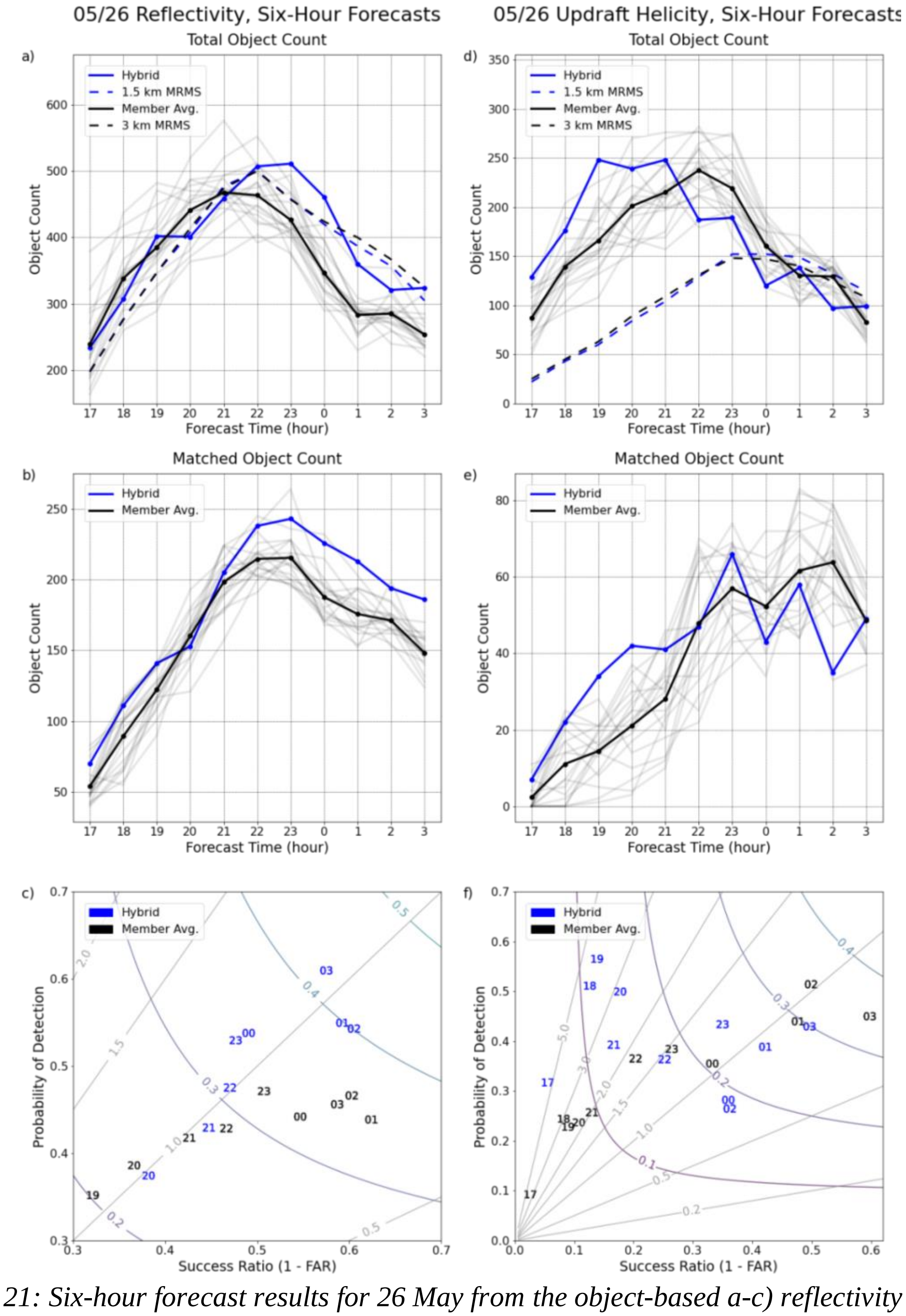


Figure 21: Six-hour forecast results for 26 May from the object-based a-c) reflectivity d-f) UH methods. a,d) Depict the total object count per event binned by initialization time, b,e) show the matched object count per event by initialization time, and c,f) display a performance diagram with each number representing an initialization time for WoF-Hybrid (solid blue), WoFS member average (solid black), and individual WoFS members (gray). a,d) Show 1.5 km MRMS (dashed blue), and 3 km MRMS (dashed black) total object counts for each six-hour forecast. c,f) Labeled gray lines represent bias contours and labeled curved lines represent CSI.

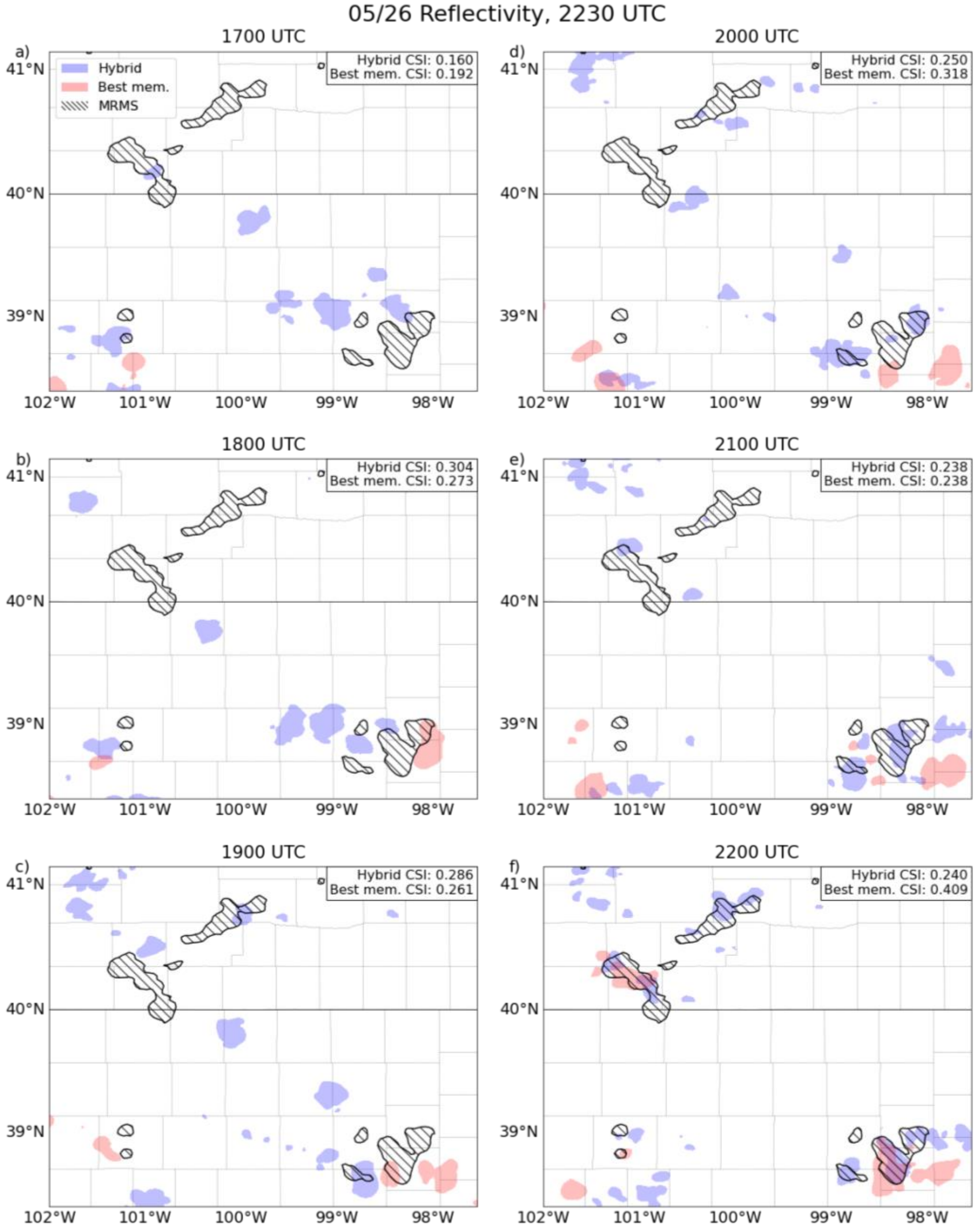


Figure 22: A spatial representation of composite reflectivity near the Kansas/Nebraska border at 2230 UTC on 26 May for six different forecasts initialized at a) 1700 UTC, b) 1800 UTC, c) 1900 UTC, d) 2000 UTC, e) 2100 UTC, and f) 2200 UTC. The domain represented on this plot is zoomed in to focus on the most impactful convection. Hatched contours are MRMS data from 2230 UTC. Red contours are the reflectivity contours above 46.1 dBZ (WoFS object threshold) of the WoFS member with the highest average CSI over the entire domain during the represented times. Blue contours are the reflectivity contours above 47.1 dBZ (WoF-Hybrid object threshold) for WoF-Hybrid. CSIs over the entire domain from the initialization time until 2230 UTC of WoF-hybrid and the highest scoring WoFS member are displayed in the upper righthand corner.

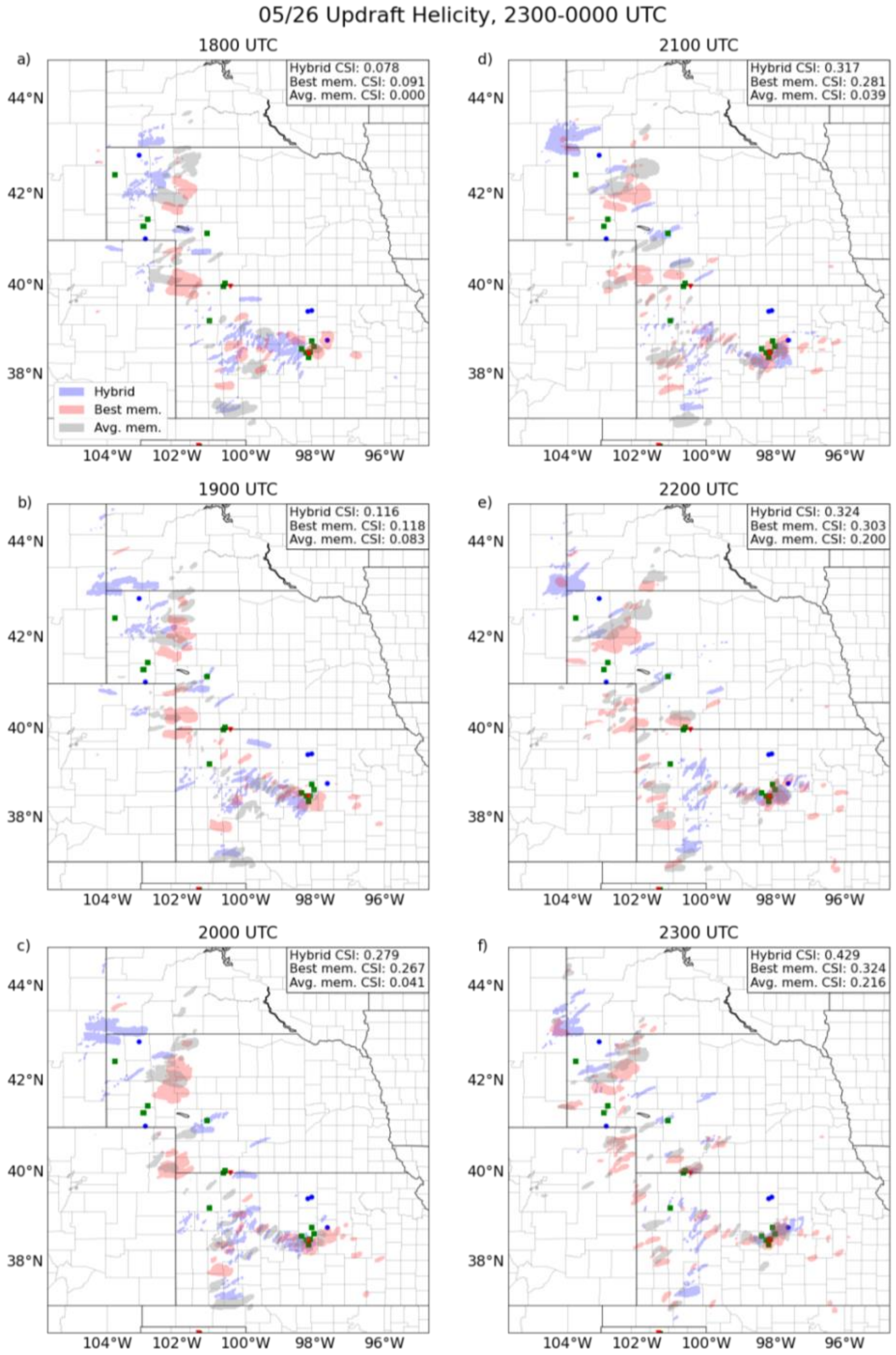


Figure 23: A spatial representation of UH from 2300-0000 UTC on 26-27 May over the WoFS domain of six different forecasts initialized at a) 1800 UTC, b) 1900 UTC, c) 2000 UTC, d) 2100 UTC, e) 2200 UTC, and f) 2300 UTC. Red (gray) contours are the UH swaths above 65 m^2s^{-2} of the WoFS member with the highest average CSI (average scoring WoFS member). Blue contours are the UH swaths above 130 m^2s^{-2} for WoF-Hybrid. The red triangles, green squares, and blue circles are tornado, hail, and wind reports from 2300-0000 UTC, respectively. CSIs of 30-minute UH swaths over the entire domain for WoF-

hybrid, highest scoring WoFS member, and average WoFS member are displayed in the upper righthand corner.

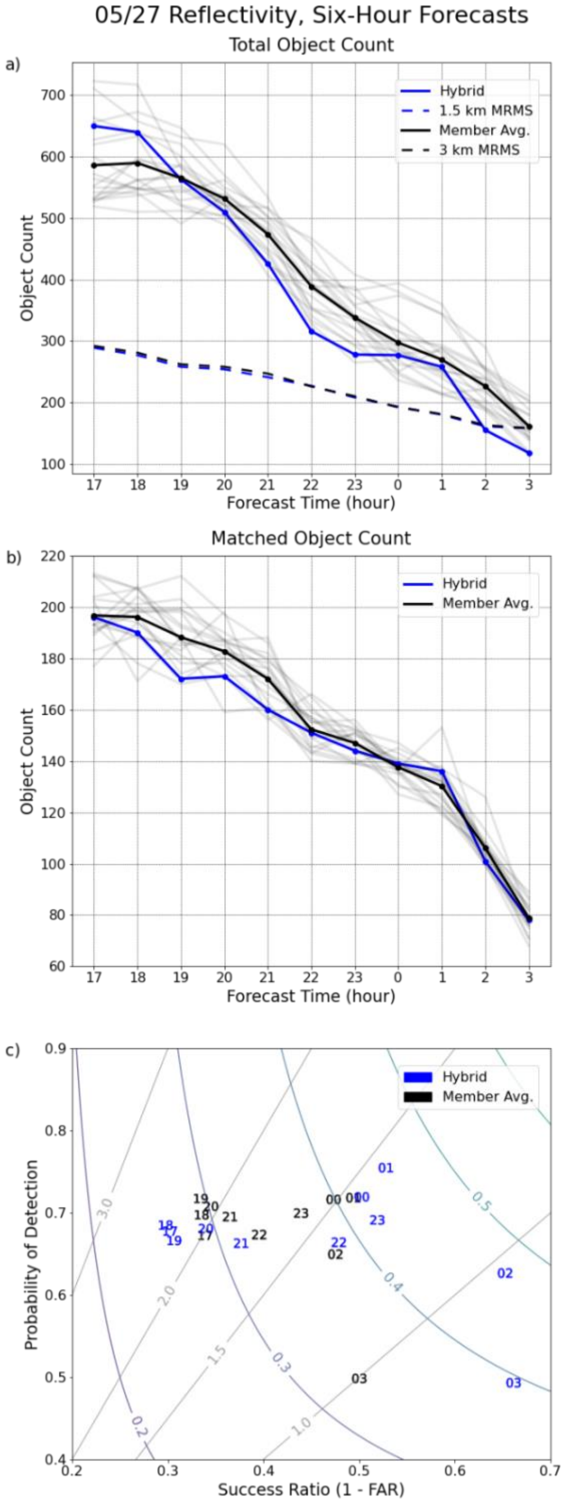


Figure 24: As in Figure 21, but for 27 May. In addition, UH results are omitted.

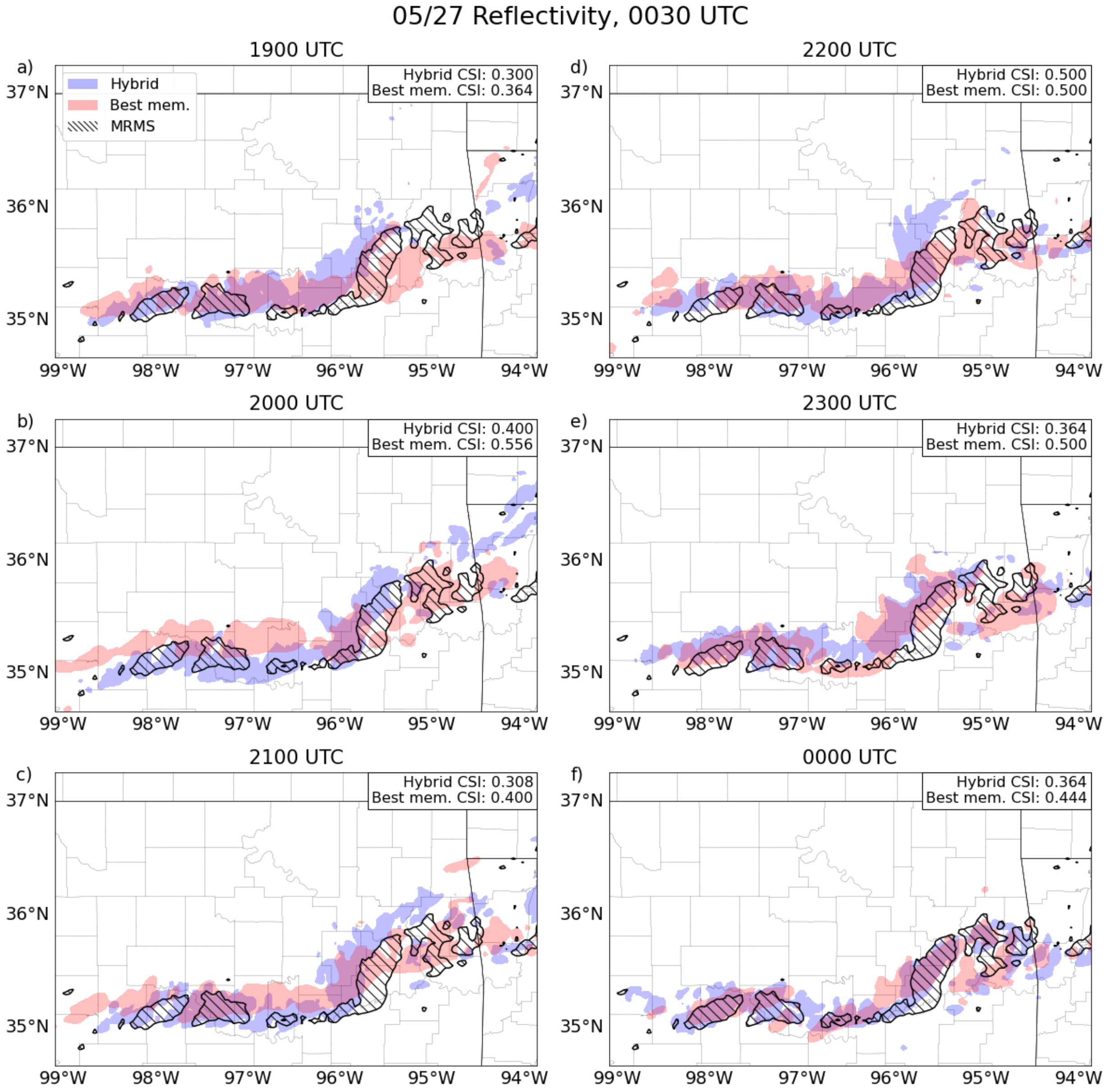


Figure 25: A spatial representation of composite reflectivity over northeastern Oklahoma and western Arkansas at 0030 UTC on 28 May for six different forecasts initialized at a) 1900 UTC on 27 May and ending with f) 0000 UTC on 28 May. The domain represented on this plot is zoomed in to focus on the most impactful convection. Hatched contours are MRMS data from 0030 UTC. Red contours are the reflectivity contours above 46.1 dBZ (WoFS object threshold) of the WoFS member with the highest average CSI over the entire domain during the represented times. Blue contours are the reflectivity contours above 47.1 dBZ (WoF-Hybrid object threshold) for WoF-Hybrid. CSIs over the entire domain from the initialization time until 0030 UTC of WoF-hybrid and the highest scoring WoFS member are displayed in the upper righthand corner.

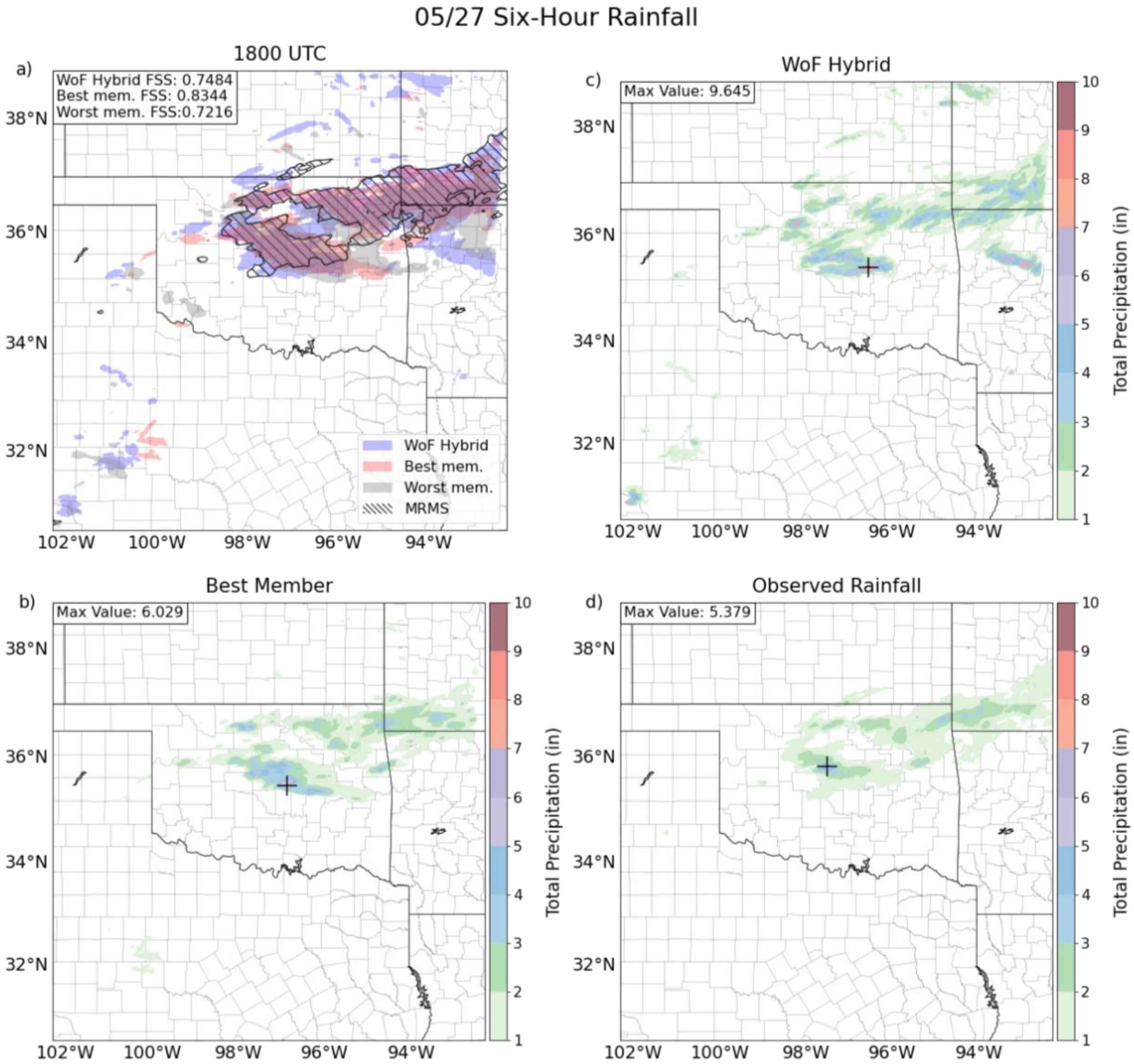


Figure 26: a) A spatial representation of six-hour rainfall higher than one inch for the WoFS member with the highest FSS (red), WoFS member with the lowest FSS (gray), WoF-Hybrid (blue), and observed rainfall (hatched) over the WoFS domain. b-d) Six-hour rainfall accumulations for b) the highest scoring WoFS member, c) WoF-Hybrid, and d) the observed rainfall over the WoFS domain. The rainfall accumulated is displayed from 1800 UTC on 27 May through 0000 UTC on 28 May. a) FSSs over the entire domain of the best WoFS member, worst WoFS member, and WoF-Hybrid are displayed in the upper lefthand corner. b-d) The maximum rainfall amount in inches over each system is displayed in the upper lefthand corner. The location of the maximum rainfall amount is marked by a black plus sign.

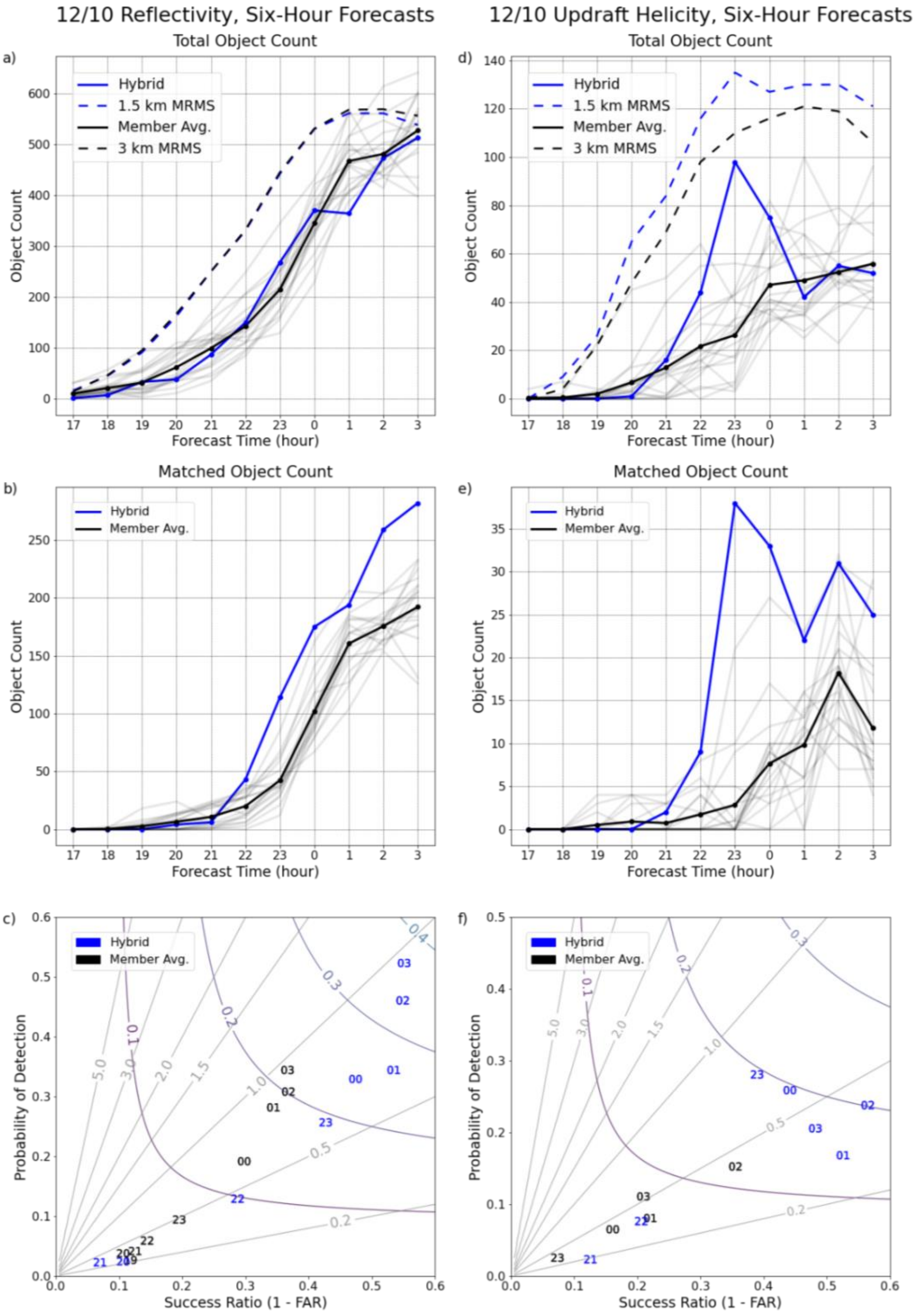


Figure 27: As in Figure 21, but for 10 December.

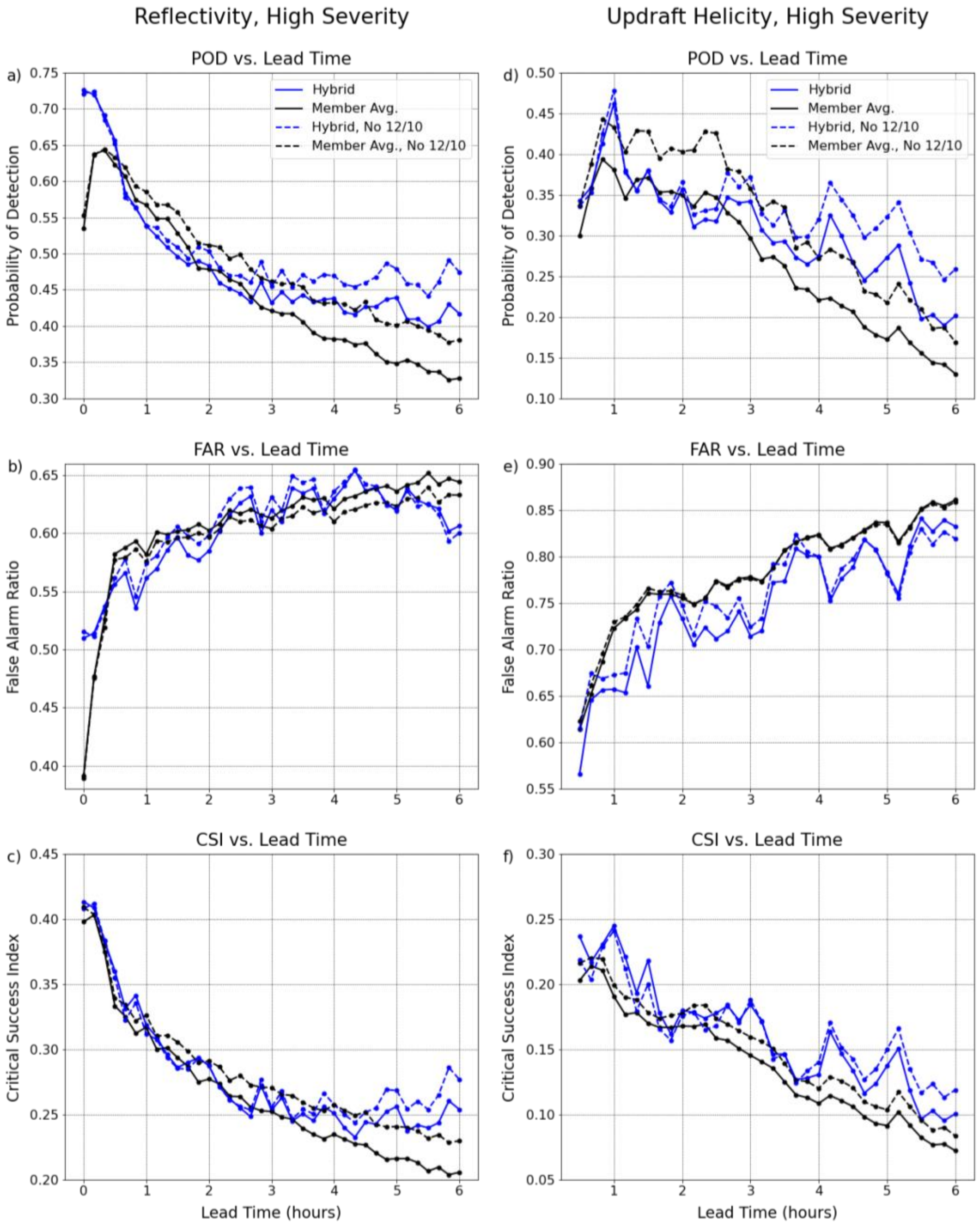


Figure 28: Reflectivity a) POD, b) FAR, c) CSI and UH d) POD, e) FAR, f) CSI for high-end events binned by lead time. Solid blue (black) lines represent high-end cases for WoF-Hybrid (WoFS member average). Dashed blue (black) lines represent high-end events without 10 December for WoF-Hybrid (WoFS member average). Each point represents a ten-minute time step.

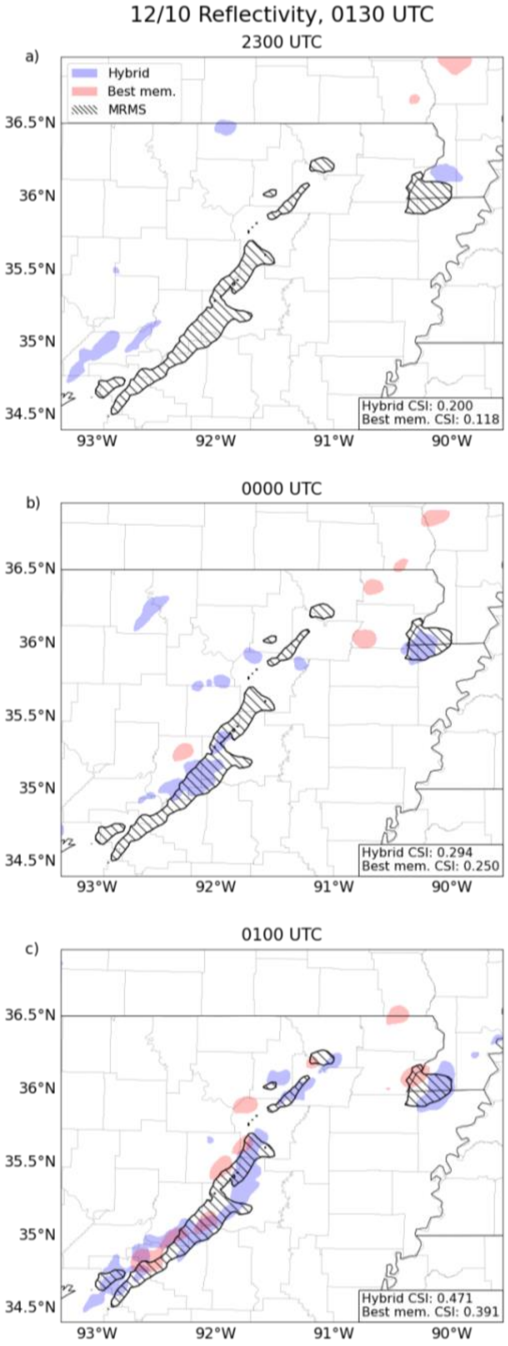


Figure 29: A spatial representation of composite reflectivity in northeastern Arkansas and southeastern Missouri at 0130 UTC on 11 December for three different forecasts starting with a) 2300 UTC on 10 December and ending with c) 0100 UTC on 11 December. The domain represented on this plot is zoomed in to focus on the most impactful convection. Hatched contours are MRMS data from 0130 UTC. Red contours are the reflectivity contours above 46.1 dBZ of the WoFS member with the highest average CSI over the entire domain during the represented times. Blue contours are the reflectivity contours above 47.1 dBZ for WoF-Hybrid. CSIs over the entire domain from the initialization time until 0130 UTC of WoF-hybrid and the highest scoring WoFS member are displayed in the upper righthand corner.

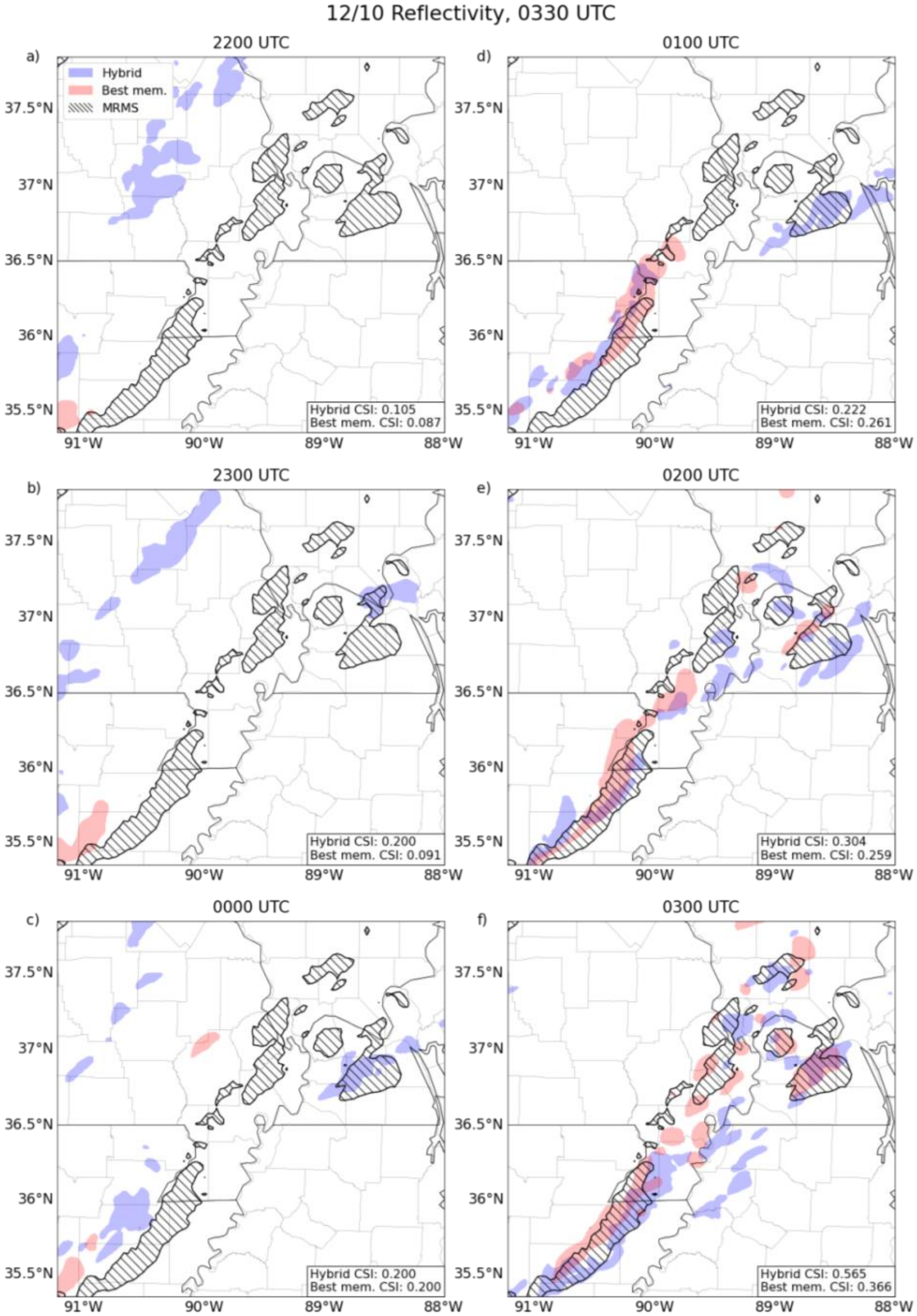


Figure 30: A spatial representation of composite reflectivity in the mid-Mississippi Valley at 0330 UTC on 11 December for three different forecasts starting with a) 2200 UTC on 10 December and ending with c) 0300 UTC on 11 December. The domain represented on this plot is zoomed in to focus on the most impactful convection. Hatched contours are MRMS data from 0330 UTC. Red contours are the reflectivity contours above 46.1 dBZ of the WoFS member with the highest average CSI. Blue contours are the reflectivity contours above 47.1 dBZ for WoF-Hybrid. CSIs over the entire domain from the initialization time until 0330 UTC of WoF-hybrid and the highest scoring WoFS member are displayed in the upper righthand corner. Mayfield, Kentucky is located at 36.74 °N, 88.64 °W.

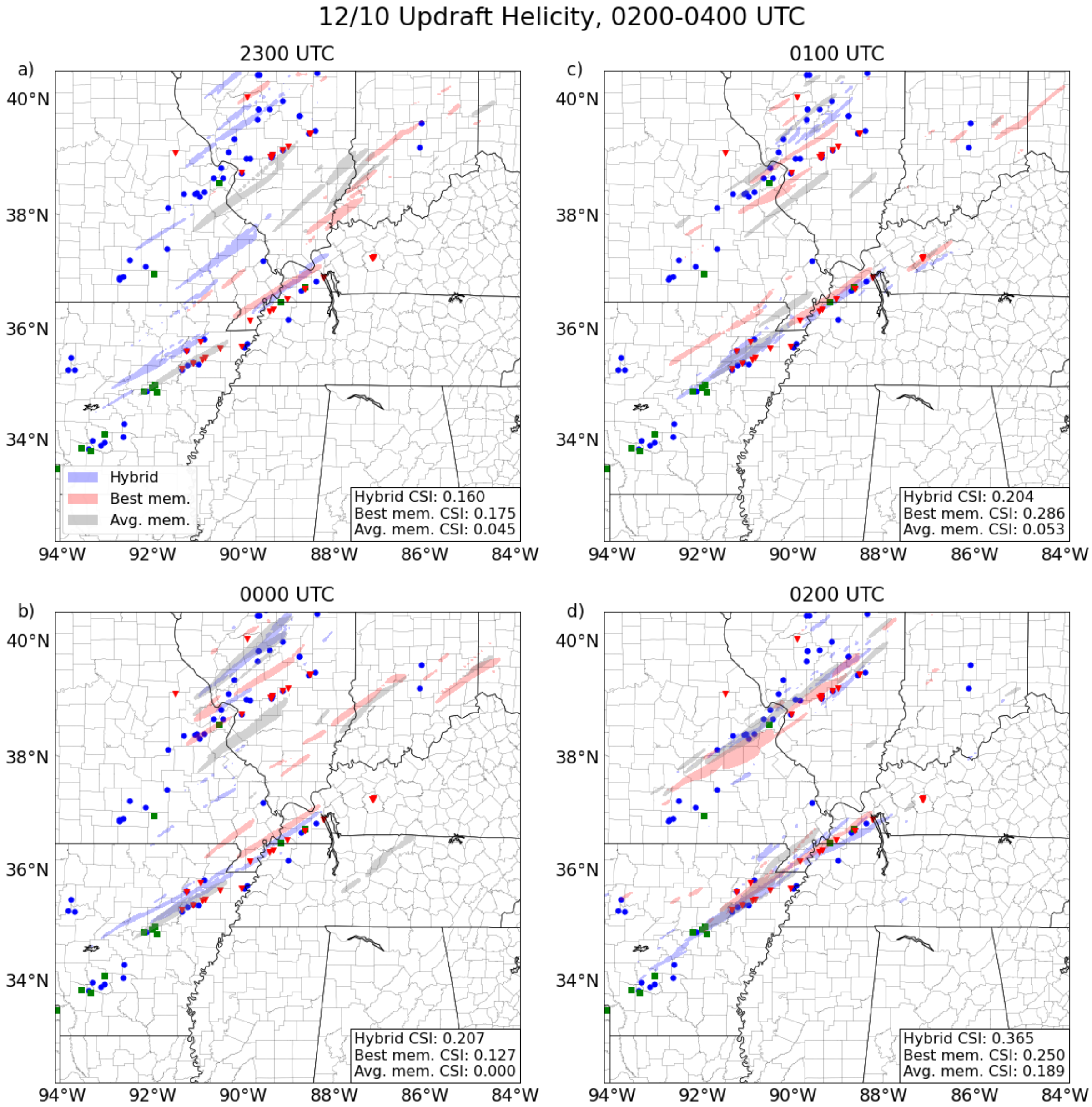


Figure 31: A spatial representation of UH from 0200-0400 UTC on 11 December over the WoFS domain of six different forecasts initialized at a) 2300 UTC, b) 0000 UTC, c) 0100 UTC, and d) 0200 on 10-11. Red (gray) contours are the UH swaths above 65 m^{2}s^{-2} of the WoFS member with the highest average (CSI average scoring WoFS member). Blue contours are the UH swaths above 130 m^{2}s^{-2} for WoF-Hybrid. The red triangles, green squares, and blue circles are tornado, hail, and wind reports from 0200-0400 UTC on 11 December, respectively. CSIs of 30-minute UH swaths over the entire domain of WoF-hybrid, the highest scoring WoFS member, and the average scoring WoFS member are displayed in the upper righthand corner. Mayfield, Kentucky is located at 36.74 °N, 88.64 °W.