UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

SENSITIVITIES OF EXPLICIT HAIL PREDICTIONS AND CONVECTIVE SCALE ENSEMBLE FORECASTING TO MICROPHYSICS PARAMETERIZATIONS AND ENSEMBLE DATA ASSIMILATION CONFIGURATIONS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

JONATHAN DAVID LABRIOLA
Norman, Oklahoma
2020

SENSITIVITIES OF EXPLICIT HAIL PREDICTIONS AND CONVECTIVE SCALE ENSEMBLE FORECASTING TO MICROPHYSICS PARAMETERIZATIONS AND ENSEMBLE DATA ASSIMILATION CONFIGURATIONS

A DISSERTATION APPROVED FOR THE
SCHOOL OF METEOROLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Ming Xue, Chair

Dr. Youngsun Jung

Dr. Nathan Snook

Dr. Cameron Homeyer

Dr. Steven Cavallo

Dr. Philip S. Harvey

Dedication

To my wife Laura and daughter Audrey, who give this work purpose.

Acknowledgments

This dissertation is a product of the support of family, friends, and coworkers. First, I owe many thanks to my research advisors Dr. Nathan Snook, Dr. Youngsun Jung, and committee chair Dr. Ming Xue. My success is attributed to these scientists. All three have worked long days (and nights) for the past seven years to lend advice, provide many rounds of edits, and offer mentorship through the many highs and lows of graduate school. I thank them for their patience as I learned the ins and outs of conducting scientific research and look forward to many more years of collaboration and friendship. I would be remiss not to also mention some of the many other CAPS scientists and collaborators who I have had the privilege to work with these past few years. Kevin Thomas provided computing support throughout my dissertation and provided me many different archived data sets. The GSI data assimilation research group (Dr. Youngsun Jung, Dr. Chengsi Liu, Dr. Rong Kong, Dr. Ming Xue, Dr. Chong Chi Tong, Lianglyu Chen, Jingyao Luo, and Jun Park) taught me how to run the GSI-based EnKF system; the hail research group (Dr. Youngsun Jung, Dr. Nathan Snook, Dr. Ming Xue, Dr. Amy McGovern, Amanda Burke, Dr. David Gagne) as well as Dr. Marcus Johnson, Dr. Ted Mansell, and Dr. Bryan Putnam provided insights to microphysics schemes, the CAPS radar emulator, and never had a dull meeting. PhD committee members: Dr. Ming Xue, Dr. Youngsun Jung, Dr. Nathan Snook, Dr. Cameron Homeyer, Dr. Steven Cavallo, and Dr. Philip S. Harvey, not only made scheduling meetings easy but took the time out of their busy schedules to attend my seminars and defenses as well as provide valuable feedback. I also thank CAPS and School of Meteorology staff members Debra Farmer, Christie Upchurch, Debbie Barnhill, Shelby Hill, and Nancy Campbell who made sure I finished all the steps needed to complete the degree and worked hard to create a sense of community in the department.

My research interests were first piqued at the University of Miami under the guidance of Dr. Sharanya Majumdar. His help opened the world of research to me. Dr. Daphane LaDue, Kiel Ortega, and Dr. Darrell Kingfield brought me to National Weather Center Research Experience for Undergraduates, provided key support during research, and showed me the power of collaborating with many of the many fantastic scientists here in Norman.

I thank my friends and family who ensure I grow both personally and academically throughout these past several years. Marcus Johnson, Rachelle Johnson, Eric Loken, Hristo Chipilsky, Tierney Harvey, Scott Harvey, Meagan Walker, and Josh King, along with many others who deserve my gratitude have made Norman home and have gone on many runs, rides, and hikes with me throughout the years. Debra and Robert Holtzman have been ready to help in any way possible including coming to Oklahoma under any circumstance. My parents Dominic and Laura Labriola have worked tirelessly and afforded me the privilege to attend college and eventually defend my dissertation. An acknowledgment does not begin to show the amount of time and effort they gave me these past three decades. I thank Audrey for allowing me to become a parent, teaching me the importance of a well-rounded life, and ensuring that even on the hardest workdays I get to be home before dinner. Finally, I thank my wife, Laura, who has stayed up late with me for the past 9 years and counting. Be it studying for an exam, Coconut Grove, letting Miley outside, finishing a manuscript, or rocking Audrey to sleep; the circumstances for why we are awake may change, but she has always emphatically been there while simultaneously dominating her own job. I couldn't think of anyone else I'd rather do this adventure with.

The research in this dissertation was supported by NSF Grant AGS-1261776 as part of the Severe Hail, Analysis, Representation, and Prediction (SHARP) project and NOAA Joint Technology Transfer Initiative (JTTI) grant NA18OAR4590385. Additional support is provided

by NOAA grant NA16OAR4590239. Computing was performed primarily on the XSEDE Stampede2 supercomputer at the University of Texas Advanced Computing Center (TACC). Supplemental computations were carried out at the Oklahoma Supercomputing Center for Education and Research (OSCER). Archived data sets used to initialize experiments in this dissertation were created under the support of NOAA CSTAR Grant NA10NWS4680001.

Table of Contents

List of Tables

List of Figures

Abstract

The explicit prediction of deep, moist convection is challenging because small model and initial condition errors rapidly grow and degrade forecast skill. Microphysics schemes employed by convection-allowing models represent a substantial source of model error because microphysical processes are poorly understood and simplifying assumptions must be made to make simulations and forecasts computationally practical. Although data assimilation systems decrease initial condition errors, analysis and forecast skill is sensitive to the experiment design. This dissertation evaluates data assimilation and ensemble forecast system performances at convection-allowing/convection-resolving resolutions, when forecast models employ different multi-moment microphysics parameterization schemes, and the data assimilation configurations are varied. We address the related issues through detailed case studies that provide insights on optimizing the configuration of convection-allowing model forecasts.

First, high-resolution hail size forecasts are made for a severe hail event on 19 May 2013 using the Advanced Regional Prediction System (ARPS). Forecasts using the National Severe Storms Laboratory (NSSL) variable density rimed ice double-moment microphysics scheme (referred to as NSSL) exhibit more skill than those using the Milbrandt and Yau double-moment (MY2) or triple-moment (MY3) schemes when verified against radar-derived hail size estimates. Although all three schemes predict severe surface hail coverage with moderate to high skill, MY2 and MY3 forecasts overpredict the maximum hail size. The NSSL scheme uses the two variable density rimed ice categories to generate large, dense hail through the wet growth of graupel. Both the MY2 and MY3 schemes predict hail to be smaller above the 0 °C isotherm because the category is primarily composed of small frozen raindrops; in the melting layer the hail quickly grows

because the rimed ice accretes excessive water. MY2 and MY3 forecasts predict the largest hail sizes to be smaller when the accretion water is eliminated beneath the 0 °C isotherm.

To improve hailstorm forecast initial conditions, CAPS Ensemble Kalman filter (EnKF) analyses are generated for the 8 May 2017 Colorado severe hail event using either the MY2 or the NSSL scheme in the forecast model. The results of the EnKF analyses are evaluated. With each microphysics scheme two experiments are conducted where reflectivity (Z) observations update either (1) only the hydrometeor mixing ratio or (2) all hydrometeor fields. Experiments that update only hydrometeor mixing ratios can create ensemble error covariances that are unreliable which increases analysis error. Despite improving initial condition estimates, experiments that update all hydrometeor fields underestimate surface hail size, which suggests additional constraint from observations is needed during data assimilation. Correlation patterns between observation prior estimates (e.g., Z) and model state variables are evaluated to determine the impact of hail growth assumptions in the MY and NSSL schemes on the forecast error covariances between microphysical and thermodynamic variables. For the MY2 scheme, Z is negatively correlated with updraft intensity because strong updrafts produce abundant, small hail aloft. The NSSL scheme predicts storm updrafts to produce fewer but larger hailstones aloft, which causes Z and updraft intensity to be positively correlated. Hail production processes also alter the background error covariances for in-cloud air temperature and hydrometeor species. This study documents strong sensitivity of ensemble data assimilation results of hailstorms to the parameterization of microphysical processes, and the need to reduce microphysics parameterization uncertainties.

To improve data assimilation configurations for potential operational implementation, EnKF data assimilation experiments based on the operational GSI system employed by the Center for Analysis and Prediction of Storms (CAPS) realtime Spring Forecast Experiments are

performed, followed by 6-hour forecasts for a mesoscale convective system (MCS) event on 28-29 May 2017. Experiments are run to evaluate the sensitivity of forecast skill to the configurations of the data assimilation system. Configurations examined include the ensemble initialization and covariance inflation as well as radar observation data thinning, covariance localization radii, observation error settings, and data assimilation frequency. Spin-up ensemble forecast surface temperatures are most skilled when the initial ensemble mean is centered upon the most recent NAM analysis, causing forecasts to predict a strong MCS. Experiments that assimilate radar observations every 5 minutes are better at the placement of high Z values near observed storms but exhibit a substantial decrease in forecast skill initially because of widespread spurious convection. Ensembles that assimilate more observations with less thinning of data or use a larger horizontal covariance localization radius for radar data overpredict the coverage of high Z values due to enhanced spurious convection. Both parameters have modestly positive impacts on forecast skill during the first forecast hour that are quickly lost due to the growth of forecast error. Forecast skill is less sensitive to the ensemble spread inflation factors and observation errors tested during this study. These results provide guidance towards optimizing the GSI EnKF system configuration, for this study the data assimilation configuration employed by the 2019 CAPS Spring Forecast Experiment produces the most skilled forecasts while remaining viable for realtime use.

Chapter 1 Introduction

1.1 Motivation and Background

In the past decade severe thunderstorms caused an estimated $163.4 billion in property damage and the frequency of severe weather events that cause more than a $1 billion in damage has steadily increased (NOAA 2019). Many of these severe weather hazards occur with relatively short warning lead-times (e.g., Brooks 2004; Stensrud et al. 2013) because the National Weather Service (NWS) issues warnings based upon the detection of a hazard (surface reports, radar detections) or an imminent threat as determined by the forecaster (i.e., warn on detection). Convection-allowing model (CAM) forecasts, which are run at relatively fine horizontal grid spacings ($\Delta x < 4$ km) to explicitly resolve deep, moist convection (Weisman et al. 1997), are increasingly used to predict the evolution of severe thunderstorms and hazards. As real-time CAM forecasts become feasible given advancements in computing infrastructure, the NWS is gradually moving toward a Warn-on-Forecast paradigm (Stensrud et al. 2009, 2013) where short-term CAM ensemble forecast guidance is incorporated into the weather decision process. A skilled CAM ensemble forecast system can potentially protect both lives and property by extending the predictive skill of severe weather hazards; however, developing a skilled prediction system represents a substantial challenge.

CAM forecast skill is highly sensitive to model and initial condition errors because small convective-scale errors quickly grow in scale to alter the evolution of the predicted storms (e.g., Lorenz 1969). A major source of model error is attributed to microphysics parameterizations (e.g., Morrison et al. 2015) that are used to explicitly predict the evolution of precipitation. This is because these parameterizations simulate poorly understood microphysical processes and must rely upon simplifying assumptions so that CAM forecasts remain computationally feasible.

Additionally, thermodynamic feedbacks from the misrepresentation of hydrometeor fields such as cold pool intensification (e.g., Snook and Xue 2008; Dawson et al. 2010) can modify the near storm environment and subsequent thunderstorm evolution (Rotunno et al. 1988).

All operational CAMs employ bulk microphysics schemes (hereafter referred to as MP schemes) to predict the evolution of bulk hydrometeor properties. These properties diagnose a fixed particle size distribution (PSD), the most common being a three-parameter gamma distribution (Ulbrich 1983) where hydrometeor mass, number concentration, and reflectivity (Z) correspond most closely to the third, zeroth, and sixth moments of the distribution. To remain computationally efficient most MP schemes predict only one or two moments of the PSD; however, in doing so the hydrometeor PSD loses flexibility because undiagnosed parameters are assigned a constant value. Double-moment schemes (e.g., Milbrandt and Yau 2005a; Morrison and Grabowski 2008; Thompson et al. 2008; Morrison and Milbrandt 2015) for example (predict hydrometeor mass and number concentration, diagnose PSD slope and intercept parameters) are prone to excessive size sorting because they are unable to modify the PSD shape parameter and narrow the PSD (e.g,. Milbrandt and Yau 2005b; Dawson et al. 2014; Johnson et al. 2016; Morrison et al. 2015). Triple-moment schemes can modify the shape parameter and mitigate excessive size sorting (Milbrandt and Yau 2005b); however, the additional state variables can dramatically increase model run times. In addition to predicting PSD moments, MP schemes can also prognose hydrometeor characteristics (e.g., rime fraction, density; Mansell et al. 2010; Milbrandt and Morrison 2013; Morrison and Milbrandt 2015) so that hydrometeor categories can better represent the diverse properties of rimed ice particles. Although these schemes are able to represent many different hydrometeor types, there are relatively few observations of microphysical relevance that can verify MP scheme accuracy.

The next-generation radar (NEXRAD; Crum et al. 1993) system is one of the only observational platforms capable of performing full-volume scans of the atmosphere at a temporal frequency (approximately 5 minutes) sufficient to capture convective storm evolution. Z is proportional to the 6th moment of the hydrometeor PSD and is most sensitive to the largest hydrometeors within a scan volume. After a recent polarimetric upgrade radars also provide observations including differential reflectivity (Z_{dr}), co-polar correlation coefficient (ρ_{hv}), and differential phase (Φ_{dp}), in addition to Z and radial velocity (V_r). With these additional observations, one can infer hydrometeor properties such as shape, size, orientation, and phase (Doviak et al. 2000; Kumjian and Ryzhkov 2008). Radar derived products such as the maximum expected size of hail (MESH, Witt et al. 1998a) and hydrometeor classification algorithm output (Heinselman and Ryzhkov 2006; Park et al. 2009) can also be used to estimate hydrometeor properties such maximum surface hail size and dominant hydrometeor type. Because they offer important microphysical information and are frequently updated, radar observations are often used for forecast verification and data assimilation.

Many CAM ensemble forecast systems (e.g., Yussouf et al. 2013; Wheatley et al. 2015; Jones et al. 2016; Snook et al. 2016; Johnson et al. 2017; Lawson et al. 2018; Labriola et al. 2019a; Stratman et al. 2020) use an ensemble Kalman filter (EnKF; Evensen 1994, 2003) to assimilate observations and initialize forecasts. This method is often used because error covariance statistics derived from the forecast ensemble allow the filter to update unobserved model state variables. An EnKF can also directly assimilate non-linear observation operators, such as Z, to update model state variables. Observation system simulation experiments (OSSEs) and real-data experiments demonstrate assimilating observed Z and V_r improves estimated thermodynamic, dynamic, and microphysical state variables (e.g., Snyder and Zhang 2003; Zhang et al. 2004; Dowell et al. 2004;

Caya et al. 2005; Tong and Xue 2005; Xue et al. 2010; Dawson et al. 2012; Jung et al. 2012; Johnson et al. 2015; Snook et al. 2015; Supinie et al. 2016; Wang and Wang 2017; Putnam et al. 2019); however, forecast initial conditions can still contain considerable errors (e.g., Xue et al. 2010) because model state variables are constrained by a limited number of relevant observations and Z can be recreated from a non-unique combination of hydrometeor moments. The configuration of the ensemble data assimilation system requires careful calibration to optimize initial condition estimates (e.g., Zhang et al. 2004; Dowell et al. 2004; Tong and Xue 2005).

1.2 Dissertation Overview

The purpose of this study is to evaluate analysis and forecast sensitivities to microphysical parameterizations and ensemble data assimilation configurations. Results are used to design a more skilled CAM ensemble forecast system for explicit hail prediction and severe weather hazards. Data assimilation is performed using the Center for Analysis Prediction of Storms (CAPS) EnKF and the gridpoint statistical interpolation (GSI) EnKF systems; forecasts are run using either the Advanced Regional Prediction System (ARPS) or the advanced research Weather Research and Forecasting (WRF-ARW) model. Experiments in this dissertation, which are all run at convection-allowing grid spacings (0.5 km $\leq \Delta x \leq$ 3 km), predict hail producing tornadic supercell thunderstorms near Oklahoma City on 19 May 2013, mountain-initiated hailstorms that impacted Denver on 8 May 2017, and a mesoscale convective system (MCS) event in the Southern United states that produced hail, wind, and tornadoes between 28 – 29 May 2017.

CAPS EnKF-initialized high-resolution (500-m horizontal grid spacing) explicit hail size forecasts of the 19 May 2013 Oklahoma City hail event are run using the ARPS in Chapter 2. Ensembles are run using multi-moment schemes that predict multiple rimed ice categories

including the Milbrandt and Yau double-moment (MY2), the Milbrandt and Yau triple-moment (MY3), and the NSSL variable density rimed ice double-moment (NSSL) schemes. Forecast maximum hail size is diagnosed from rimed ice PSD information and is evaluated against hydrometeor classification algorithm output indicated maximum hail size. To better understand biases in explicit hail size forecasts, microphysical budgets extracted from the model forecast evaluate hail growth and decay processes for each scheme. Microphysical processes in the MY2 and MY3 schemes are updated to isolate the cause of hail size forecast biases in this study.

In Chapter 3, CAPS EnKF-generated analyses are evaluated for the 8 May 2017 mountain-initiated hailstorms. Four separate experiments are run using either the MY2 or NSSL scheme where the EnKF updates only hydrometeor mixing ratio or all microphysical state variables. Analysis estimated hail size and coverage for each experiment is verified against hydrometeor classification algorithm output to evaluate EnKF rimed ice PSD estimates. Ensemble simulated Z and V_r root mean square innovations during data assimilation are also evaluated to determine which experiment configuration most closely fits radar observations. For experiments run using both MP schemes, the correlation between observation priors (Z and updraft intensity) and model state variables (hail and rain state variables, air temperature, updraft intensity) is analyzed. The correlation analyses conducted in this chapter provide insight into the sensitivity of the background error covariance to MP scheme.

The 2019 CAPS storm scale ensemble forecast GSI EnKF system is used to assimilate radar and convectional observations and initialize short-term (0-6 hour) forecasts for the 28 – 29 May 2017 MCS event in Chapter 4. Experiments evaluate probabilistic forecasts for regions of high Z (> 40 dBZ) both subjectively and objectively using reliability diagrams and the brier skill score. Additional experiments determine which aspects of the experiment configuration forecast

skill are most sensitive to. Each experiment methodically alters a different aspect of the experiment configuration including ensemble initialization method (e.g., initial condition perturbations), spread inflation during data assimilation, data thinning, covariance localization radius, observation error, and data assimilation frequency. Forecast verification of the different experiments provides insight into which configuration produces the most skilled forecasts.

Chapter 5 summarizes and discusses the performance of the CAM ensemble forecast systems across the different case studies. Future work that can further improve ensemble forecast system skill is also explored.

Chapter 2 [1]Explicit Ensemble Prediction of Hail in 19 May 2013 Oklahoma City Thunderstorms and Analysis of Hail Growth Processes with Several Multi-Moment Microphysics Schemes

2.1 Introduction

Each year, hail causes on average more than $1.4 billion in property and crop damage in the United States (Changnon et al. 2009). Densely populated regions in particular are susceptible to costly hail damage; for example, the 1995 "Mayfest" hailstorm near Fort Worth Texas caused more than $2 billion in damage and injured 109 people at an outdoor festival (Edwards and Thompson 1998). The cost of hail-related damage is expected to increase as cities expand; most hail events that have caused more than $1 billion in inflation-adjusted insured losses have occurred since 2001 and the frequency of occurrence of such events is increasing (Changnon 2009; NCEI 2017).

Extending the warning lead-time for severe hail has the potential to mitigate hail-related damage, though this is challenging, as severe hail events often develop quickly. To extend severe weather warning lead time, the National Weather Service (NWS) is shifting from issuing severe weather warnings based upon detection by, e.g. radar (warn-on-detection), to a paradigm where warnings will be issued based upon high-resolution numerical weather prediction (NWP) model guidance (warn-on-forecast; Stensrud et al. 2009, 2013) Many studies analyze the skill of convective-scale forecasts, however only a handful of studies (e.g., Milbrandt and Yau 2006; Snook et al. 2016, hereafter S16; Luo et al. 2017; Gagne et al. 2017; Labriola et al. 2017, hereafter

[1] A paper based on the material presented in this chapter has been accepted for publication as: Labriola, J., N. Snook, Y. Jung, and M. Xue, 2019: Explicit ensemble prediction of hail in 19 May 2013 Oklahoma City thunderstorms and analysis of hail growth processes with several multimoment microphysics schemes. Mon. Wea. Rev., 147, 1193–1213, https:// doi.org/10.1175/MWR-D-18-0266.1.

L17; Adams-Selin and Ziegler 2016; Luo et al. 2018) assess the skill of explicit hail prediction (hail size and coverage). Hail prediction using storm-scale NWP forecasts remains understudied due to the difficult and complex nature of hail storm prediction, hail growth processes, and challenges associated with hail forecast verification.

The skill of explicit hail prediction is dependent upon the NWP model's ability to predict the complex processes and the storm environment that supports the growth of hail; this requires an accurate representation of the hailstorm structure, dynamic, thermodynamic, and microphysical processes, as well as the surrounding environment. Hail grows where supercooled liquid accretes to the surface of rimed ice; however, growth is limited if a large number of hail embryos depletes the supercooled liquid (e.g., Heymsfield 1983). Hail particle growth is strongly correlated with storm updraft strength (e.g., Rasmussen and Heymsfield 1987). Strong updrafts suspend large rimed ice particles in regions of supercooled liquid; however, if an updraft is too strong the particles are more rapidly advected outside of the optimal growth zone. Strong vertical wind shear also contributes to hail growth (e.g., Nelson 1983; Foote 1984; Dennis and Kumjian 2017). Increased deep layer shear extends the storm updraft horizontally and increases the residence time of hail within the growth zone. Due to the complex nature of riming, a large number of microphysical parameterizations that have different treatments for rimed ice were developed with the goal of improving the representation of rimed ice.

Microphysical processes in convective-scale NWP models are mostly parameterized using bulk microphysics (MP) schemes (e.g., Lin et al. 1983; Milbrandt and Yau 2005a). Bulk MP schemes predict the bulk characteristics for hydrometeor species x such as mixing ratio (q_x), total number concentration (N_{tx}), or reflectivity (Z_x) to define a commonly used three-parameter gamma particle size distribution (PSD):

$$N(D) = N_{0x} D^{\alpha_x} e^{-\lambda_x D} \quad (1)$$

where D is the hydrometeor diameter, λ_x is the slope parameter, α_x is the shape parameter, and N_{0x} is the intercept parameter (Ulbrich 1983). q_x is converted to mass concentration (m_x) when multiplied by air density. A few MP schemes (Mansell et al. 2010; Morrison and Milbrandt 2015; Morrison et al. 2015; Milbrandt and Morrison 2016) also have the capacity to predict hydrometeor volume (v_x) and thus explicitly derive density (ρ_x); this is primarily done for rimed ice hydrometeor categories, which undergo large fluctuations in density during growth and melting. Variable density MP schemes (e.g., Mansell et al. 2010) have the ability to update particle fall speeds and rimed ice production.

To save computational time, one or more PSD parameters are often assumed to be constant, though such assumptions can limit the model's ability to realistically represent microphysical processes. For example, most double-moment MP schemes predict the 3rd (q_x) and 0th (N_{tx}) moments of a PSD to diagnose λ_x and N_{0x}, but assume α_x to be constant. Double-moment schemes are able to simulate qualitatively polarimetric signatures (Jung et al. 2012; Johnson et al. 2016; Putnam et al. 2017b), but the schemes often suffer from excessive size sorting (e.g,. Milbrandt and Yau 2005b; Dawson et al. 2014; Johnson et al. 2016; Morrison et al. 2015). Triple-moment MP schemes, which additionally predict the 6th moment of the PSD (Z_x) can diagnose α_x. The shape parameter narrows the hydrometeor size spectra and limits size sorting by having the weighted fall speeds of a hydrometeor type converge towards a singular value (Milbrandt and Yau 2005b; Dawson et al. 2014).

Convective-scale forecasts have been shown to predict spatial extent and size of hail at the surface with some skill. Multiple studies (e.g., S16; L17; Luo et al. 2017; Luo et al. 2018) have used predicted hail quantities (mass, number concentration) to identify regions where hail reaches

the surface. Simulated radar products such as the maximum expected size of hail (MESH; Witt et al. 1998a) have also been used to derive surface hail size from model output (S16; L17; Luo et al. 2017; Luo et al. 2018). Another method, the Thompson hail method (Thompson et al. 2018), uses model diagnosed hail PSDs to approximate the largest "observable" hailstone. Variants of this method have been used to diagnose the maximum surface hail size and distinguish regions of hail growth (Milbrandt and Yau 2006; S16; L17; Luo et al. 2018). The skill of hail forecasts derived directly from model output is strongly dependent upon choice of MP scheme; to account for model biases machine learning algorithms can be applied to model output (e.g., Gagne et al. 2015, 2017; McGovern et al. 2017; Herman and Schumacher 2018a,b; Burke et al. 2020).

Convective-scale explicit hail forecasts are subject to significant error associated with the NWP model (e.g., initial conditions, microphysics). Many recent convective-scale studies use cycled ensemble Kalman filter (EnKF; Evensen 1994, 2003) data assimilation (DA) to generate initial conditions (e.g., Dowell et al. 2004; Snook et al. 2011, 2012, 2015; Dawson et al. 2012; Jung et al. 2012; Putnam et al. 2014, 2017, Yussouf et al. 2013, 2016; Wheatley et al. 2014; Schwartz et al. 2015; Skinner et al. 2018). This technique is preferred because error covariances derived from the forecast ensemble allow the DA system to update unobserved variables such as temperature, pressure, and microphysical variables from available observations (e.g., Tong and Xue 2005; Tong and Xue 2008a). An EnKF DA system can thus use radar observations, which indirectly observe hydrometeor information, to improve the microphysical state variables of hail producing storms (S16; L17).

Hail forecast verification is a substantial challenge. Hail events are underreported in rural areas, especially away from major highways, as well as in cases when a more extreme severe weather event (e.g., one or more tornadoes) occurs in addition to hail (Doswell et al. 2005; Witt et

al. 1998b; Allen and Tippett 2015). Additionally, hail sizes that correspond to familiar circular or spherical objects (e.g., dimes, softballs) are over-reported (Sammler 1993), as such objects are commonly used as size references in reports from the public. New observational databases attempt to mitigate observed hail size biases. The meteorological phenomena identification near the ground (mPING; Elmore et al. 2014) phone application requires users to report hail in size increments of 0.25 inches, and the severe hail analysis and verification experiment (SHAVE; Ortega et al. 2009) directly interviews the public to create a hail observation database with high spatial resolution. Nevertheless, surface-based reports remain inadequate for objective hail forecast verification.

The next-generation radar (NEXRAD; Crum et al. 1993) system is the only observational platform that performs full-volume scans of the atmosphere at a temporal frequency (approximately 5 minutes) that captures the rapid evolution of convective storms. Radar derived hail products, such as MESH, serve as a proxy for hail size and can be used for forecast verification (e.g., S16). Although MESH varies in skill throughout the United States, it is not subject to population bias and thus superior to ground based reports (Cintineo et al. 2012).

Polarimetric radars provide observations such as differential reflectivity (Z_{dr}), co-polar correlation coefficient (ρ_{hv}), and differential phase (Φ_{dp}), in addition to Z and radial velocity (V_r). With these additional observations, one is able to infer hydrometeor properties such as shape, size, orientation, and phase (Kumjian and Ryzhkov 2008). Hydrometeor classification algorithms (HCAs; Park et al. 2009) apply a fuzzy logic algorithm to both single-and dual-polarization radar data to classify the dominant hydrometeor type. This study implements the Putnam et al. (2017b) HCA, along with hail size discrimination algorithm membership functions from Ortega et al. (2016) to classify the dominant hail size into one of three bins: non-severe (5 mm), severe (25

mm), and significant severe (50 mm). HCAs often classify hail size with more skill than MESH (Ortega et al. 2016), although biased observations degrade the skill of the classifications.

Previous studies (i.e. S16; L17; Luo et al. 2017, 2018) evaluated hail forecasts produced using single-moment, double-moment, and triple-moment MP schemes; however, no studies have evaluated hail forecasts produced using a variable density MP scheme, which has the ability to represent a spectrum of rimed ice particle characteristics. In this study double-moment, triple-moment, and variable-density double-moment MP schemes are used to produce explicit ensemble hail forecasts for a severe hailstorm event that occurred in the Oklahoma City Metropolitan area. Surface hail size forecasts are subjectively and objectively verified against HCA output (Ortega et al. 2016) using the Brier skill score (Brier 1950), reliability diagrams, and relative operating characteristic curves. In addition to forecast evaluations, a microphysical budget analysis is conducted. This study analyzes hail growth and decay processes predicted by each MP scheme to identify sources of forecast error, and determine the impact of microphysical processes on the prediction of hail.

The rest of the chapter is organized as follows. In section 2.2 we provide a brief overview of the 19 May 2013 Oklahoma City severe hail event and describe the experimental design, including the DA and forecast model settings. Hail forecast verification statistics and an in-depth analysis of hail growth and decay processes are presented and discussed in section 2.3. A summary and some further discussions on the results are given in section 2.4.

2.2 Case, experiment configuration, and verification procedures

2.2.1 Case overview

On 19 May 2013 an upper-level low pressure system developed in the Northern Plains of the United States as a negatively-tilted trough moved over the Southern Plains. An 1800 UTC (1300 LST) sounding launched from the Norman Weather Forecast Office (WFO) indicated the environment was both highly unstable (CAPE ~ 4878 j kg^{-1}) and strongly sheared (0 – 6 km shear ~ 49 kts), ideal for the development of supercell thunderstorms. Thunderstorms initiated along a dryline boundary west of Oklahoma City starting at approximately 1945 UTC. Three supercell thunderstorms (Fig. 2.1a) produced large regions of severe hail, and localized regions of significant-severe hail, in the southwestern Oklahoma City Metropolitan area (Fig. 2.1b, c). Aside from producing large hail throughout the region, these storms also produced two long-track tornadoes that produced EF-3 damage in Carney, OK and EF-4 damage in Shawnee, OK. For further discussion on this event we refer the reader to Carlin et al. (2017) and Wienhoff et al. (2018).

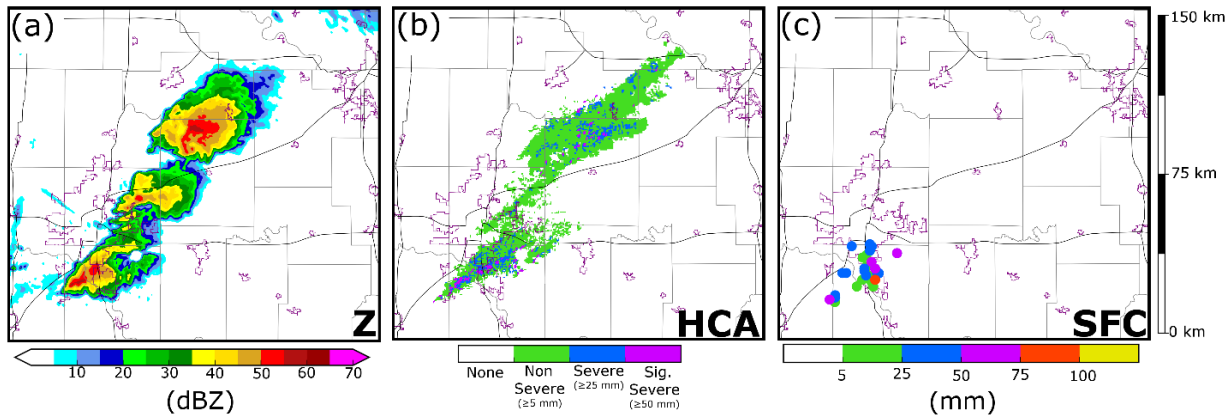

Fig. 2.1. (a) KTLX observed *Z* at the lowest radar tilt (0.5°), valid for 2220 UTC. (b) A swath of the largest observed hailsize between 2155 – 2320 UTC. Hail size is determined by applying a HCA to the lowest radar tilt of KTLX and KINX (0.5°) and merging the subsequent output. (c) Surface based hail size reports (SPC and mPING) between 2155 -2320 UTC. On the background maps thin black lines are highways and purple lines are urban boundaries. The central and right panel color scales correspond to the same range of hail sizes.

2.2.2 Prediction model configurations

All experiments use the Advanced Regional Prediction System (ARPS; Xue et al. 2000, 2001) as the numerical weather prediction model. The model domain consists of 573 × 515 × 53 grid points with a 500-m horizontal grid spacing. The model grid is stretched in the vertical, with a minimum vertical grid spacing of 50-m at the surface and an average vertical grid spacing of 425-m. The ARPS model settings follow those of S16 and L17. Model physics (Xue et al. 2000, 2001) include NASA Goddard Flight Center short- and longwave radiation (Chou 1990, 1992; Chou and Suarez 1994), surface fluxes calculated from surface drag coefficients, surface temperature and volumetric water, a two-layer soil model, and a 1.5-order turbulent kinetic energy-based sub grid-scale turbulence parameterization (Deardorff 1980; Klemp and Wilhelmson 1978; Moeng 1984).

Hail forecasts are produced using three different MP schemes: the Milbrandt and Yau (2005a) double-moment MP scheme (MY2), the Milbrandt and Yau (2005a) triple-moment MP scheme (MY3), and the National Severe Storms Laboratory (NSSL) double-moment variable-density rimed ice MP scheme (Mansell et al. 2010). Forecast maximum surface hail size (D_{max}) is calculated using a variant of the Thompson hail algorithm (Gagne et al. 2019). D_{max} is calculated at the first model level above the surface (~25 m AGL) and is defined as the largest diameter for which the hail PSD has at least 1 stone per 10,000 m^4 (S16).

The MY2 scheme is selected for this experiment because it has been used in multiple recent hail prediction studies and serves as a baseline for esplicit hail prediction (e.g., S16, L17, Luo et al. 2017, 2018). The MY2 and MY3 schemes use the same prognostic equations for hydrometeor

mass mixing ratio and number concentration, however MY3 also predicts hydrometeor Z. Z is used to diagnose the PSD shape parameter (Table 2.1); this parameter narrows the size distribution and limits further size sorting by causing weighted fall speeds to converge towards a singular value (Dawson et al. 2014). The NSSL scheme (Table 2.1) includes mass mixing ratio and number concentration, as well as prognostic equations for both hail and graupel volume (and thus density). Variable density is used to update particle fall speeds and conversion rates between hydrometeor species (e.g., dense wet-growth graupel is converted to hail; Mansell et al. 2010). The minimum number concentration threshold is set to 10^{-8} m^{-3} for the MY2 and MY3 schemes; the default setting (10^{-3} m^{-3}) led to excessive removal of hail, including some large hail near the surface. Unlike the MY2 and MY3 schemes, which threshold via number concentration and mass, the NSSL scheme only thresholds via mixing ratio, and thus does not remove as many hail gridpoints near the surface.

	MY2	**MY3**	**NSSL**
Graupel	q_g, N_{tg}	q_g, N_{tg}, Z_g	q_g, N_{tg}, V_g
Hail	q_h, N_{th}	q_h, N_{th}, Z_h	qh, N_{th}, V_h

Table 2.1. The prognostic rimed ice variables included in each MP scheme

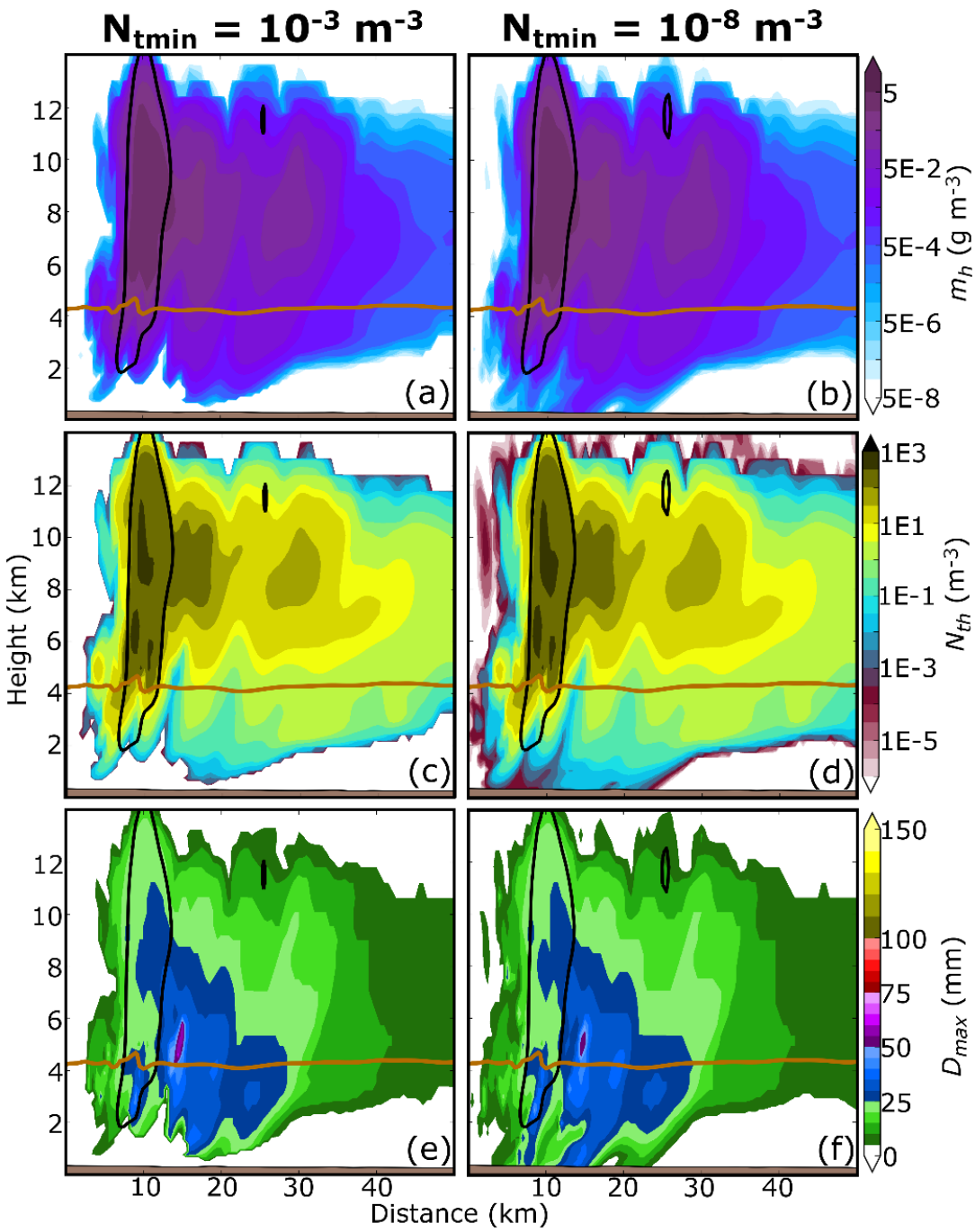

Fig. 2.2. Vertical cross-sections taken through the hail core of a supercell thunderstorm after a 60 minute forecast. Forecasts are run using the MY2 scheme where the minimum number concentration is set to 10^{-3} m^{-3} (a, c, e) and 10^{-8} m^{-3} (b, d, f). Cross-sections are of m_h (a, b), N_{th} (c, d), and D_{max} (e, f). In each plot a brown horizontal line represents the 0 °C isotherm, black contours indicate regions where updraft velocity exceeds 5 m s^{-1}.

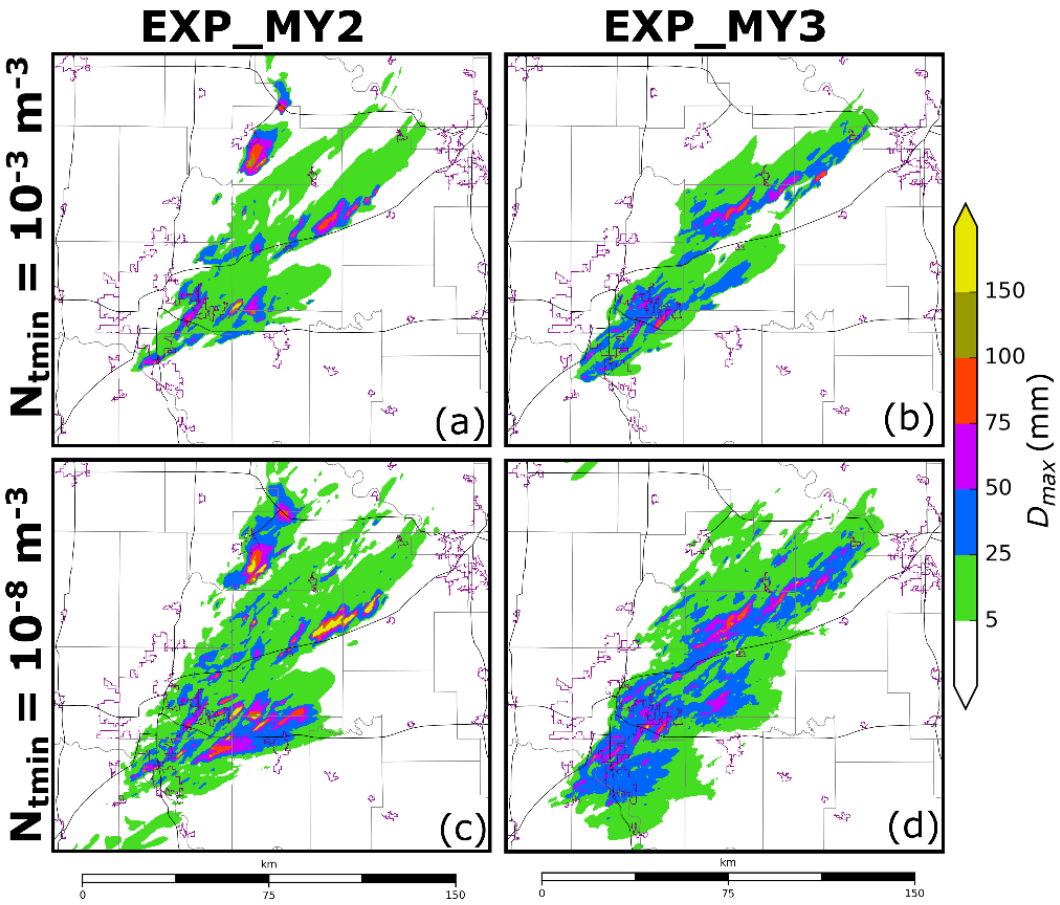

Fig. 2.3. Swaths of forecast maximum surface D_{max} calculated at every model time step between 2155 - 2320 UTC. Forecasts are run using the MY2 (a, c) and MY3 (b, d) schemes, where the minimum number concentration threshold is set to 10^{-3} m^{-3} (a, b) or 10^{-8} m^{-3} (c, d). Background maps are the same as in Fig. 2.1

Vertical cross-sections taken through the hail core of a thunderstorm, after a 60-minute forecast, are used to demonstrate the impact of reducing the minimum number threshold. The default MY2 scheme (Fig. 2.2a ,c, e) removes near-surface hail that is relatively sparse in both mass ($< 5 \times 10^{-3}$ g m^{-3}) and number ($< 5 \times 10^{-5}$ m^{-3}); however, the updated scheme (Fig. 2.2b, d, f) predicts more hail to reach the surface that is classified by the D_{max} algorithm (Fig. 2.2f). Decreasing the minimum number threshold in the MY2 and MY3 schemes substantially increases surface hail size and coverage (Fig. 2.3c, d) compared to the default scheme (Fig. 2.3a, b) because the lower threshold limits the removal of near-surface hail. Altering the minimum number

threshold affects mostly near surface hail size forecasts (Fig. 2.3), while model state variables (e.g., w, q_h, N_{th}) remain relatively similar between the forecasts (Fig. 2.2a-d) in the rest of the storm. Although additional precipitation in the simulation can alter the storm environment, reducing the minimum number concentration threshold does not increase the number of spurious storms for this study.

2.2.3 Radar observation operators and EnKF DA configurations

The EnKF DA requires accurate and efficient radar reflectivity observational operators consistent with the MP schemes used. The Rayleigh scattering approximation (Jung et al. 2008a,b) is typically used to simulate the Z of ice hydrometeors in prior EnKF studies (Dawson et al. 2010; Snook et al. 2011, 2015, 2016, Jung et al. 2012; Putnam et al. 2014; L17). While this observation operator is computationally efficient, it frequently overestimates the Z of large ice particles that cause Mie scattering (e.g., hail). The T-matrix radar simulator/observation operator (Jung et al. 2010a) accounts for Mie scattering; it uses the T-matrix calculations (Vivekanandan et al. 1991; Bringi and Chandrasekar 2001) to derive the scattering amplitude of hydrometeors for different size bins and Z is then calculated by integrating the scattering amplitude over a given PSD. Recent updates have optimized the computational efficiency of the T-matrix method; for example, Putnam et al. (2018) calculates scattering amplitude for select PSD parameters and uses lookup tables and interpolation to derive Z. This study uses the Putnam et al. (2018) T-matrix-based operator, which accounts for Mie scattering at computational speeds comparable to the Rayleigh scattering approximation.

Adjustments are made to the radar forward observation operator to accommodate the different MP schemes. To reduce the size of scattering amplitude look-up tables, the Rayleigh

scattering approximation is assumed for the NSSL snow category, since snow particles are small enough that we can assume Rayleigh scattering with little to no error. To simulate the melting layer, a variant of the Jung et al. (2008a) melting model is used to create ice-water mixtures for both graupel and hail. The ice-water hybrid category (which is generated offline) is determined via a ratio between ice and rainwater mixing ratios. For example, the hail-water mixing ratio (f_{wh}) is:

$$f_{wh} = \frac{q_r}{q_r + q_h}, \tag{2}$$

while computationally efficient, this calculation does not account for shedding and thus allows too much liquid to accumulate on the surface of an ice particle. In regions where q_r is large, this can cause the diameter of melting ice to overinflate. To maintain computational efficiency but to produce more realistic Z we modify the ice-water mixture number concentration so that the mean mass diameter (D_{mx}) of the melting ice species is preserved during melting.

A 40-member ensemble is used for both the EnKF DA and forecast periods during this experiment, and one ensemble experiment is run using each of the microphysical schemes listed above (MY2, MY3, NSSL). Hereafter, the ensembles run using the MY2, MY3, and NSSL schemes are referred to as EXP_MY2, EXP_MY3, and EXP_NSSL, respectively. We employ a 40-member ensemble because the initial and boundary conditions are provided by the Center for Analysis and Prediction of Storms (CAPS) EnKF-initialized storm scale ensemble forecast (SSEF; Jung et al. 2014). The EnKF-initialized SSEF is a 40-member ensemble of convection allowing (4-km horizontal grid spacing) forecasts that were run using the Advanced Research Weather Research and Forecasting model (ARW; Skamarock et al. 2008) for the NOAA Hazardous Weather Testbed Spring Forecasting Experiment. The SSEF model domain spans the contiguous

United States, forecasts are initialized on 19 May 2013 at 1800 UTC and are run until 20 May 2013 at 0000 UTC. Additional information about the SSEF is provided in CAPS (2013).

At 1900 UTC, our 500-m domain ensemble is initialized by interpolating from the 1-hour forecast leadtime SSEF. To introduce additional storm-scale perturbations, a 2-dimensional recursive filter is used to create smoothed Gaussian perturbations with horizontal and vertical correlation scales of 6 km and 3 km, respectively, which are added to the interpolated 1900 UTC SSEF ensemble. The storm scale perturbations have standard deviations of 2 m s^{-1}, 2 K, and 0.5 g kg^{-1} for horizontal wind components (u, v), potential temperature (θ), and water vapor mixing ratio (q_v), respectively. Sixty-minute spin-up ensemble forecasts are launched from the perturbed initial conditions and run until 2000 UTC when EnKF DA is first performed. The CAPS EnKF system (Xue et al. 2006; Tong and Xue 2008b), which is based upon the Whitaker and Hamill (2002) ensemble square-root filter (EnSRF) algorithm, is used to assimilate observations every 10 minutes from 2000 to 2150 UTC, the time of final EnKF analysis. This configuration resulted in the improved suppression of spurious storms.

The assimilated conventional observations are surface (Oklahoma Mesonet, ASOS, and AWOS) and profiler observations. Conventional observations are not used to update vertical velocity (w) because in preliminary tests, unreliable covariances led to the development of spurious updrafts. A list of the model state variables updated by each observation type is provided in Table 2.2. Observation error settings largely follow S16: surface observation errors are assumed to be 1.5 m s^{-1} for u and v, 2.0 K for θ, and 2.0 K for dewpoint temperature (T_d). The surface observation covariance localization radius is set to 300 km in the horizontal and 6 km in the vertical. Wind profiler observations are used to update background u and v winds and assume an error of 2.5 m

s^{-1} for both variables and a covariance localization radius of 80 km in the horizontal and 6 km in the vertical.

Observation	Updated State Variables
Surface	u, v, θ, q_v
Profiler	u, v
Sounding	u, v, θ, q_v
KTLX Z	u, v, w, θ, q_v, q_x
KTLX V_r	u, v, w, q_v

Table 2.2. A list of the assimilated observations along with the corresponding model state variables that are updated by each observation.

Radar observations can infer important microphysical information and are assimilated. Data from the nearest Oklahoma City NEXRAD radar (KTLX) are interpolated horizontally to the model grid column locations but are preserved in the vertical at the height of the radar beam (Xue et al. 2006); Z and V_r observations are also thinned in the horizontal to one observation every 1 km in regions of precipitation (observed $Z > 5$ dBZ) and to one observation every 2 km in regions of clear air (observed $Z < 5$ dBZ). Radar data are thinned horizontally to control the ensemble spread reduction due to assimilation of very dense observations. KTLX observation errors are assumed to be 4.0 m s^{-1} for V_r and 6.0 dBZ for Z (Snook et al. 2013), and the covariance localization function (Gaspari and Cohn 1999) cutoff radius is set to 3 km in both the horizontal and vertical.

As multi-category microphysical variables are under-constrained by observed Z, large uncertainties may exist within analyzed variables. Uncertainties may increase for multi-moment schemes because of the larger number of predicted variables (Xue et al. 2010); for example, the MY2, MY3, and NSSL schemes predict 12, 17, and 15 state variables, respectively. Multi-moment

schemes better represent the non-monotonic relationship between q_x and Z, which improves the representation of simulated radar variables (e.g., Putnam et al. 2014), and motivates the use of multi-moment schemes in an EnKF during radar DA (Jung et al. 2012). To limit the influence of unreliable covariances, we update only hydrometeor mixing ratio during DA (Table 2.2). Updating only hydrometeor mixing ratios may lead to a slower filter convergence rate but the system produces fewer spurious storms than updating more variables.

The relaxation-to-prior-spread method (Whitaker and Hamill 2012) is used to inflate the posterior ensemble spread to 95% of the prior ensemble spread for all updated variables. Although N_{tx}, v_x, and Z_x are not updated (and thus never inflated), the variables are modified to avoid unrealistic PSD behavior during inflation. During inflation, N_{tx} is updated to preserve hydrometeor mean mass diameter, v_x is updated to preserve particle density, and Z_x is updated to preserve PSD shape parameter. Subjective comparisons between MY3 predicted Z after inflation and radar observed Z (not shown) indicate that the diagnostic update can produce a posterior Z that compares well with observations.

2.2.4 Verification procedure

Hail size classifications obtained from the output of the Ortega et al. (2016)HCA applied to the two nearest NEXRAD radars (KTLX and KINX) are used to verify hail size forecasts. Observations from KTLX and KINX are interpolated to the 500-m model grid and a 9-point smoother is applied to the Z_{dr} and ρ_{hv} fields to reduce noise. The HCA is performed for radar data on the lowest elevation angle within 120 km of a given radar — 120 km is the maximum distance Ortega et al. (2016) used to evaluate hail size. Although the HCA is applied to the lowest radar tilt, the membership functions derived by Ortega et al. (2016) are tuned to detect surface hail size.

Next, the data are merged (the largest hail size is selected where radar volumes overlap). Due to small-scale noise, a final smoothing filter is applied to the merged HCA output. The smoothing filter decreases (increases) hail size detections when the four closest grid points are smaller (larger); when a detection is updated, it is set to match the four surrounding grid points. Merged HCA output (Fig. 2.1b) is used for both subjective and objective forecast verification in this study.

Hail forecast verification is challenging in part because large surface hail is a highly localized phenomenon (most hail cores are at most several kilometers wide). When using a traditional neighborhood ensemble probability (NEP; Schwartz et al. 2010) approach, small position errors between ensemble members can lead to mostly non-zeros probabilities that are difficult to verify. Neighborhood maximum ensemble probability (NMEP; Ben Bouallègue and Theis 2014; Schwartz and Sobash 2017) considers the probability of an event occurrence within distance *i*; this expands both the number of probabilistic and observed events and increases the sharpness of the forecast. For objective verification the NMEP method defined in Schwartz and Sobash (2017) is used to verify the probability of an event occurrence within 10 km of a grid point with a smoothing Gaussian filter of 10 km. This distance was selected because it accounts for small forecast storm displacement errors, but maintains the high spatial resolution necessary for warn-on-forecast applications.

Probabilistic hail size forecasts are evaluated using the Brier skill score (Brier 1950), reliability diagrams, and relative operating characteristic (ROC; Mason 1982) curves. The Brier skill score is defined by three separate components: forecast reliability, resolution, and uncertainty. Forecast reliability compares predicted probability of occurrence to observed frequency, which is used to infer model bias and is most commonly displayed in terms of a reliability diagram. In an unbiased system observed frequency and predicted probability are similar; if there is a mismatch,

the model suffers an overprediction (underprediction) bias when the predicted probability of occurrence is larger (smaller) than the observed frequency. ROC curves determine the model's ability to discern between events and non-events. ROC curves are generated by plotting probability of detection (POD) against probability of false detection (POFD) for increasing probability thresholds. Forecast skill is usually determined by the area under the ROC curve (AUC), if the AUC exceeds 0.7 then a general consensus suggests the forecast is skilled at predicting the event (Buizza et al. 1999). Higher AUC values suggests the forecast exhibits additional skill.

2.3 Results

2.3.1 Forecast evaluation and verification

90-minute-long hail size forecasts starting from the final EnKF analyses at 2150 UTC are verified between 2155 - 2320 UTC, a period during which multiple supercell thunderstorms produce severe and significant severe hail over the Oklahoma City metropolitan area (Fig. 2.1b, c). Accumulated swaths of the forecast maximum D_{max} (calculated at every model time step) are verified against accumulated swaths of observed maximum hail size as indicated by HCA output (calculated every radar volume scan). We note that radars observe the atmosphere far less frequently (~ 5 minutes) than a typical storm-scale model time step (~ 1 second), however the higher temporal resolution of the model output better captures the extent and evolution of hail in the model storms, and thus choose to perform verification with the highest temporal frequency data available. Hail size forecasts are verified for two separate diameter thresholds including severe hail (diameter > 25 mm) and significant severe hail (diameter > 50 mm).

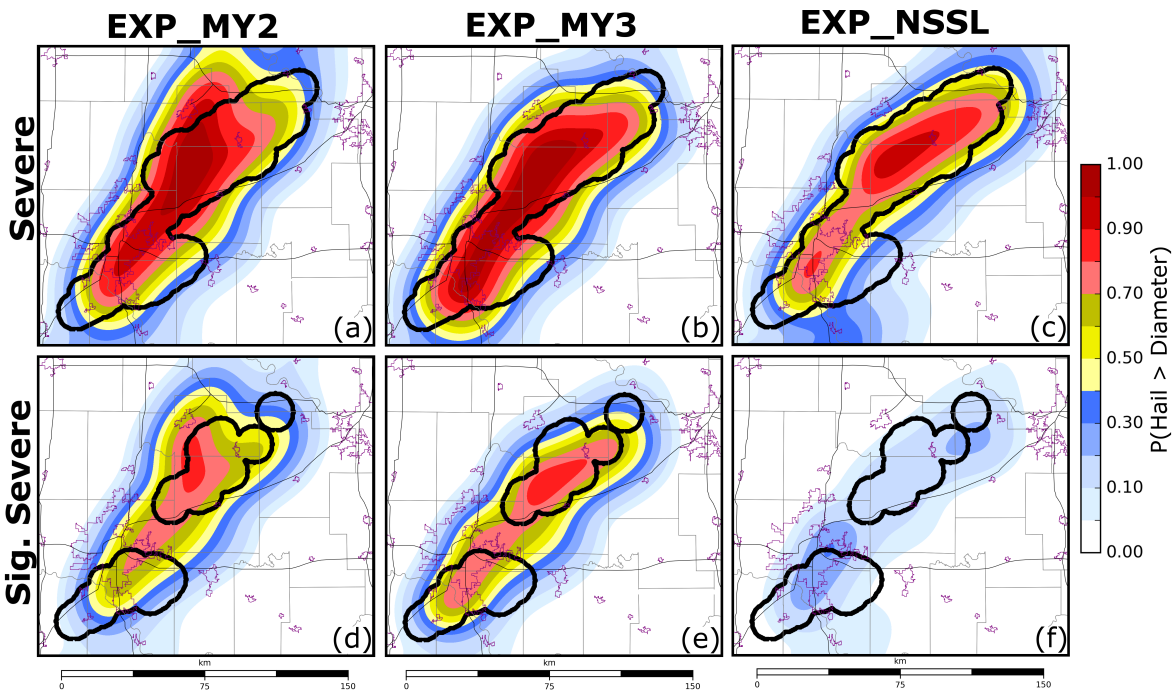

Fig. 2.4. The NMEP of (a - c) severe hail and (d - f) significant severe hail predicted using the (a, d) MY2, (b, e) MY3, and (c, f) NSSL MP schemes between 2155 – 2320 UTC. Thick black contours represent the locations of observed severe (>25 mm) or significant severe (> 50 mm) hail according to HCA output. Background maps are the same as in Fig. 2.1.

For this study, we first subjectively evaluate the NMEP of severe and significant severe hail and note forecast biases. The predicted probability of severe hail (P[severe hail]) (Fig. 2.4a - c) is generally high (> 0.7) in regions where severe hail is observed. Both EXP_MY3 (Fig. 2.4b) and EXP_NSSL (Fig. 2.4c) predict the northernmost hail swath to be oriented towards the northeast, suggesting these ensembles predict the rightward deviation in storm motion commonly observed in supercell thunderstorms (Bunkers et al. 2000). Unlike the above two experiments, EXP_MY2 (Fig. 2.4a) predicts a bifurcation in maximum P(severe hail) values for the northernmost storm. This bifurcation is due to the production of spurious storms in EXP_MY2 forecasts. EXP_MY2 forecasts produce many spurious storms, this increases the overall coverage of severe hail and the total area where P(severe hail) > 0.4 (Fig. 2.4a). Both EXP_MY2 (Fig. 2.4d)

and EXP_MY3 (Fig. 2.4e) predict P(significant severe hail) to exceed 0.5 where HCA output (Fig. 2.1b) indicates the occurrence of significant severe hail.

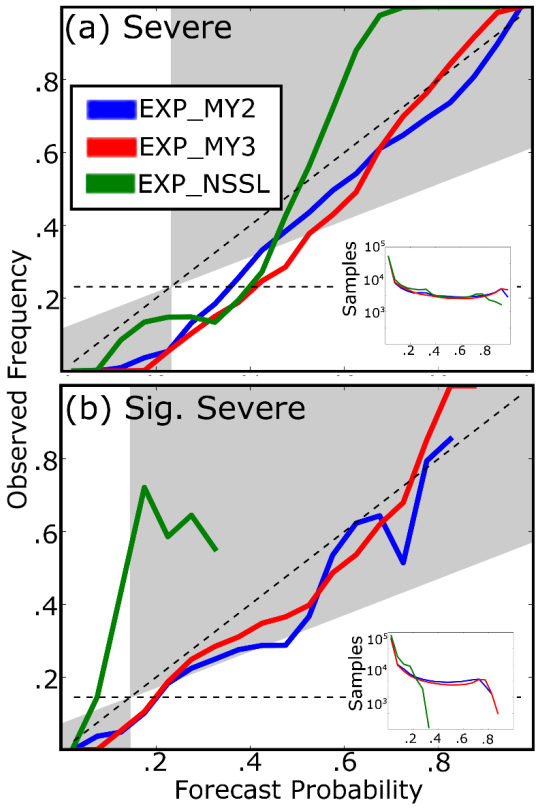

Fig. 2.5. Reliability diagrams for the probabilistic (a) severe hail forecasts and (b) significant severe hail forecasts shown in Fig. 2.4. In the diagrams blue lines correspond to EXP_MY2, red lines to EXP_MY3, and green lines to EXP_NSSL. Forecasts exhibit skill when the forecast reliability falls within the gray shading. The number of samples per forecast probability bin is provided for both (a) severe hail and (b) significant severe hail.

Reliability diagrams indicate that ensembles predict the spatial coverage of severe hail (Fig. 2.5a) with relatively little bias. Although Brier skill scores (Table 2.3) vary slightly (0.532 – 0.633), all three ensembles produce skillful severe hail forecasts. For significant severe hail forecasts, the Brier skill score (Table 2.3) decreases for all three ensembles (0.132 – 0.463), indicating the ensembles predict more extreme events with less skill. EXP_NSSL predicts significant severe hail with the least skill (Table 2.3), in part because the ensemble produces under

confident significant severe hail size forecasts (Fig. 2.5b) and rarely predicts P(significant severe) > 0.2 (Fig. 2.4f). Despite lower Brier skill scores (Table 2.3), EXP_MY2 and EXP_MY3 produce reliable significant severe hail forecasts (Fig. 2.5b) and predict significant severe hail with higher confidence than EXP_NSSL.

	Severe	Significant Severe
EXP_MY2	0.532	0.352
EXP_MY3	0.633	0.463
EXP_NSSL	0.629	0.132

Table 2.3. The Brier skill score for EXP_MY2, EXP_MY3, and EXP_NSSL. Skill scores are calculated for severe hail and significant severe hail forecasts between 2155 – 2320 UTC. The corresponding NMEP used to derive the Brier skill score is provided in Fig. 2.4.

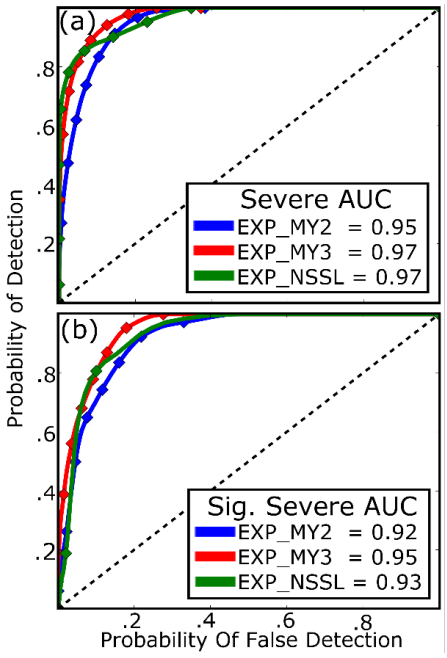

Fig. 2.6. ROC curves for the proabilistic (a) severe and (b) significant severe hail forecasts shown in Fig. 2.4. In the diagrams blue lines correspond to EXP_MY2, red lines to EXP_MY3, and green lines to EXP_NSSL. Forecast AUC scores are provided, scores that are greater than 0.7 predict severe or significant severe hail with skill.

Contrary to the other verification metrics, ROC curves (Fig. 2.6) indicate the ensembles predict both severe and significant severe hail with a high level of skill (AUC > 0.92). Large variations in P(significant severe) between the ensembles does not appear to have a large impact on the AUC score. It is noted that these verifications are performed for only a single event in this study, additional case studies conducted show similar results.

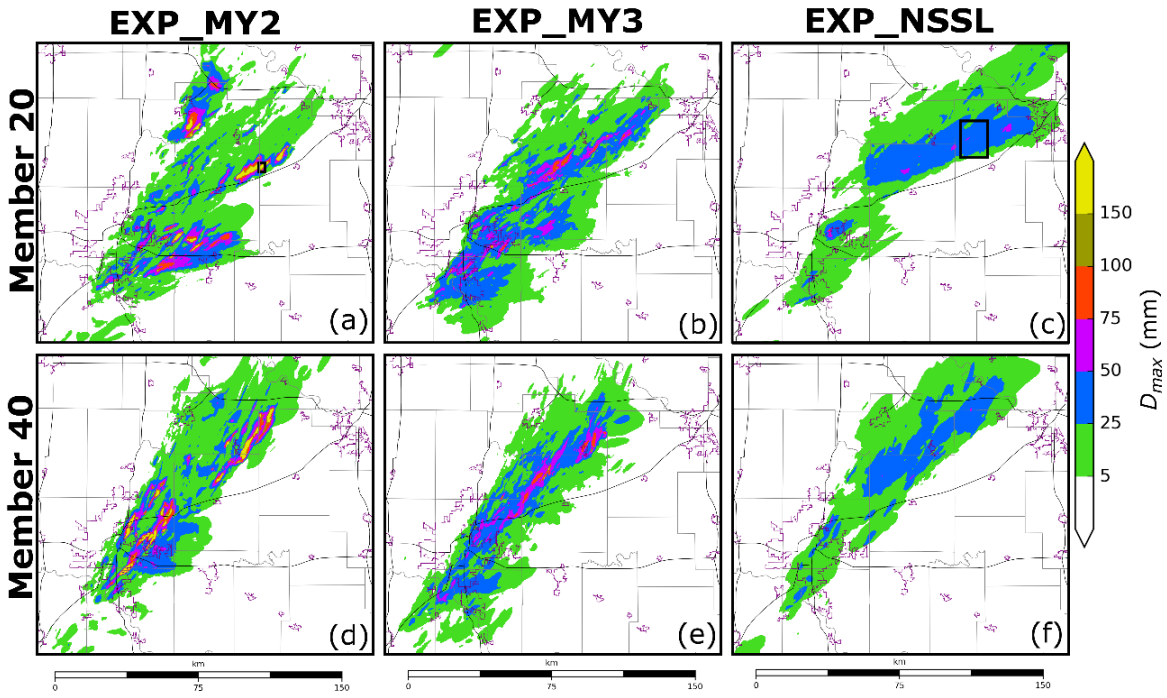

Fig. 2.7. Swaths of forecast maximum surface D_{max} calculated at every model time step between 2155 - 2320 UTC. Swaths are created for ensemble members (a - c) 20 and (d - f) 40 which are run using the (a, d) MY2, (b, e) MY3, and (c, f) NSSL MP schemes. These members are representative of the other ensemble member forecasts. Black polygons in (a) and (c) highlight the locations where hail processes are analyzed later in the study. Background maps are the same as in Fig. 2.1.

Swaths of D_{max} produced by single ensemble members (Fig. 2.7) provide insight into hail predictions. EXP_MY2 overpredicts the size of hail at the surface, with both members 20 (Fig. 2.7a) and 40 (Fig. 2.7d) predicting hail to exceed 150 mm (~ 6 inches) in diameter. The HCA output (Fig. 2.1b) is unable to determine the upper limit on observed hail size for hail exceeding

50 mm because the algorithm classifies all hail 50 mm and larger as significant severe; however, mPING reports suggest that hail, which occurred during this event, did not exceed 3 inches (~ 75 mm) (Fig. 2.1c) in diameter. The MY2 scheme surface hail size overprediction bias is noted in several previous studies; both L17 and Luo et al. (2018) note the scheme produces excessively large surface hail. The MY3 scheme employs the same mass and number concentration tendency equations as the MY2 scheme; as such, EXP_MY3 also overpredicts hail size (Fig. 2.7b, e). Despite an overprediction bias, EXP_MY3 members (Fig. 2.7b, e) predict smaller D_{max} values than EXP_MY2 members (Fig. 2.7a, d). Forecast differences are in part because microphysical processes (e.g., sedimentation) are sensitive to variations in the shape parameter (Milbrandt and Yau 2005b).

EXP_NSSL forecasts (Fig. 2.7c, f) qualitatively most closely match HCA output in terms of maximum hail size. EXP_NSSL members predict the northernmost storm to produce more severe hail than what is observed in HCA output (Fig. 2.1b), though this is partly because we compare low temporal frequency (~ 5 minutes) radar observations to high temporal frequency (~ 1 second) NWP output. Additionally, the NSSL scheme predicts the storms to produce hail cores that are wider than observations.

2.3.2 Hail production processes

Similar to the microphysical budget analysis conducted in L17, this study analyzes hail growth and decay process predicted by each MP scheme to identify sources of forecast error. Hail microphysical tendency terms (e.g., accretion, sublimation, melting) from a single time step are integrated along a horizontal plane that spans the verification domain (Fig. 2.4). Terms are plotted as a function of average model level height in order to illustrate the model's treatment of hail both

above and below the 0 °C isotherm. Vertical cross-sections are taken through the hail core of a supercell thunderstorm during the microphysical budget analysis to illustrate the impact of microphysical processes on the representation of hail. We conduct the microphysical analysis at 2250 UTC; this is one hour into the forecast period when multiple supercell thunderstorms are producing hail.

Hail growth assumptions in MP schemes impact the representation of hail within the model. In EXP_MY2 (Fig. 2.8b) and EXP_MY3 (Fig. 2.8d), hailstones are almost exclusively created by the three-component accretion of rain; this was first reported in Johnson et al. (2016). Three-component accretion of rain occurs when rainwater freezes to the surface of a frozen hydrometeor species to produce a subsequent third ice hydrometeor type (e.g., rain freezing onto snow to create hail). A static threshold is used to determine the resulting hydrometeor category. The NSSL scheme hail category does not contain frozen raindrops (Mansell et al. 2010), instead hail is created from dense graupel (ρ_g > 800 kg m^{-3}) in wet growth conditions (Fig. 2.8f). EXP_NSSL N_{th} tendencies (Fig. 2.8f) are approximately 1 – 2 orders of magnitude smaller than EXP_MY2 (Fig. 2.8b) or EXP_MY3 (Fig. 2.8d) N_{th} tendencies because dense, wet-growth graupel particles are much less numerous than frozen raindrops.

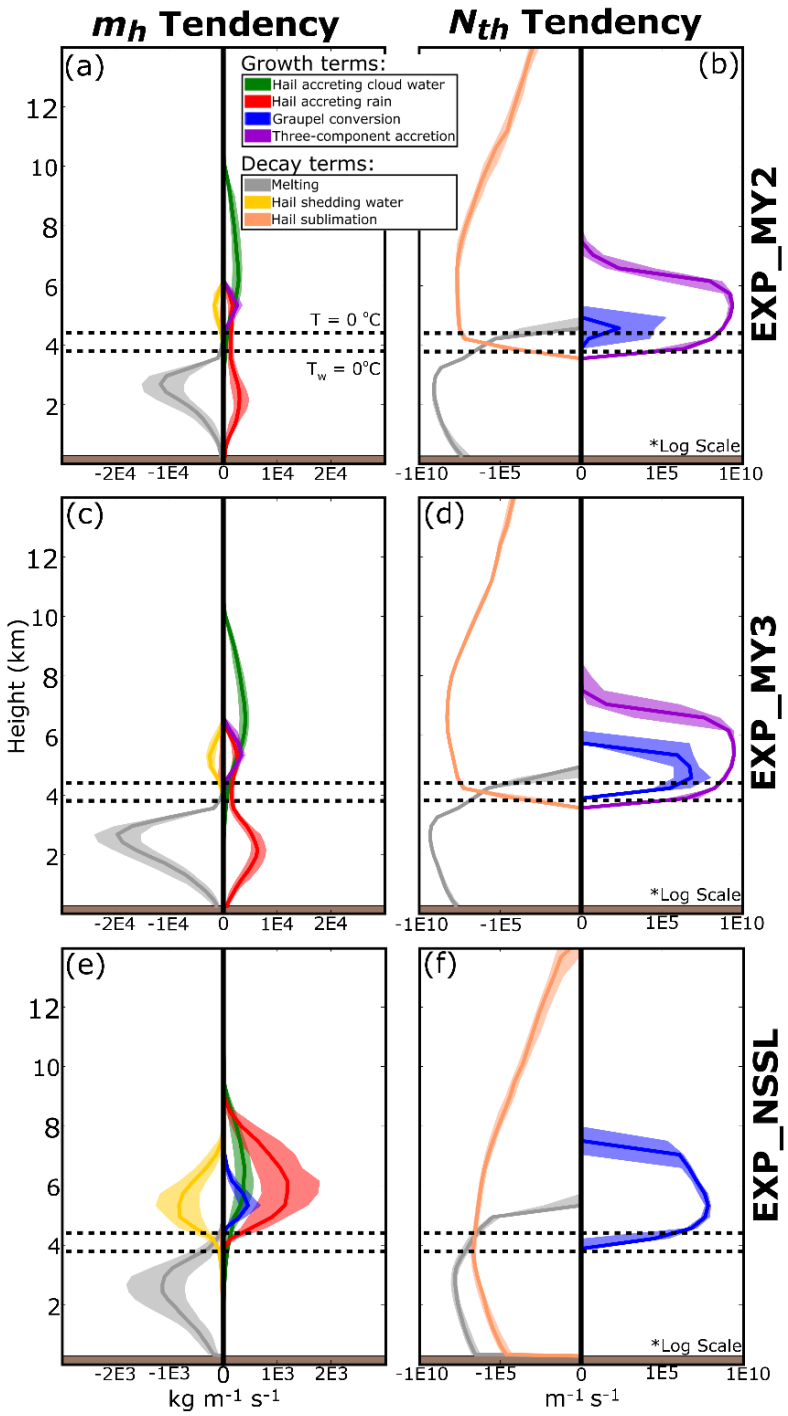

Fig. 2.8. Microphysical source and sink terms for m_h (a, c, e) and N_{th} (b, d, f) predicted by (a - b) EXP_MY2, (c - d) EXP_MY3, and (e - f) EXP_NSSL at 2250 UTC. Microphysical terms are integrated over each model vertical level spanning the verification domain shown in Fig. 2.4. The terms are plotted as a function of average model height per vertical level. Terms are summed for each ensemble member, the interquartile range of the ensemble is shaded and median value is denoted as a dark line. Only terms that contribute greater than 10 kg m^{-1} for m_h and 10 m^{-1} for N_{th} are plotted. In each plot the upper horizontal dashed black line is the average 0 °C isotherm height and the lower horizontal dashed line is the average wet-bulb temperature 0 °C isotherm height.

The three-component accretion of rain increases the number of hailstones above the 0 °C isotherm in the EXP_MY2 (Fig. 2.8b) and EXP_MY3 (Fig. 2.8d) forecasts; however this process contributes relatively little to the total mass of hail aloft because frozen raindrops are relatively small (Fig. 2.8a, c). The two largest m_h growth processes for the MY2 (Fig. 2.8a) and MY3 (Fig. 2.8c) schemes are the accretion of rain and cloud water. Similar to L17, the MY2 (Fig. 2.8a) and MY3 (Fig. 2.8c) schemes both predict hail to accrete rainwater near the surface despite the wet-bulb temperature exceeding 0 °C. The consequences of this process are discussed later in section c of this chapter. Unlike MY2 and MY3, the NSSL scheme assumes all accreted water is shed in and beneath the melting layer for hail, eliminating the possibility of m_h growth. EXP_NSSL m_h tendencies (Fig. 2.8e) are smaller in magnitude than either EXP_MY2 (Fig. 2.8a) or EXP_MY3 (Fig. 2.8c) because the rate of accretion and melting is proportional to hail number concentration. The NSSL scheme predicts fewer hailstones than the MY2 and MY3 schemes, and thus predicts less rain and cloud water droplets are accreted.

Vertical cross-sections of hail-related microphysical variables (Fig. 2.9) are evaluated to understand the impact of microphysical processes on the behavior of hail. Although the EXP_MY2 (Fig. 2.7a, d) and EXP_MY3 (Fig. 2.7b, e) predict storms to produce $D_{max} > 100$ mm at the surface, cross-sections indicate hail is much smaller above the 0°C isotherm (Fig. 2.9g, h). The NSSL scheme predicts storms to produce smaller D_{max} values at the surface (<=75 mm) (Fig. 2.7c, f); however, the hail mean mass diameter (D_{mh}) remains large throughout the vertical depth of the updraft (Fig. 2.9i). Differences in microphysical assumptions cause EXP_MY2 and EXP_MY3 to predict larger hail than EXP_NSSL beneath the 0 °C isotherm, but smaller hail aloft.

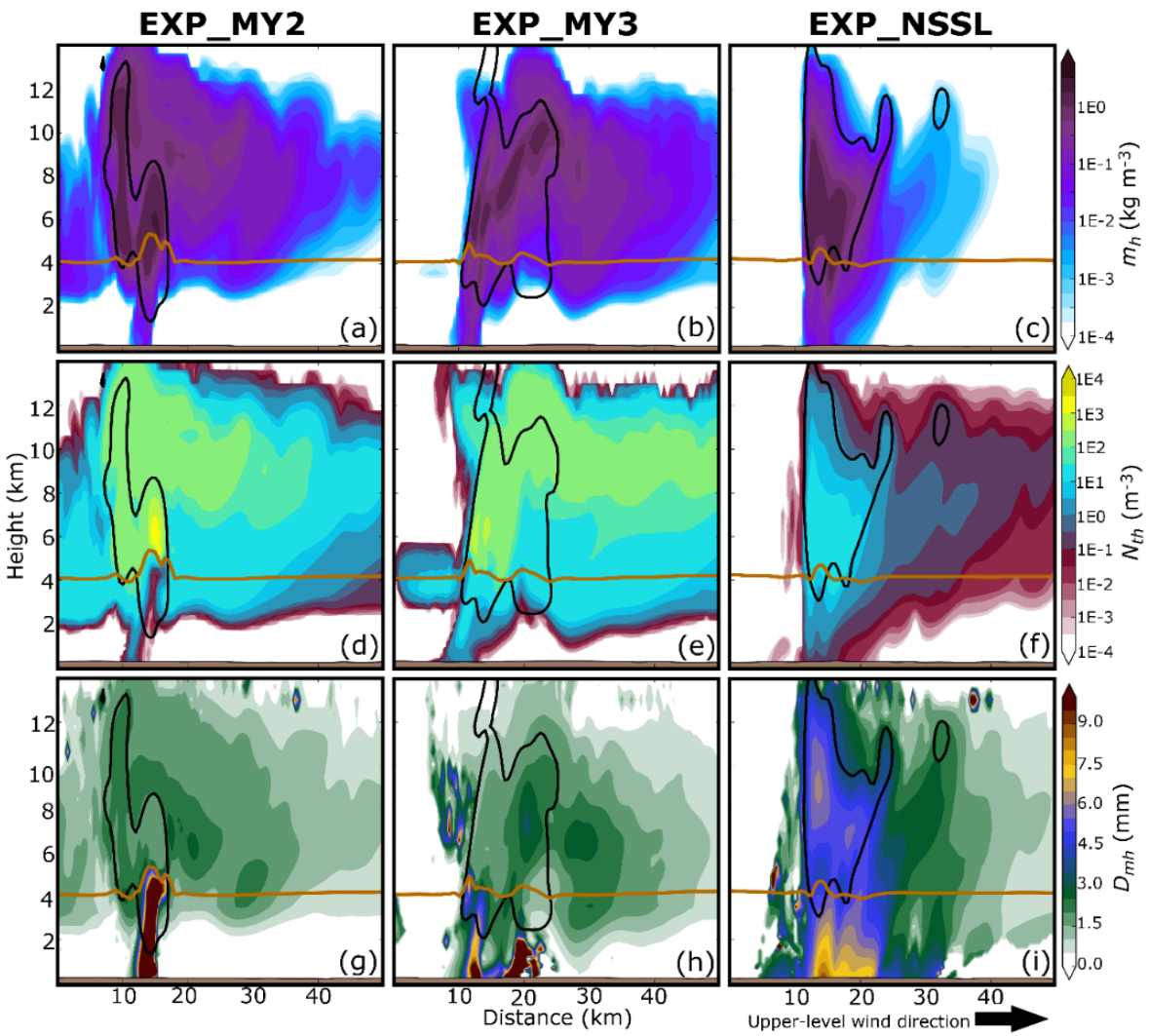

Fig. 2.9. A vertical cross-section taken through the hail core of northernmost supercell thunderstorm predicted by member 20 of (a, d, g) EXP_MY2, (b, e, h) EXP_MY3, and (c, f, i) EXP_NSSL at 2250 UTC. Cross-sections are of (a - c) m_h, (d - f) N_{th}, and (g - i) hail mean mass diameter (D_{mh}). In each plot a brown horizontal line represents the 0 °C isotherm, black lines contour regions where updraft velocity exceeds 5 m s^{-1}.

Both the MY2 (Fig. 2.9d) and MY3 (Fig. 2.9e) schemes predict N_{th} to be large above the 0 °C isotherm at 5 km above mean sea level; this is the approximate level where the three-component accretion of rain is maximized (Fig. 2.8b, d). Since frozen raindrops are relatively small in diameter, the three-component accretion of rain decreases the D_{mh} above the 0 °C isotherm for both EXP_MY2 (Fig. 2.9g) and EXP_MY3 (Fig. 2.9h) members. The small hail particles aloft

are advected downwind of the updraft, creating a plume of hail that extends more than 40 km. Although the MY2 and MY3 schemes produce large plumes of hail aloft, the small hail particles in the plumes mostly melt before reaching the surface. S16 previously noted the MY2 scheme melts hail too quickly and under-predicts the spatial coverage of hail at the surface. The underprediction bias noted in S16 is mitigated when a less stringent minimum N_{th} threshold (10^{-8} m^{-3}) is used, this allows hailstones that are more sparse in number concentration to reach the surface.

Above the 0 °C isotherm, the NSSL scheme produces fewer but larger hailstones than either the MY2 or MY3 schemes because hail is created from wet-growth graupel. Large hail behaves differently from small rimed ice particles: first, upper-level winds do not advect the NSSL scheme hail as far downstream, and most large hail ($D_{mh} > 4.5$ mm) thus remains confined to the boundaries of the storm updraft region (Fig. 2.9i). Second, the larger hailstones have an increased terminal velocity, and thus higher fall speeds. An increased fall speed causes a higher percentage of hailstones to reach the surface before melting and produces relatively wide swaths of severe hail (Fig. 2.7c, f) that closely match observations (Fig. 2.1b, c). Despite producing qualitatively more skilled forecasts, we are unable to verify the predicted hail fall speed.

2.3.3 Hail melting processes

In this study, EXP_MY2 forecasts (Fig. 2.7a, d) predict D_{max} to frequently exceed 150 mm (approximately 6 inches), despite HCA output (Fig. 2.1b) indicating that surface hail is mostly non-severe or severe (< 50 mm in diameter). Much of the hail growth in EXP_MY2 forecasts occurs in the melting layer; EXP_MY2 member 20, which is representative of most EXP_MY2 members, predicts D_{max} to increase from 25 mm to more than 150 mm while traversing the 0 °C

isotherm (Fig. 2.10a). L17 suggests much of this growth is because hail in the MY2 scheme accretes rain and cloud water in and beneath the melting layer.

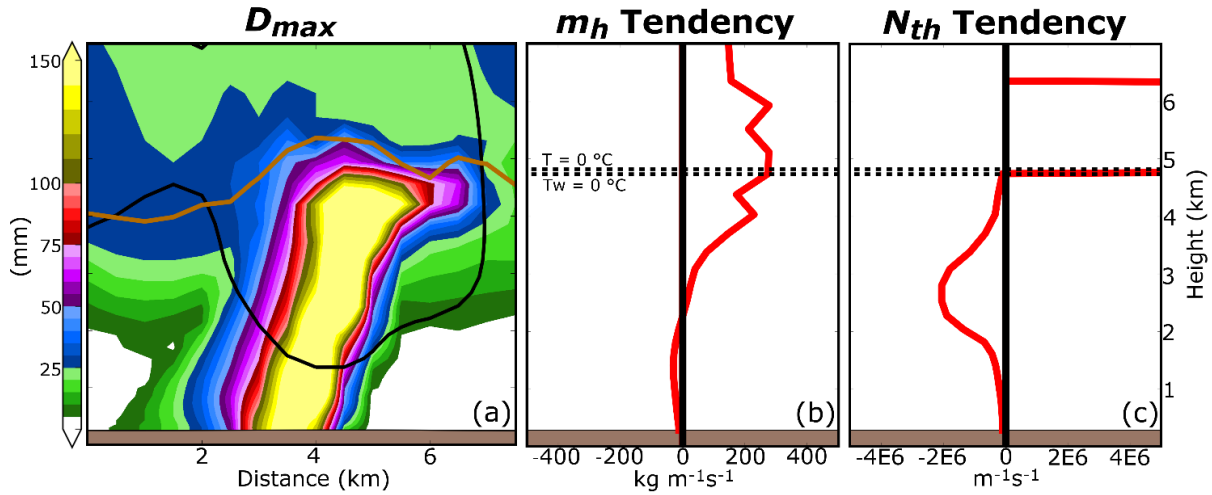

Fig. 2.10. (a) A vertical cross-section of D_{max} taken through the hail core of the northernmost storm predicted by EXP_MY2 member 20 at 2250 UTC. The total microphysical tendency for m_h and N_{th} are also provided in (b) and (c), respectively. The total microphysical tendency is integrated over each model vertical level in the storm hail core, which is highlighted by the black polygon in Fig. 2.7a. The tendencies are plotted as a function of average model level height. In the vertical cross-section (a) the horizontal brown line represents 0 °C isotherm and the black contour represents where updraft velocity exceeds 5 m s^{-1}. In (b, c) the upper horizontal dashed black line is the 0 °C isotherm and the bottom dashed line is the wet-bulb 0 °C isotherm.

To better understand MY2 scheme hail growth, we integrate the total microphysical tendency of m_h and N_{th} over a subdomain that encompasses only the hail core of EXP_MY2 member 20 (Fig. 2.7a). Microphysical tendencies vary substantially throughout a single storm, integrating the tendencies over the hail core (Fig. 2.10b, c) provides more insight into hail growth processes than integrating over the verification domain (Fig. 2.8a, b). Above the 0 °C isotherm N_{th} increases rapidly due to the three-component accretion of rain (Fig. 2.10c). The introduction of many small, frozen raindrops decreases the mean size of hail aloft. Beneath the 0 °C isotherm, m_h increases (Fig. 2.10b) because hail accretes more rain and cloud water than it melts and sheds. At approximately the same layer below the freezing level, N_{th} slowly decreases due to melting

(Fig. 2.10c), this is because accretional growth does not increase N_{th}. When m_h increases but N_{th} decreases the slope parameter of the hail PSD λ_h becomes shallow (Fig. 2.11b) and causes D_{max} to exceed 150 mm (Fig. 2.10a). Accretional growth maintains or expands hail size, allowing more hail to reach the surface.

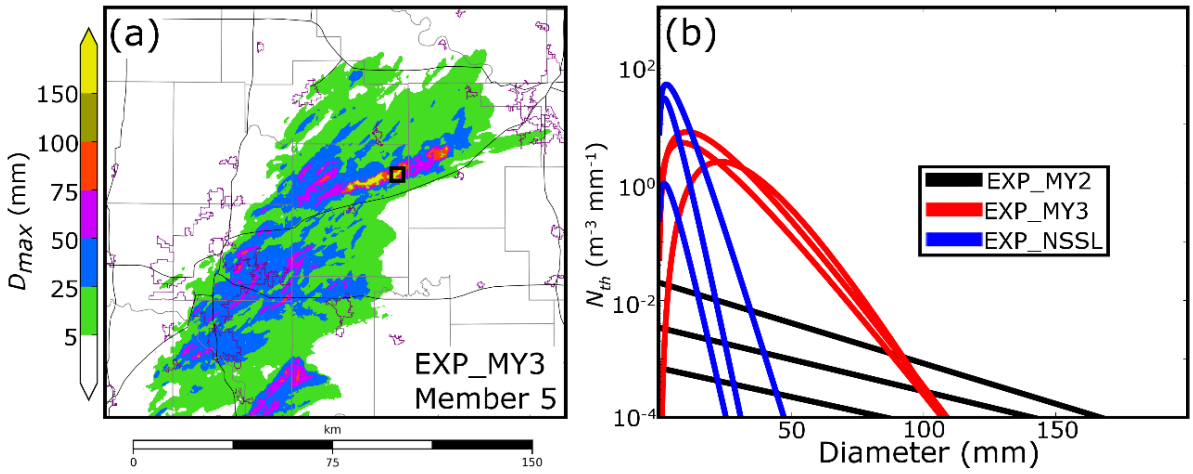

Fig. 2.11. (a) A Swath of forecast maximum surface D_{max} calculated at every model time step between 2155 - 2320 UTC for EXP_MY3 member 5. A black polygon indicates where hail PSDs are sampled for the MY3 scheme. Background maps are the same as in Fig. 2.1(b) A comparison of hail core PSDs for EXP_MY2 member 20 (Fig. 2.7a), EXP_MY3 member 5 (Fig. 2.11a), and EXP_NSSL member 20 (Fig. 2.7c) at 2250 UTC. PSDs are sampled from within the black polygons in Fig. 2.7a, Fig. 2.11a, and Fig. 2.7c, respectively. Because the minimum y-axis value is 10^{-4} m^{-3} mm^{-1} the x-intercept is considered to be the maximum observable hail size D_{max}. For EXP_MY3, a rightward shift in the PSD peak is caused by an increased shape parameter.

EXP_MY3 predicts D_{max} (Fig. 2.7b, e) to exceed 150 mm less frequently than EXP_MY2 (Fig. 2.7a, d), even though both MP schemes use the same m_h and N_{th} tendency equations. Hail PSDs selected from three gridpoints within a hail core of EXP_MY3 member 5 (Fig. 2.11a) indicate that when $\alpha_h > 0$ the PSD follows a gamma distribution which narrows the PSD and moderates D_{max} (Fig. 2.11b). Multiple studies (e.g., Milbrandt and Yau 2005b; Kumjian and Ryzhkov 2012; Dawson et al. 2014) have documented that α_h increases due to sedimentation. Although the MY3 scheme moderates D_{max}, hail continues to accrete liquid in and beneath the

melting layer, which causes EXP_MY3 to predict hail greater than 100 mm in diameter (Fig. 2.7b, e).

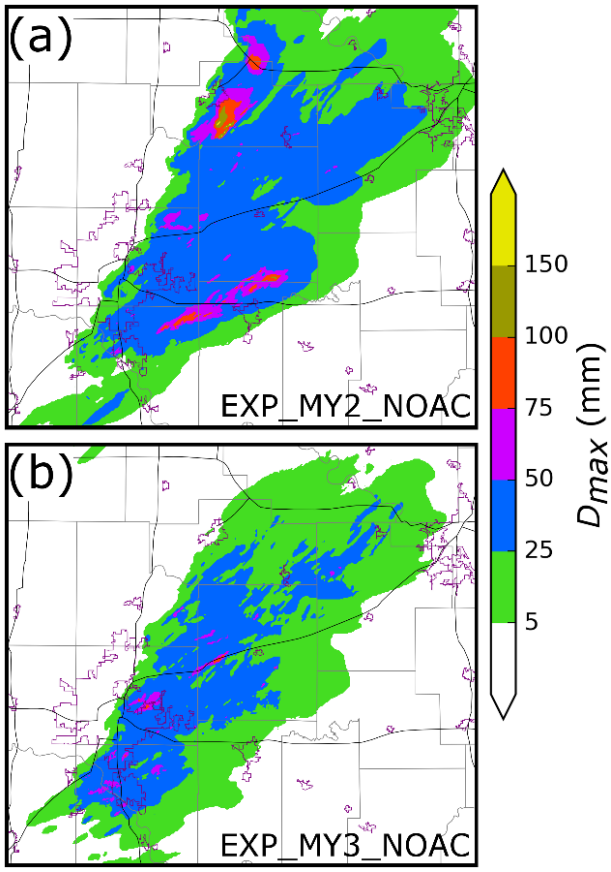

Fig. 2.12. Swaths of forecast maximum surface D_{max} calculated at every model time step between 2155 - 2320 UTC. Forecasts are initialized from member 20 analyses of (a) EXP_MY2 and (b) EXP_MY3. Forecasts (a) EXP_MY2_NOAC and (b) EXP_MY3_NOAC are run with modified MY2 and MY3 schemes, respectively, where the accretion of rainwater and cloud water does not occur beneath the 0 °C isotherm. Background maps are the same as in Fig. 2.1.

EXP_MY2 and EXP_MY3 member 20 forecasts are rerun without the accretion of rain and cloud water beneath the 0 °C isotherm (EXP_MY2_NOAC and EXP_MY3_NOAC, respectively) to demonstrate the sensitivity of surface hail size to the accretion of water beneath the 0 °C isotherm. EXP_MY2_NOAC (Fig. 2.12a) and EXP_MY3_NOAC (Fig. 2.12b) predict the maximum surface hail size to be smaller in diameter than EXP_MY2 (Fig. 2.7a) and

EXP_MY3 (Fig. 2.7b) forecasts, and more closely resemble observations (Fig. 2.1b, c). It is noted that EXP_MY2_NOAC (Fig. 2.12a) predicts hail to exceed 100 mm; however, most of this hail is attributed to initialization from the EXP_MY2 member 20 analysis, and within 10 minutes of forecast initialization most of the storms cease to produce hail greater than 100 mm in diameter. Forecast experiments EXP_MY2_NOAC (Fig. 2.12a) and EXP_MY3_NOAC (Fig. 2.12b) are in agreement with the conclusions of this study and suggest the accretion of liquid water beneath the 0 °C isotherm is important to the production of large hailstones (D_{max} > 100 mm).

Coverage of severe hail increases at the surface for EXP_MY2_NOAC (Fig. 2.12a) and EXP_MY3_NOAC (Fig. 2.12b) when compared to EXP_MY2 (Fig. 2.7a) and EXP_MY3 (Fig. 2.7b) forecasts. In all three MP schemes the N_{th} melting tendency term is used to preserve D_{mh} during melting; however, both the MY2 and MY3 schemes reduce the *Nth* melting term by 90% when accretional growth occurs. Except in updrafts where large quantities of liquid water are accreted, when the N_{th} melting term is reduced D_{mh} decreases because m_h decreases more quickly than N_{th}. When accretion of liquid water is eliminated beneath the 0°C isotherm the N_{th} melting tendency is not reduced, and thus EXP_MY2_NOAC and EXP_MY3_NOAC predict hail to more frequently exceed 25 mm in diameter (Fig. 2.12).

EXP_NSSL exhibits limited hail growth across the 0 °C isotherm. EXP_NSSL member 20, for example, predicts a change of only 0 - 5 mm across the melting layer (Fig. 2.13a), compared to a change of 150 mm across the melting layer in EXP_MY2 member 20 (Fig. 2.11a). Microphysical tendency terms in the hail core of EXP_NSSL member 20 (Fig. 2.7c) suggest m_h (Fig. 2.13b) and N_{th} (Fig. 2.13c) decrease at approximately the same rate beneath the 0 °C isotherm. Both terms decrease at the same rate because hail sheds all accreted liquid when the air temperature is warmer than 0 °C. Although the NSSL scheme limits hail growth within the melting layer,

EXP_NSSL forecasts severe and significant severe hail (Fig. 2.4c, f) because large hailstones are created from the wet-growth of graupel above the 0 °C isotherm.

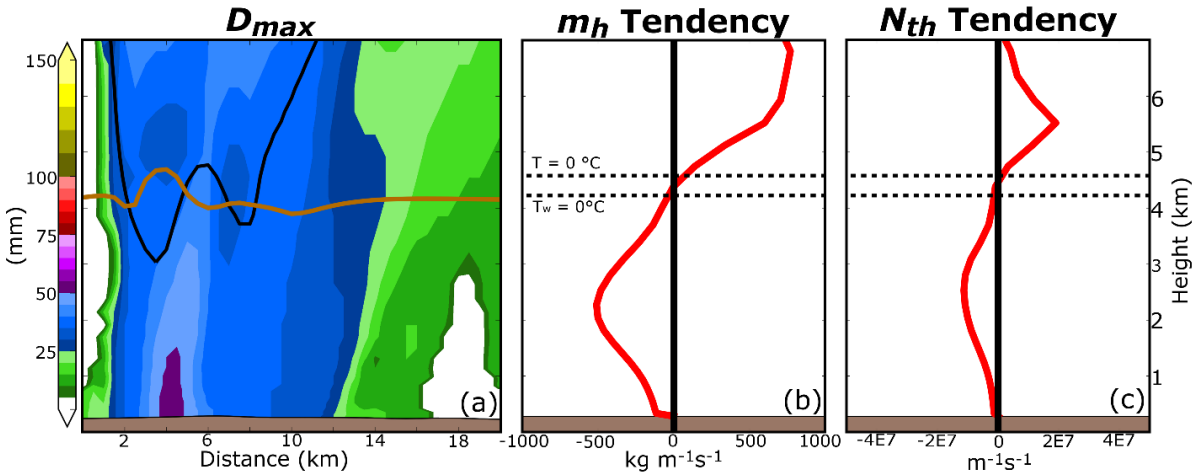

Fig. 2.13. (a) A vertical cross-section of D_{max} taken through the hail core of the northernmost supercell predicted by EXP_NSSL member 20 at 2250 UTC. The total microphysical tendency for m_h and N_{th} are also provided in (b) and (c), respectively. The total microphysical tendency is integrated over each model vertical level in the storm hail core, which is highlighted by a polygon in Fig. 2.7c. The total tendency is plotted as a function of average model level height. In the vertical cross-section (a) the horizontal brown line represents 0 °C isotherm and the black contour represents where updraft velocity exceeds 5 m s^{-1}. In (b, c) the upper horizontal dashed black line is the 0 °C isotherm and the bottom dashed line is the wet-bulb 0 °C isotherm.

Most MP schemes assume ice is dry; this assumption has important consequences on the treatment of hail beneath the 0 °C isotherm. Collection of water beneath the 0 °C isotherm will increase the ice water fraction and potentially hail size (Lesins and List 1986); however, if too much liquid water is collected then the hailstone will shed and evaporate excess liquid (Rasmussen and Heymsfield 1987; Chong and Chen 1974). Instead of representing the collection of water, the MY2 and MY3 schemes freeze rain and cloud water onto the surface of hail (i.e., accretion). Since the accreted water is frozen, excess liquid cannot be removed via shedding, and instead m_h increases even when the air temperature is well above freezing (Fig. 2.10b). In contrast, the NSSL scheme does not allow any liquid water collection beneath the 0 °C isotherm (Fig. 2.13b); this

limits a potential hail growth process within the scheme and thus potentially limits the maximum surface hail size. Without the ability to represent ice-water hybrid categories these MP schemes must continue to make assumptions on the treatment of hail during melting.

2.4 Summary and discussion

In this study we evaluate 90 minute-long EnKF-initialized explicit hail ensemble forecasts at 500-m grid spacing for the 19 May 2013 Oklahoma City supercell hailstorm event. For the initial conditions, we assimilate all available surface, profiler, and radar (reflectivity and radial velocity) observations every 10 minutes, over a 110 minute period ending at the final analysis time of 2150 UTC. Three ensembles, consisting of 40 members each, are run for both the data assimilation and forecast periods. Each ensemble is similarly configured but uses one of three microphysics (MP) schemes: the Milbrandt and Yau (2005a) double-moment MP scheme (MY2), the Milbrandt and Yau (2005a) triple-moment MP scheme (MY3), or the NSSL double-moment variable density rimed ice MP scheme (Mansell et al. 2010). Ensembles run using these schemes are hereafter referred to as EXP_MY2, EXP_MY3, and EXP_NSSL, respectively.

Surface hail size forecasts are verified against output from a hydrometeor classification algorithm (HCA). We determine the forecast maximum hail size (D_{max}) via hail particle size distributions (PSDs), which approximates the largest observable hailstone at each model grid point. Area under the relative operating characteristic (AUC) scores indicate all three MP schemes predict the coverage of severe (diameter > 25 mm) and significant severe (diameter > 50 mm) hail with a high level of skill (≥ 0.92); albeit Brier skill scores suggest the schemes predict significant severe hail coverage with less skill than severe hail coverage. The ensembles produce reliable hail size forecasts, with the exception of EXP_NSSL, which produces under confident significant

severe hail forecasts. Although EXP_MY2 and EXP_MY3 produce skillful severe and significant severe hail forecasts, both ensembles overestimate the maximum surface hail size. Qualitatively, EXP_NSSL most closely resembles HCA output but the scheme predicts limited significant severe hail coverage.

Hail production tendencies are evaluated during the forecast period to understand hail size forecast differences and biases. Storms predicted by either EXP_MY2 or EXP_MY3 generally produce a large number of small hail particles above the 0 °C isotherm because the MP schemes create hail via the three-component accretion of rain (i.e., frozen raindrops). The small hail particles are advected downstream of the updraft to produce large plumes of hail aloft that extend more than 40 km downstream of the updraft. EXP_NSSL predicts storms to produce fewer but relatively larger hail particles above the 0 °C isotherm because the NSSL scheme hail category consists of only dense graupel that has undergone wet growth.

Hail size forecast skill is also determined by how MP schemes model hail in and beneath the melting layer. Because water fraction is not predicted by the MY2, MY3, or NSSL schemes, neither scheme is able to represent the collection of liquid water in the melting layer. Both the MY2 and MY3 MP schemes assume all water collected by hail is converted to ice both above and beneath the melting layer; this process causes model predicted hail to exceed 150 mm (~6 inches) in diameter beneath the 0 °C isotherm. In regions where hail is exceptionally large the MY3 scheme often increases the shape parameter (α_h) of hail. Increasing α_h narrows the hail PSD and moderates the most extreme hail sizes (diameter > 150 mm), however EXP_MY3 continues to overpredict hail size. The NSSL scheme sheds all accreted liquid water when the air temperature is warmer than 0 °C. While this assumption prevents the rapid growth of hail in the melting layer,

it is not realistic when compared to observations (e.g., Rasmussen et al. 1984) and neglects potential hail growth via the collection of water (Lesins and List 1986).

To further improve the hail forecast skill, we must use MP schemes that accurately model hail growth processes. MP schemes such as the predicted particle properties (P3; Morrison and Milbrandt 2015; Morrison et al. 2015; Milbrandt and Morrison 2016) scheme and the Ice-Spheroids Habit Model with Aspect-ratio Evolution (ISHMAEL; Jensen et al. 2017) scheme explicitly predict the mass and volume of accreted rime to avoid converting between pre-defined hydrometeor species. Hydrometeor conversion between un-rimed and rimed ice categories is a large source of model error, this is most evident in EXP_MY2 and EXP_MY3 forecasts when frozen rain is converted to hail. ISHMAEL also predicts particle shape, which is shown to impact the spatial distribution of ice due to size sorting (Jensen et al. 2018). Further investigation into the inclusion of prognostic water fraction equations such as those in Ferrier (1994) could also improve representation of ice within the melting layer. Integrating these novel MP schemes into a hail prediction system could be part of an investigation of the benefit of including additional microphysical variables and processes.

Data assimilation can also be used to improve the microphysical state of hailstorms, however we must determine if microphysical variables are being realistically updated by the DA system. Performing data assimilation when using a multi-moment MP scheme is non-trivial. Due to the large number of predicted hydrometeor properties, microphysical variables are typically under constrained by observations (i.e., reflectivity). Microphysical variables also have very nonlinear relations with observations; Xue et al. (2010) notes this introduces large errors into the initial conditions. For this study, we update only hydrometeor mixing ratios to limit the degrees of freedom of the model that are updated. Although this technique produces a stable configuration,

the resulting analysis is suboptimal because D_{max} is often too large in early cycles when forecast errors are large. An optimal data assimilation procedure for multi-moment MP schemes should be developed to improve initial condition estimates and produce a more skilled hail size forecast.

A limited number of studies have analyzed hail forecasts for storm modes other than supercell thunderstorms. Many non-rotating storms produce large hail and cause extensive damage, one such hail event caused more than $1.4 billion in hail damage in Colorado (Fritz 2017). Studies such as Luo et al. (2017) have evaluated the skill of next-day CAM forecasts for a severe-hail-producing pulse thunderstorm, but to our knowledge no studies have evaluated hail forecasts for non-rotating storms using a frequently updated, high-resolution ensemble forecast framework.

Chapter 3 [2]Evaluating Ensemble Kalman Filter Analyses: Effects of State Variable Updating and Multi-Moment Microphysics Schemes on State Variable Cross-Covariances

3.1 Introduction

Hail causes significant damage to both crops and personal property in the United States—often exceeding $1 billion in damage annually (Changnon 2009). Increasing the warning lead time of severe hail events can potentially mitigate damage by allowing more time for preparation. Current warning lead times for severe hail are limited because the National Weather Service (NWS) issues warnings based upon the detection of severe hail (diameter > 1.0 inch). Generally, hail detections are based upon either surface-based reports or radar signatures that are indicative of severe hail. To extend severe hail warning lead times the NWS is investigating a warn-on-forecast framework (Stensrud et al. 2009, 2013), where instead of issuing warnings based upon the detection of severe hail, the NWS will issue warnings based upon high-resolution, frequently-updated NWP model guidance. While convection-allowing forecasts can skillfully predict convective hazards (e.g., Kain et al. 2008, 2010; Clark et al. 2012; Dawson et al. 2012; Snook et al. 2016, 2018; Yussouf et al. 2015; Yussouf and Knopfmeier 2019; Johnson et al. 2015; Johnson and Wang 2017; Schwartz et al. 2015; Skinner et al. 2018; Dawson et al. 2017; Labriola et al. 2017, 2019a; Jones et al. 2016, 2019; Supinie et al. 2016, 2017; Gallo et al. 2019; Stratman et al. 2020), forecast skill is sensitive to initial conditions and model physics errors.

Previous studies have shown, either through observing system simulation experiments

[2] A paper based on the material presented in this chapter has been accepted for publication as: Labriola, J., N. Snook, Y. Jung, and M. Xue, 2020: Evaluating Ensemble Kalman Filter Analyses of Severe Hailstorms on 8 May 2017 in Colorado: Effects of State Variable Updating and Multi-Moment Microphysics Schemes on State Variable Cross-Covariances. Mon. Wea. Rev.

(OSSEs) or real-data experiments (e.g., Snyder and Zhang 2003; Zhang et al. 2004, 2006; Dowell et al. 2004; Tong and Xue 2005; Caya et al. 2005; Xue et al. 2006; Snook et al. 2011; Dawson et al. 2012; Romine et al. 2013; Schwartz et al. 2015; Johnson and Wang 2017; Lawson et al. 2018; Stratman et al. 2020) that an EnKF (Evensen 1994, 2003) can be successfully applied to convection-resolving forecasts. EnKF methods employ an ensemble of forecasts to sample error covariance. Flow-dependent error covariances derived from ensemble forecasts are used to correct errors in unobserved variables during data assimilation (DA) that strongly influence convective-scale dynamics such as updraft speed, in-cloud temperature, and microphysical properties (Tong and Xue 2005, hereafter TX05; Tong and Xue 2008a). This feature of EnKF is particularly useful when assimilating radar data (e.g., Snyder and Zhang 2003; Zhang et al. 2004; Dowell et al. 2004; Tong and Xue 2005; Caya et al. 2005; Dowell and Wicker 2009; Aksoy et al. 2009), which provide indirect but high spatial and temporal resolution observations.

Even when forecasts are initialized from relatively accurate initial conditions, forecast error will continue to increase with time due to model errors including those associated with sub-grid scale processes (Zhu and Navon 1999; Houtekamer et al. 2005; Zhang et al. 2006; Hawblitzel et al. 2007; Melhauser and Zhang 2012; Zhang et al. 2015). For convective-allowing model simulations, microphysical parameterizations are a large source of model uncertainty is (e.g., Morrison et al. 2015; Johnson et al. 2016). These parameterizations represent fine, sub-grid scale processes that undergo rapid, non-linear transformations at convective scales and often rely upon ad-hoc procedures due to the limited understanding of microphysical processes.

Bulk microphysics schemes (referred to as MP schemes) are the most commonly used type of microphysical parameterization in convective modeling studies. MP schemes predict the evolution of an assumed hydrometeor PSD that typically follows a gamma distribution (Ulbrich

1983). The 0^{th}, 3^{rd}, and 6^{th} moments of a PSD are most commonly predicted by MP schemes and are proportional to hydrometeor number concentration (N_{tx}), mixing ratio (q_x), and reflectivity (Z_x), respectively. The number of predicted moments corresponds to the number of diagnosed PSD parameters. All undiagnosed parameters are typically assumed to be constant.

While single-moment MP schemes (e.g., Lin et al. 1983) are generally computationally efficient because they predict only hydrometeor mixing ratio, they are unable to accurately represent microphysical processes such as selective melting/evaporation and size sorting. Multi-moment MP schemes (e.g., Ferrier 1994; Milbrandt and Yau 2005a; Morrison et al. 2005; Morrison and Grabowski 2008; Thompson et al. 2008; Mansell et al. 2010; Milbrandt and Morrison 2013; Morrison and Milbrandt 2015; Milbrandt and Morrison 2016; Lim and Hong 2010) can replicate these microphysical processes and improve the representation of storm structure. Double-moment schemes that predict both q_x and N_{tx} (diagnose N_{0x}, λ_x), can reproduce observed polarimetric radar signatures (Jung et al. 2012; Johnson et al. 2016; Putnam et al. 2014, 2017a,b); however, multiple studies (e.g. Milbrandt and Yau 2005b; Dawson et al. 2014; Johnson et al. 2016; Morrison et al. 2015) note these schemes suffer from excessive size sorting. Excessive size sorting can be mitigated using a correction mechanism (Thompson et al. 2008; Mansell 2010), a diagnostic shape parameter, or by using a triple-moment scheme (Milbrandt and Yau 2005b).

In some newer MP schemes, instead of predicting many static hydrometeor categories (e.g., Straka and Mansell 2005), the scheme predicts the evolution of particle characteristics such as density (Mansell et al. 2010; Milbrandt and Morrison 2013; Morrison and Milbrandt 2015; Morrison et al. 2015; Milbrandt and Morrison 2016). MP schemes generally predict the density of rimed ice; this is because ice particle density can vary substantially, and because ice particles can undergo large fluctuations in density during the riming process. Prognostic density equations

allow the MP scheme to update rimed ice fall speeds and improve the representation of hail production processes (Labriola et al. 2019a).

Predicted microphysical variables can be used to diagnose surface hail size in a NWP forecast, either via simulated radar-derived hail products (e.g., Snook et al. 2016; Luo et al. 2017, 2018; Labriola et al. 2017) such as the MESH (Witt et al. 1998a), or by using hail PSDs to diagnose the largest observable hail size (e.g., Snook et al. 2016; Labriola et al. 2017, 2019a,b; Luo et al. 2018; Gagne et al. 2019). Hail size forecasts not only provide useful information for forecasters but are also useful in evaluating MP scheme treatment of hail growth and decay processes. Due to the limited coverage and biased nature of surface hail reports (e.g. Sammler 1993; Witt et al. 1998b; Doswell et al. 2005), most hail size forecasts are evaluated against radar-derived hail products that serve as a proxy for hail size such as MESH or output from a HCA (Heinselman and Ryzhkov 2006; Park et al. 2009; Ryzhkov et al. 2013; Ortega et al. 2016).

Over the continental United States, the NEXRAD system (Crum et al. 1993) provides full volumetric scans at a temporal resolution sufficient to observe the rapid evolution of convective storms. Reflectivity (Z) is proportional to the 6th moment of a hydrometeor PSD and is strongly influenced by larger particles within a radar volume. After a recent polarimetric upgrade to the NEXRAD system, radars now emit both horizontally and vertically polarized signals. Dual-polarization variables such as differential reflectivity (Z_{dr}), co-polar cross-correlation coefficient (ρ_{hv}), and specific differential phase (K_{dp}), which in addition to single-polarization products, Z and radial velocity (V_r), provide information about hydrometeor size, orientation, phase, and shape (Doviak et al. 2000; Kumjian and Ryzhkov 2008). Due to the large amount of hydrometeor information that can be inferred from polarimetric radar products, several recent studies have used simulated or observed polarimetric variables (e.g., Jung et al. 2010a, 2012; Ryzhkov et al. 2011;

Kumjian et al. 2014; Dawson et al. 2014; Putnam et al. 2014, 2017a,b, 2019 Johnson et al. 2016, 2018, 2019; Snook et al. 2016; Snyder et al. 2017a,b) to evaluate microphysical parameterizations.

Generally, Z is the primary observation of microphysical relevance assimilated during DA. Despite a limited number of observations, OSSEs have shown that an EnKF system can reasonably update single-moment MP scheme variables when assimilating Z and V_r (TX05). Multi-moment schemes, which predict approximately twice as many microphysical variables as a single-moment scheme, are even more under-constrained by observations. Although Xue et al. (2010) (hereafter XJZ10) obtained a reasonably good analysis using a double-moment scheme in an OSSE, it was noted that Z alone may be insufficient to constrain the increased number of variables in multi-moment schemes because many different microphysical configurations can correspond to the same Z. Additionally, the large number of predicted variables introduces error into the analysis because increasing the number of error-containing parameters decreases the accuracy of the estimation (Aksoy et al. 2006; Tong and Xue 2008b; Jung et al. 2010b)

Relatively few studies have considered developing an optimal EnKF-based framework for updating multi-moment MP scheme variables during DA. Previously, to improve constraint of microphysical variables, XJZ10 assimilated both radar Z and V_r, and Labriola et al. (2019a) used an EnKF system to update only hydrometeor mixing ratio; however, both studies conclude that there is a need to further analyze the ability of an EnKF system to update multi-moment scheme variables. This study evaluates EnKF analysis sensitivity to selective state variable updating and the microphysical assumptions made within two double-moment MP schemes used in the prediction model. Knowledge gained through this study provides information that can be used to optimize EnKF configurations when using such multi-moment MP schemes. To this end, EnKF analyses are evaluated for the 8 May 2017 Colorado severe hail event. The rest of the chapter is

organized as follows: in section 3.2 a brief overview of the 8 May 2017 Colorado severe hail event is provided along with the experimental design, including the EnKF and prediction model settings. Verification statistics and ensemble correlation analyses are provided in section 3.3. Finally, section 3.4 contains a summary and discussion of the results.

3.2 Methods

3.2.1 Case overview

On 8 May 2017, multiple hail producing thunderstorms occurred over North-Central Colorado. According to surface-based hail reports received by the Storm Prediction Center (SPC), these storms produced severe (diameter > 25 mm) and significant severe (diameter > 50 mm) hail starting at approximately 2000 UTC. One storm that impacted downtown Denver (Fig. 3.1) caused a particularly large amount of damage; due to heavy traffic volume during evening rush hour, a large number of cars sustained major hail damage during the storm (Fritz 2017). As of 2018, an estimated $2.3 billion in insurance claims (NOAA 2018) have been filed in Colorado for hail damage sustained during this severe weather event.

Atmospheric conditions were conducive for the development of supercell thunderstorms with moderate shear (0-6 km shear: 30 – 40 knots) and instability (mixed layer CAPE: 1000 – 1500 J kg^{-1}) (Marsh 2017). On 8 May 2017, Colorado was located downstream of a trough associated with an upper-level low-pressure system located over Baja California. Cyclonic vorticity advection from the trough paired with low-level convergence along a weak frontal boundary initiated multi-cellular storms to the south and east of Denver at approximately 1930 UTC. Additional, more isolated storms initiated along the front-range of the Rocky Mountains via upslope flow, one of these storms later produced significant severe hail over Denver (Fig. 3.1c).

For additional information regarding this severe weather event we refer the reader to Labriola et al. (2019b).

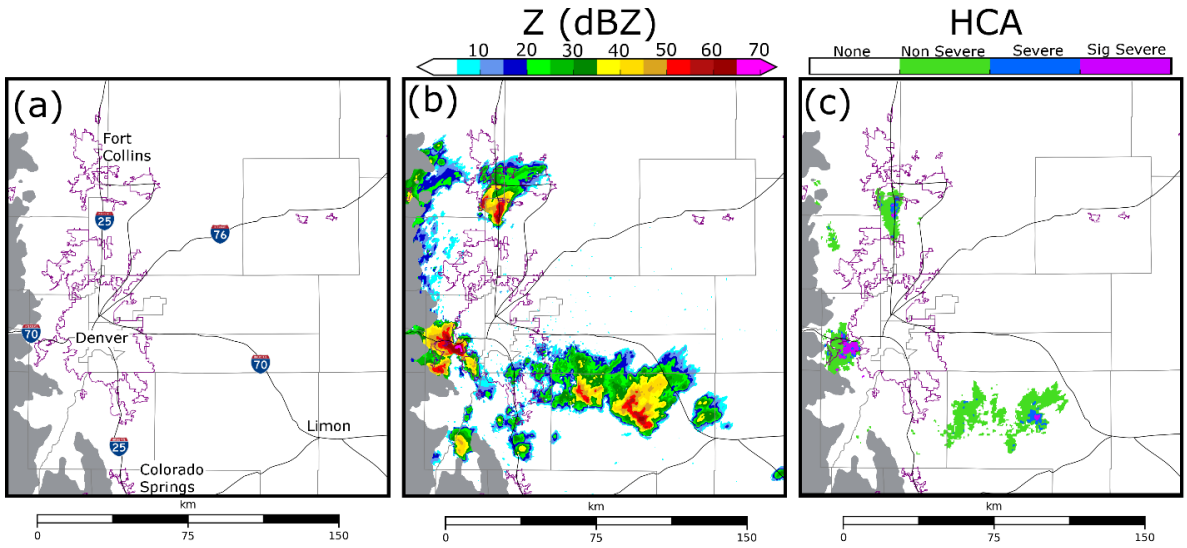

Fig. 3.1. (a) An areal map of the Denver region. (b) Observed *Z* from the lowest radar tilt (0.5°) of KFTG (Denver) at 2040 UTC. (c) Merged HCA output between 2000 – 2040 UTC. The HCA is applied to the lowest radar tilts (0.5°) of the two closest radars: KFTG (Denver) and KCYS (Cheyenne). Purple contours represent urban boundaries, thin black lines represent major highways. All regions of the domain that are more than 2.5 km above mean sea level are shaded gray.

3.2.2 Prediction model settings

All experiments use the ARPS; Xue et al. 2000, 2001) NWP model on a domain of 483 × 443 × 53 grid points with a horizontal grid spacing of 500 m. The model grid is stretched in the vertical, with a minimum grid spacing of 50 m at the surface and an average vertical grid spacing of 425 m. Model physics include NASA Goddard Flight Center short- and longwave radiation parameterization (Chou 1990, 1992; Chou and Suarez 1994); surface fluxes calculated from surface drag coefficients, surface temperature, and volumetric water; a two-layer soil model; and a 1.5-order turbulent mixing and planetary boundary layer parameterization (Xue et al. 2000, 2001).

Ensemble forecasts are run using either the Milbrandt and Yau (2005a) double-moment (MY) MP scheme or the NSSL double-moment variable density rimed ice MP scheme (Mansell et al. 2010). These MP schemes were selected for this study because they are commonly-used multi-moment MP schemes used for hail prediction (Milbrandt and Yau 2006; Snook et al. 2016; Luo et al. 2017, 2018, Labriola et al. 2017, 2019a,b) and the hail growth processes are well understood as demonstrated in the previous chapter. Following Labriola et al. (2019a), the minimum hydrometeor number concentration threshold is set to 10^{-8} m^{-3} for both schemes to avoid excessive removal of hail near the surface which can result when using the MY default setting (10^{-3} m^{-3}) .

Initial and boundary conditions are interpolated from the Center for Analysis and Prediction of Storms (CAPS) EnKF storm scale ensemble forecast (SSEF; Jung et al. 2018)—a 40-member ensemble of forecasts run using the Advanced Research version of the Weather Research and Forecasting model (WRF-ARW; Skamarock et al. 2008). The SSEF uses 3-km horizontal grid spacing, and its model domain spans the contiguous United States. The SSEF was initialized via EnKF on 8 May 2017 at 1800 UTC from the North American Mesoscale Forecast System (NAM; Environmental Modeling Center 2017) analysis, with initial perturbations introduced by the Short-Range Ensemble Forecast (SREF; Du et al. 2015). The SSEF is run until 9 May 2017 at 0000 UTC. At 1903 UTC the 500-m domain is interpolated from the SSEF, with boundary conditions obtained from the SSEF every 9 minutes; these times were selected because the SSEF outputs model data every 9 minutes. Following (Labriola et al. 2017), no microphysical information is provided along domain boundaries to avoid development of spurious convection,. For additional information regarding the 2017 CAPS SSEF, we refer the reader to CAPS (2017).

3.2.3 DA and observation operator settings

After initialization, the 500-m ensemble forecasts undergo a ~60-minute spin-up period until 2000 UTC, when DA cycling begins (Fig. 3.2). The spin-up period is constrained to only 60 minutes because the CAPS SSEF is initialized at 1800 UTC. In the future it is recommended additional tests examine the optimal spin-up period length for high-resolution forecasts. The CAPS EnKF system (Xue et al. 2006; Tong and Xue 2008b), which is based upon the Whitaker and Hamill (2002) ensemble square-root filter (EnSRF) algorithm, assimilates all available observations every 5 minutes between 2000 and 2040 UTC (Fig. 3.2).

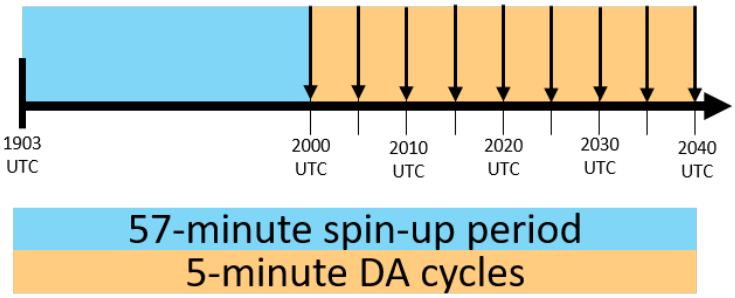

Fig. 3.2. The experiment configuration for ensembles MY-Q, MY-ALL, NSSL-Q, and NSSL-ALL. The blue region represents the approximate 1-hour spin-up period and the orange region represents the 40-minute period when 5-minute DA cycling occurs. Downward pointing arrows indicate where assimilation occurs.

Surface (ASOS and AWOS) observations and data from the Denver NEXRAD radar (KFTG) are assimilated during the experiment. Observation error settings follow Snook et al. (2016). For surface observations, errors are assumed to be 1.5 m s^{-1} for horizontal wind components u and v, 2.0 K for potential temperature θ, and 2.0 K for dewpoint temperature T_d. The covariance localization radius is set to 300 km in the horizontal and 6 km in the vertical. Surface observations update only select model state variables (i.e., u, v, θ, q_v), as was done in Labriola et al. (2019a) because unreliable covariances led to the intensification of spurious updrafts

during DA. KFTG observation errors are assumed to be 4.0 m s^{-1} for V_r and 6.0 dBZ for Z. These values are larger than typical instrument errors associated with NEXRAD radars (e,g., Doviak and Zrnić 1993; Ryzhkov et al. 2005), but for data assimilation purposes, the errors also include representative errors and other sources of uncertainties. Also, specifying relatively large observation errors can help alleviate under-dispersion problems of the EnKF ensemble and past sensitivity studies have shown better performance of the ensemble analyses and subsequent ensemble forecasts (e.g., Dowell et al. 2004; Snook et al. 2013). The Gaspari and Cohn (1999) covariance localization cutoff radius is set to 3 km in both the horizontal and vertical. KFTG data are interpolated horizontally to the model domain, but the height of the radar beam is preserved in the vertical (Xue et al. 2006). Z and V_r observations are pre-processed in the horizontal directions to 1 km grid spacing in regions of precipitation (Z > 5 dBZ) and to 2 km spacings in regions of clear air (Z < 5 dBZ). Given that the radar observations have average resolution of about 1 km, such observation density preserves most useful information in the observations. The relaxation-to-prior spread (Whitaker and Hamill 2012) algorithm is used to inflate the posterior ensemble spread to 95% of the prior spread for all thermodynamic variables (i.e., u, v, w, θ, p) and hydrometeor mixing ratio q_x. Inflating all hydrometeor state variables (i.e., q_x, N_{tx}, and v_x) can potentially cause PSDs to become unbalanced, and cause hydrometeor properties to becomes unrealistic (e.g., hail is too dense). Following Labriola et al. (2019a), N_{tx} and v_x are updated during inflation to preserve EnKF estimated mean mass diameter and density, respectively.

The radar observation operator used during this experiment, which is a variant of the Jung et al. (2010a) T-matrix method, uses look-up tables based upon PSD parameters to calculate the hydrometeor scattering amplitude; an in-depth discussion of this observation operator is provided in Putnam et al. (2019). Generally, MP schemes do not track melt water, instead the water fraction

is diagnosed in the observation operator. The Jung et al. (2008a) melting model is used to create an ice-water hybrid mixture for both hail and graupel in the observation operator. The mixing ratio of the ice-water hybrid category is determined via a ratio between ice and rainwater mixing ratios, the number concentration is updated to preserve particle mean mass diameter and prevent the excessive accumulation of rainwater on the surface of ice particles (Labriola et al. 2019a).

A total of four experiments are conducted during this study; ensembles run using either the MY or NSSL schemes will assimilate observations where either Z updates: (1) hydrometeor mixing ratios only or (2) all microphysical state variables (q_x, N_{tx}, v_x); in addition to dynamic and thermodynamic variables (u, v, w, θ, q_v). To determine the impact of assimilating Z observations, no other observations update microphysical state variables. Experiments (Table 3.1) are named after the MP scheme used and the microphysical variables that are updated (i.e., MY-Q, MY-ALL, NSSL-Q, NSSL-ALL). In addition to assimilating Z, V_r is used in all experiments to update u, v, and w, along with water vapor mixing ratio (q_v). The V_r forward observation operator does not consider hydrometeor terminal velocities due to the large uncertainties in hydrometeor size distribution and fall speed assumptions and small elevation angles of radar beams.

	***Z* Updates Only q_x**	***Z* Updates all Hydrometeor State Variables**
MY Scheme	MY-Q	MY-ALL (q_x, N_{tx})
NSSL Scheme	NSSL-Q	NSSL-ALL (q_x, N_{tx}, $v_{g,h}$)

Table 3.1. A list of the experiments conducted in the chapter.

Analyses and forecasts during the DA period (2000 – 2040 UTC) are objectively and subjectively evaluated to determine which experiment produces the most accurate initial state

estimate. This time period was selected for evaluation because radar derived hail products (i.e., HCA output) indicate that multiple thunderstorms were producing significant severe surface hail, the largest HCA hail size bin. While more substantial damage reports were received from the public for these storms after 2040 UTC when they impacted downtown Denver, they produced similarly large hailstones both during and after the DA period. Forecast evaluation after 2040 UTC is beyond the scope of this study, instead we refer the reader to previous studies (e.g., Snook et al. 2016; Luo et al. 2018) that evaluate surface hail size forecasts.

3.3.4 Analysis evaluation procedure

Due to the nature of this severe weather event, estimated surface hail size is verified against radar-based hail size estimates. Surface-based hail reports are generally too sparse and unreliable to be used for verification; severe hail occurrences are underreported in rural areas and away from highways, as well as when a more severe weather event, such as a tornado, occurs nearby (Doswell et al. 2005; Witt et al. 1998b; Allen and Tippett 2015). For example, there are only 4 hail reports recorded by the SPC during the DA period (2000 – 2040 UTC) and all reports are in Denver. Radar observations during the same period (Fig. 3.1c) indicate multiple hail producing storms were present in the experiment domain. Further, hail sizes that correspond with familiar circular objects (e.g., quarters, golfballs) are over-reported by the public (Sammler 1993), skewing reported hail sizes. Radar derived hail products such as MESH (Witt et al. 1998a) and HCA output (Heinselman and Ryzhkov 2006; Park et al. 2009; Ryzhkov et al. 2013; Ortega et al. 2016; Putnam et al. 2017b) have been shown to be superior to surface based reports for hail size verification because they have a high spatial resolution and fewer systematic biases (Cintineo et al. 2012).

When working with unbiased observations, HCA output is shown to better discriminate

hail size than the MESH algorithm (Ortega et al. 2016). Subjective comparisons also demonstrate the HCA output closely resembles surface-based hail reports (when available) both in terms of hail size and the location of the report. Output from the Ortega et al. (2016) HCA (Fig. 3.1c) along with radar observed Z (Fig. 3.1b) is used to evaluate ensemble analyses and forecasts during the DA period.

In this study, the HCA is applied to NEXRAD radar observations from both Denver (KFTG) and Cheyenne (KCYS); these are the two closest radars to the hail producing storms. Radar observations are interpolated to the 500-m model grid for verification, with a 9-point smoother applied to Z_{dr} and ρ_{hv} to reduce noise. Environmental information is obtained from the ensemble forecasts. HCA output is considered within a 120 km radius of a given radar, this is the maximum distance Ortega et al. (2016) uses to diagnose surface hail size. HCA output from KFTG and KCYS is merged by taking the maximum detected hail size at each grid point and applying a smoothing filter that decreases hail size detections by one size bin (e.g., significant severe to severe) when the majority of grid points within a 1 km radius are smaller than the given HCA detection (Labriola et al. 2019a).

NWP models do not explicitly predict surface hail size, instead hail PSDs are frequently used to diagnose the maximum observable hail size at a grid point (e.g., Snook et al. 2016; Labriola et al. 2017, 2019a,b; Luo et al. 2018; Gagne et al. 2019). This study uses the Snook et al. (2016) maximum hail size definition, which defines the maximum hail size (D_{max}) as the largest diameter for which the PSD predicts at least 1 hailstone in a 100 m × 100 m box located in the lowest meter of the atmosphere ($N_{tmin} = 10^{-4}$ m^{-4}). This criteria is similar to minimum number concentration thresholds defined for hail in previous studies (e.g., Milbrandt and Yau 2006; Gagne et al. 2019). Although the minimum number concentration threshold can be modified, the average D_{max} value

does not substantially change (approximately 1 – 2 mm) when the threshold is increased or decreased by an order of magnitude. The spatial coverage of analysis estimated D_{max} at the lowest model height (~25 m AGL) is compared subjectively and objectively to the HCA output (Fig. 3.1c).

3.3 Results

3.3.1 Evaluating ensemble forecast and analysis innovations

The root-mean square innovation (RMSI) and ensemble spread of Z and V_r (Fig. 3.3) quantitatively evaluate EnKF analyses and forecasts. The RMSI is defined as:

$$\text{RMSI} = \sqrt{\langle d^2 \rangle} \tag{1}$$

where $\langle d^2 \rangle$ is the mean squared innovation or difference between the observation and model mapped to observation space and averaged across the ensemble. d is defined as:

$$\text{d} = \text{y}^{\text{o}} - \overline{H(x)} \tag{2}$$

where y^o is the observation, H is the forward operator that maps the model state vector to observation space, and x is either the model state forecast or analysis vector, respectively. Innovations are averaged over the KFTG volume in regions where either the observed or ensemble mean simulated Z exceeds 15 dBZ. This criterion includes both observed precipitation and spurious echoes, but eliminates potentially large regions of clear air and light precipitation from the statistics (Snook et al. 2011; Jung et al. 2012).

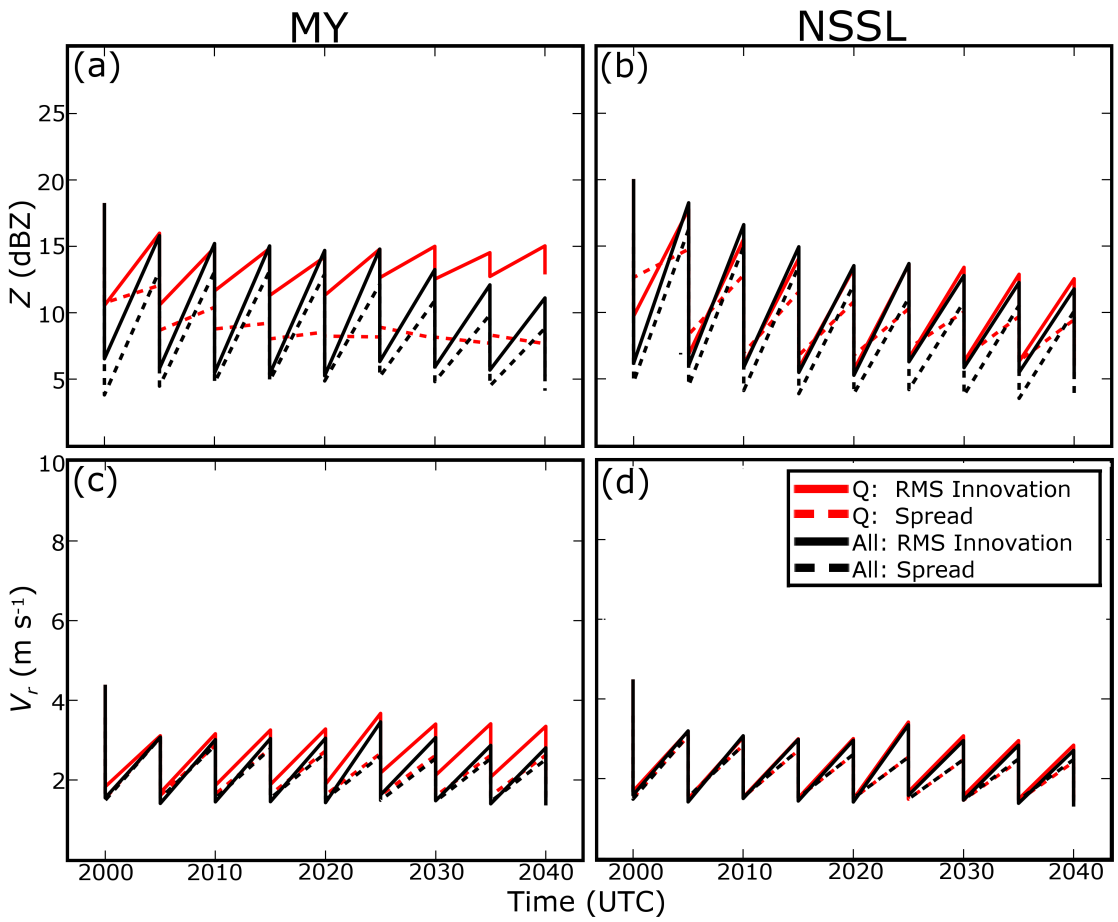

Fig. 3.3. The ensemble RMSI (solid lines) and spread (dotted lines) for Z (a,b) and V_r (c,d). The performance of ensembles MY-Q (a,c) and NSSL-Q (b,d) are marked with red lines, MY-ALL (a,c) and NSSL-ALL (b,d) are marked with black lines. Statistics are calculated over the experiment domain from 2000 to 2040 UTC, calculations are limited to locations where the observed and/or model (ensemble mean) Z exceeds 15 dBZ.

Although innovations are relatively similar between NSSL-Q and NSSL-ALL (Fig. 3.3b,d), RMSIs differ substantially between MY-Q and MY-ALL for Z (Fig. 3.3a). The MY-Q and MY-ALL RMSIs for V_r are relatively similar (Fig. 3.3c); this is partly because only winds in the precipitation regions are included in the statistics while differences in the outflow regions are not included. MY-Q RMSIs for Z exhibit only a small reduction during DA cycling, suggesting updating only one PSD parameter for multi-moment schemes may lead to an unreasonable PSD in MY-Q (Fig. 3.3a). RMSIs for MY-Q and MY-ALL V_r are more similar than RMSIs for Z and

decrease with time during DA. Wind speed is indirectly influenced by microphysical state variables through thermodynamic feedbacks, and therefore unlike Z, is less affected by updates to total number concentrations during assimilation. When only parts of PSD variables are updated, the RMSI for MY-Q V_r does not change much with cycles (Fig. 3.3c). This behavior is not observed in NSSL-Q for reasons that will be discussed later in this section.

Analysis mean simulated Z (Fig. 3.4) and KFTG observations (Fig. 3.1b) are subjectively compared at the time of the final DA cycle (2040 UTC) to identify why the RMSI for MY-Q Z (Fig. 3.3a) is larger than any other ensemble (Fig. 3.3a,b). The ensemble analysis mean is analyzed instead of individual members because the mean is used to derive the RMSI. Although there are slight variations in simulated Z intensity, MY-ALL (Fig. 3.4b), NSSL-Q (Fig. 3.4c), and NSSL-ALL (Fig. 3.4d) analyses closely match observed Z (Fig. 3.1b). MY-Q (Fig. 3.4a) Z is somewhat different from observations (Fig. 3.1b); storm structure is less organized, and the MY-Q analysis frequently underestimates Z. Throughout the storms MY-Q ensemble spread is often small (< 5 dBZ), suggesting the EnKF gives less weight to observations during DA. Additionally, MY-Q predicts reflectivity to be spuriously intense in regions of relatively little observed precipitation, such as to the east of Denver (Fig. 3.4a). Poor MY-Q Z estimates that increase the RMSI for MY-Q Z (Fig. 3.3a), are due in part to unreliable error covariances that develop as a consequence of unbalanced updates of hydrometeor mixing ratios and number concentrations.

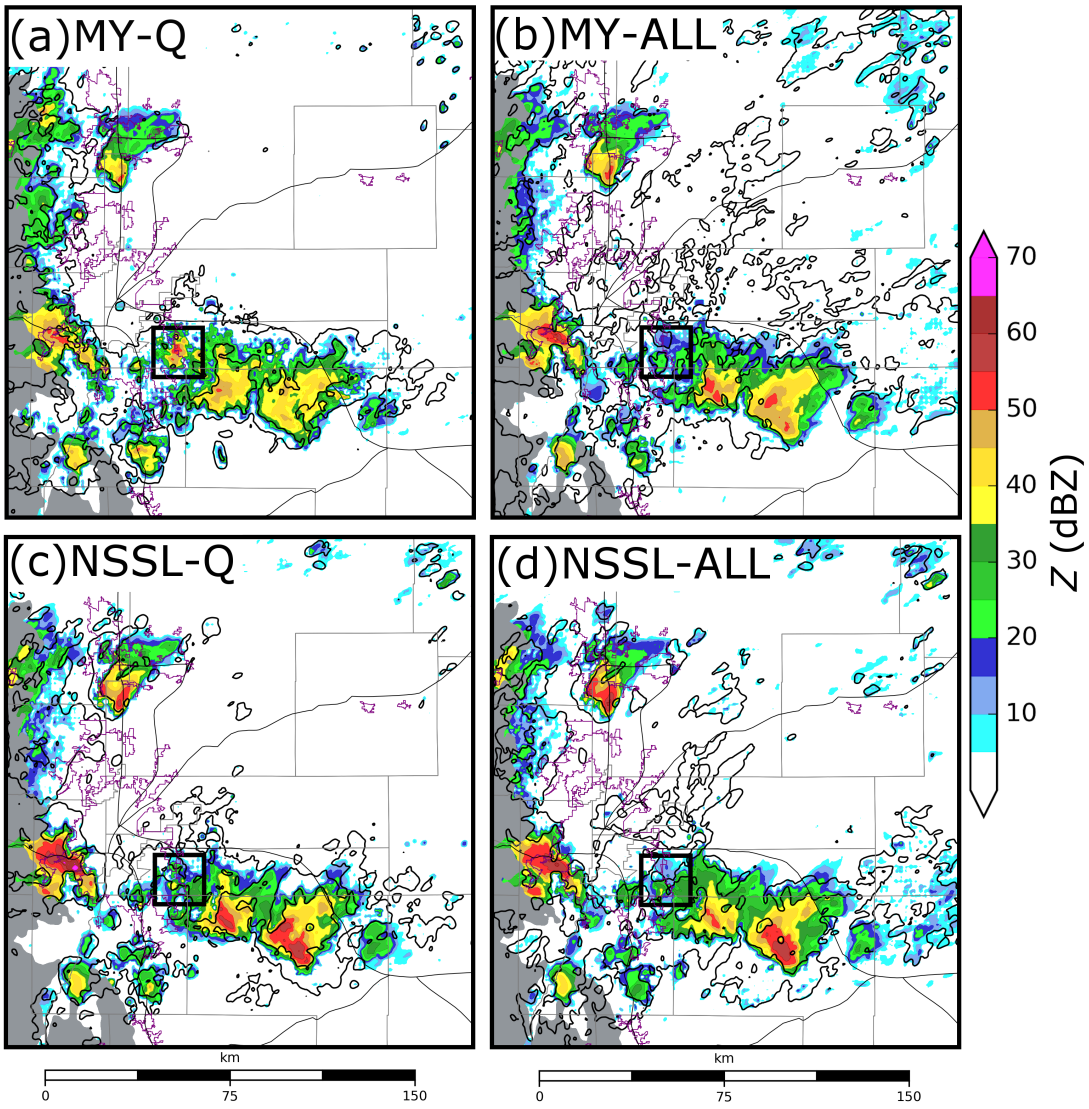

Fig. 3.4. Ensemble mean simulated *Z* interpolated to the lowest tilt (0.5°) of KFTG at the time of the final analysis (2040 UTC) for ensembles MY-Q (a), MY-ALL (b), NSSL-Q (c), and NSSL-ALL (d). 5 dBZ ensemble spread is contoured in black. A black square in each figure marks the subdomain analyzed in Fig. 3.6, the background map is the same as Fig. 3.1.

Despite predicting weaker *Z* values in the convective core of the observed storms, MY-Q predicts the total mass and number of rain (Fig. 3.5a, c) and hail (Fig. 3.5b, d) to be larger than MY-ALL. This is partly because MY-Q predicts more spurious *Z* throughout the domain during DA (Fig. 3.4a); however, the ensemble also predicts observed storms to produce more precipitation. NSSL-Q (Fig. 3.4c) and NSSL-ALL (Fig. 3.4d) predict more intense *Z* than either

MY ensemble (Fig. 3.4a, b) despite both experiments predicting substantially less hail in terms of both mass and number (Fig. 3.5b, d). This is because although the NSSL scheme predicts storms to produce relatively large hail aloft that contributes to high *Z* values. Results agree with previous studies (e.g., Johnson et al. 2016, 2019) which note the MY scheme predicts storms to produce a large number of small hailstones aloft while the NSSL scheme predicts storms to produce fewer but larger hailstones. Differences in simulated *Z* between the experiments demonstrate that *Z* is not a monotonic function of hydrometeor mixing ratio but is sensitive to changes in the hydrometeor particle size distribution.

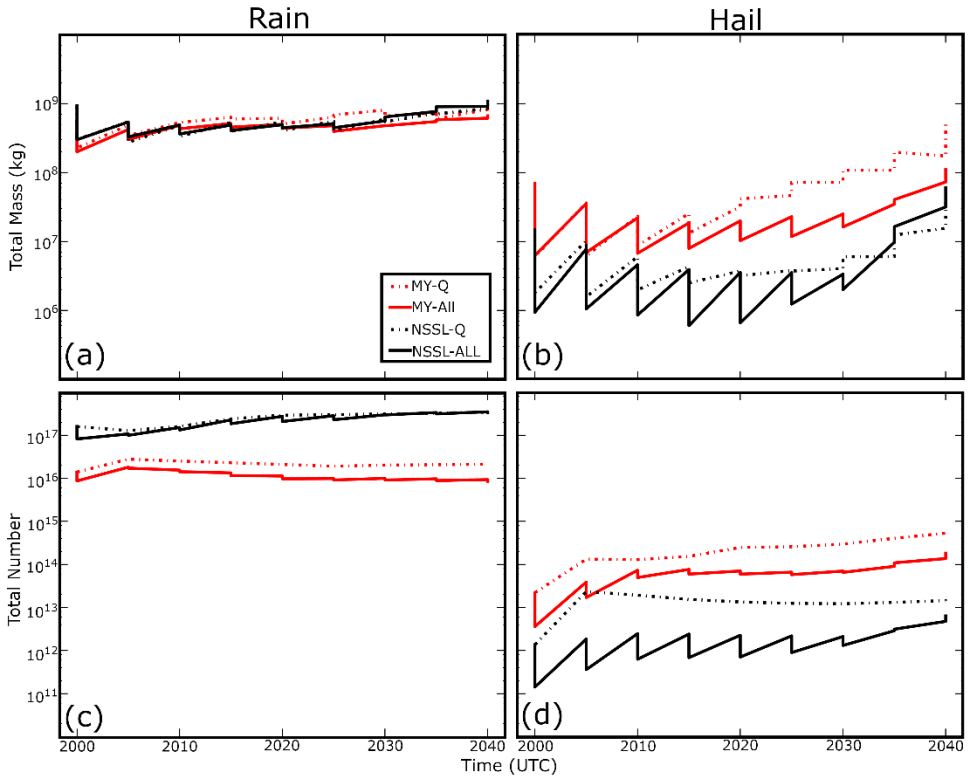

Fig. 3.5. Ensemble mean (forecast and analysis) total mass (a-b) and number (c-d) of rain (a, c) and hail (b, d) between 2000 – 2040 UTC. Hydrometeor concentrations are summed over a volume that spans the experiment subdomain.

For an EnKF system to accurately estimate model state variables that are not directly

observed, the filter must develop reliable multivariate covariances during the assimilation period. The cross-covariance between state variables and observation priors is used to retrieve unobserved variables and modify storm structure. Previous studies, such as TX05 and XJZ10, analyzed forecast cross-correlations between model state variables and an observation prior at an assumed observation location. Such correlation analyses demonstrate the expected impact of assimilating that observation on analyzed model state variables; a positive (negative) correlation between state variables and observation prior suggests the EnKF will adjust both variables in the same (opposite) direction. For instance, if the *Z* observation prior and vertical velocity are positively correlated the EnKF will strengthen the storm updraft (i.e., intensify the storm) when the *Z* prior is adjusted upward toward the observed Z value by the filter. Figure 1 of Snyder and Zhang (2003) illustrates this idea for radial velocity observation and vertical velocity."

Examples of an ensemble correlation analysis are provided in Fig. 3.6a-h. The ensemble correlation analysis is performed over a subdomain containing spuriously strong *Z* values in MY-Q, highlighted in Fig. 3.4. The assumed *Z* observation prior (sampled at the location of the white star) and hydrometeor mixing ratios are interpolated to the lowest radar tilt of KFTG (0.5°) from the closest model grid points both above and below the radar beam prior to calculating correlation; this allows for comparison between simulated (Fig. 3.6j-m) and observed *Z* (Fig. 3.6i). In previous OSSEs (e.g., Zhang et al. 2004, Caya et al. 2005, TX05), analyzed state variable estimates are often poor during the first several DA cycles because the error covariance between model state variables and observation priors is unreliable. Ensemble covariances are typically examined when the state estimation and ensemble covariance become reasonably reliable in later cycles (e.g., TX05). In this study, all correlation analyses are performed on the ensemble forecasts prior to the final DA at 2040 UTC when multivariate ensemble covariances can be reasonably good (Fig. 3.2).

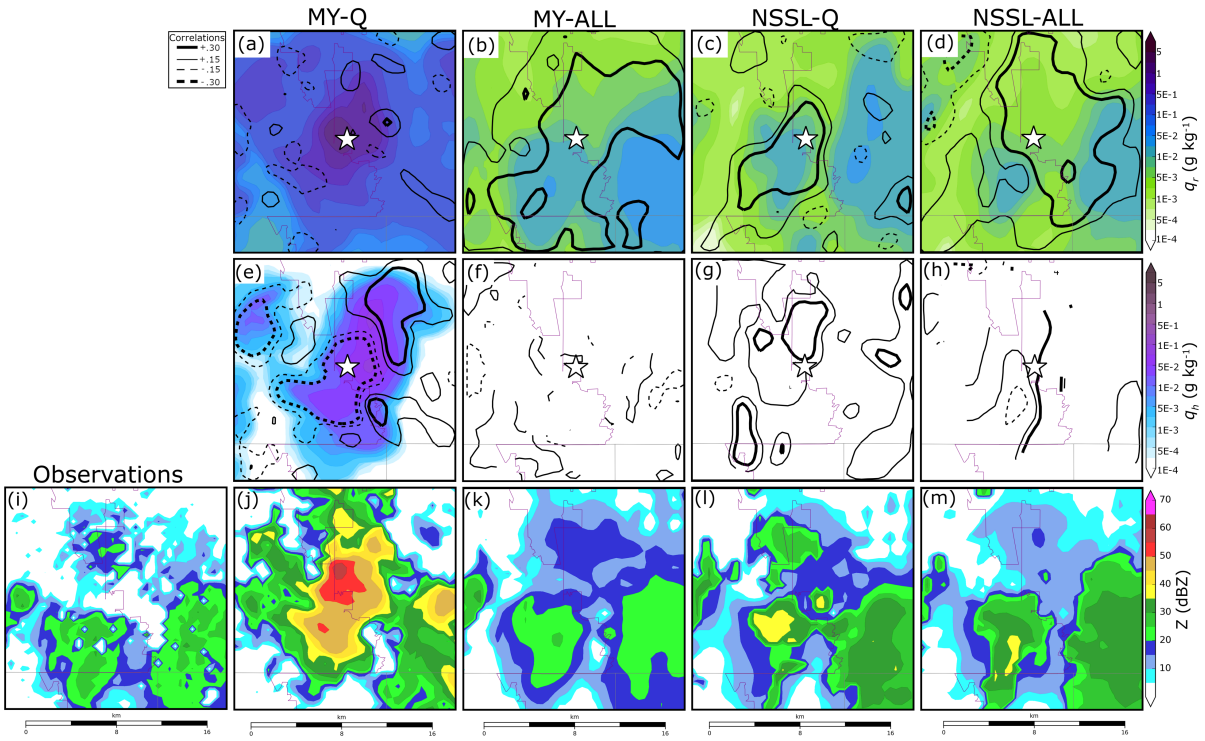

Fig. 3.6. Forecast error correlations for ensembles MY-Q (a,e), MY-ALL (b,f), NSSL-Q (c,g), and NSSL-ALL (d,h). The location of the subdomain is shown in Fig. 3.4. Correlations are calculated between the mass of rain q_r (a – d) or mass hail q_h (e -h) and an assumed Z observation (white star) interpolated to the lowest tilt of (0.5°) KFTG prior to the final assimilation cycle at 2040 UTC. Positive correlations (solid lines) and negative correlations (dashed lines) are plotted in increments of 0.15 between -0.3 and 0.3. Stronger correlations (0.3 and -0.3) are contoured with a thick black line, weaker correlations (0.15, -0.15) are contoured with a thin black line. To reduce noise, the correlation field is smoothed using a 9-point filter. Color shading represents ensemble forecast mean q_r (a-d) or q_h (e -h). Observed Z (i) and ensemble analysis mean Z for ensembles MY-Q (j), MY-ALL (k), NSSL-Q (l), NSSL-ALL (m) are provided at the same time and location.

MY-Q predicts high Z values where no organized convection is observed (Fig. 3.4a); such spurious precipitation develops within the model during the DA cycles near 2030 UTC and is not effectively suppressed by the filter in this case. A correlation analysis is conducted to understand why the MY-Q EnKF intensifies Z in a spurious storm to the east of Denver. Although MY-Q analyses predict the storm to produce large Z values, the updraft is weak and disorganized which remains unchanged during DA. In the vicinity of the spurious storm, Z and q_r simulated in MY-Q are relatively uncorrelated (Fig. 3.6a), consequently the EnKF is unable to decrease q_r in the

spurious storm when assimilating Z observations. In the same location MY-Q Z and q_h are negatively correlated (Fig. 3.6e), this relationship intensifies reflectivity in the spurious storm during DA (Fig. 3.6j) because the EnKF increases q_h when decreasing spuriously large model diagnosed Z. While experiment MY-Q exhibits the weakest correlations between Z and q_r (Fig. 3.6a), the other ensembles exhibit mostly strong positive correlations throughout much of the subdomain (>0.3) (Fig. 3.6b-d). Differences in the forecast error covariance structure between MY-Q (Fig. 3.6a) and MY-ALL (Fig. 3.6b) suggest the importance of correctly updating all model state variables. MY-ALL, NSSL-Q, and NSSL-ALL analyses predict weaker Z values that better fit observations in the subdomain (Fig. 3.6k-m). This is in part because the background better fits observations (i.e., no spurious storm prior to assimilation) and the background error covariance produces a more optimal analysis.

Updates to microphysical state variables, such as q_r, also impact thermodynamic state variables. For example, enhanced evaporational cooling attributed to an increase in precipitation during data assimilation (Fig. 3.5a,b) may cause MY-Q to predict more intense cold pools (Fig. 3.7a,b) and stronger (Fig. 3.7c,d) winds than MY-ALL. Winds modified by spurious storms also increase the RMSI for MY-Q V_r (Fig. 3.7c) more than for MY-ALL V_r (Fig. 3.7d).

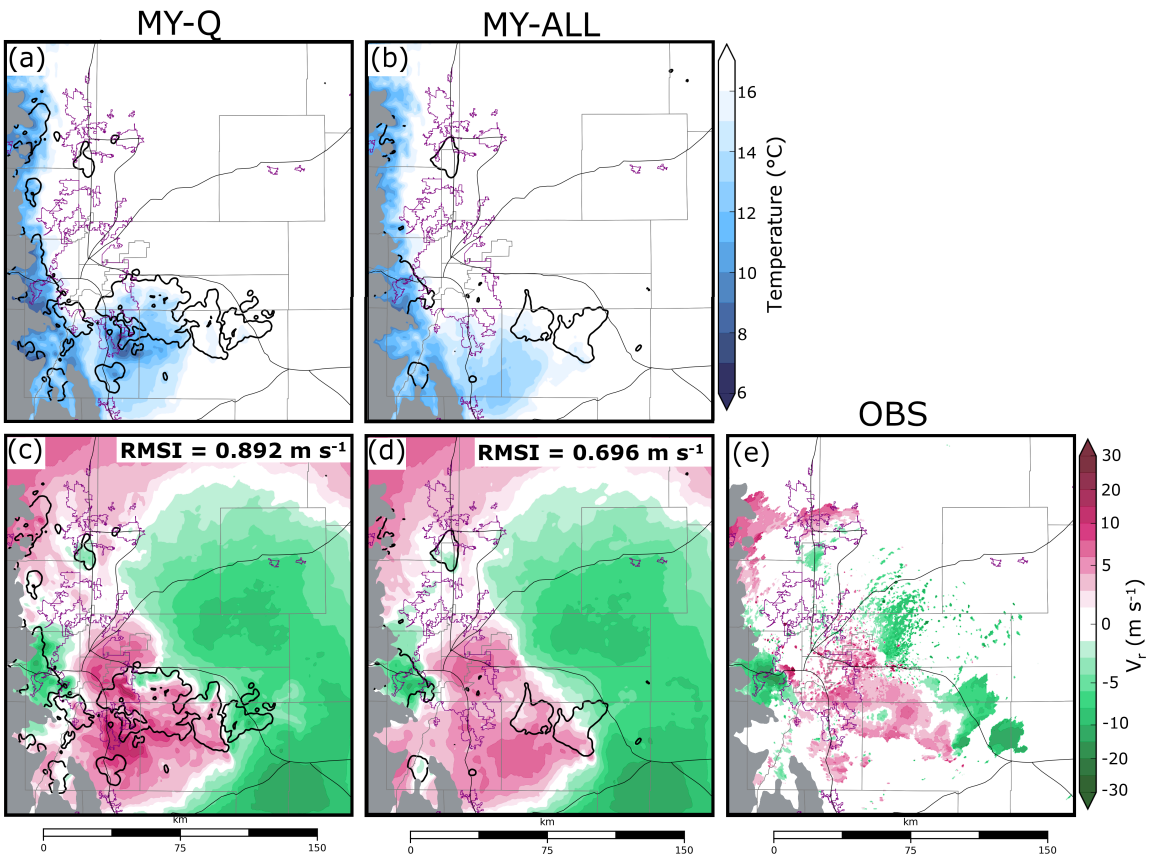

Fig. 3.7. Ensemble mean forecast temperature (a-b) at the lowest model grid level above the surface and radial velocity (c-d) at the lowest KFTG grid tilt for MY-Q (a,c) and MY-ALL (b,d) prior to the final assimilation cycle at 2040 UTC. Observed KFTG radial velocity (e) at the same time is provided. Thick black contours represent the 30 dBZ ensemble forecast mean Z smoothed using a 9-point filter. Background maps are the same as Fig. 3.1.

3.3.2 Evaluating surface hail size

For an analysis to skillfully estimate surface hail size, microphysical variables must be properly updated so that the analyzed hail PSD produces reasonable hail size estimates. Box-and-whisker plots (Fig. 3.8) compare the ensemble member analysis estimated areal surface coverage of non-severe, severe, and significant severe hail to HCA output (Fig. 3.1c). Additional subjective comparisons are performed by comparing the probability-matched mean (Ebert 2001) of the ensemble analysis D_{max} (Fig. 3.9) to HCA output (Fig. 3.1c). Unlike a simple ensemble mean, which typically smooths out extreme values (i.e, large hail sizes), the probability-matched mean

maintains the spatial distribution of the mean and the frequency distribution of the ensemble to preserve extreme values and create a single best-guess forecast.

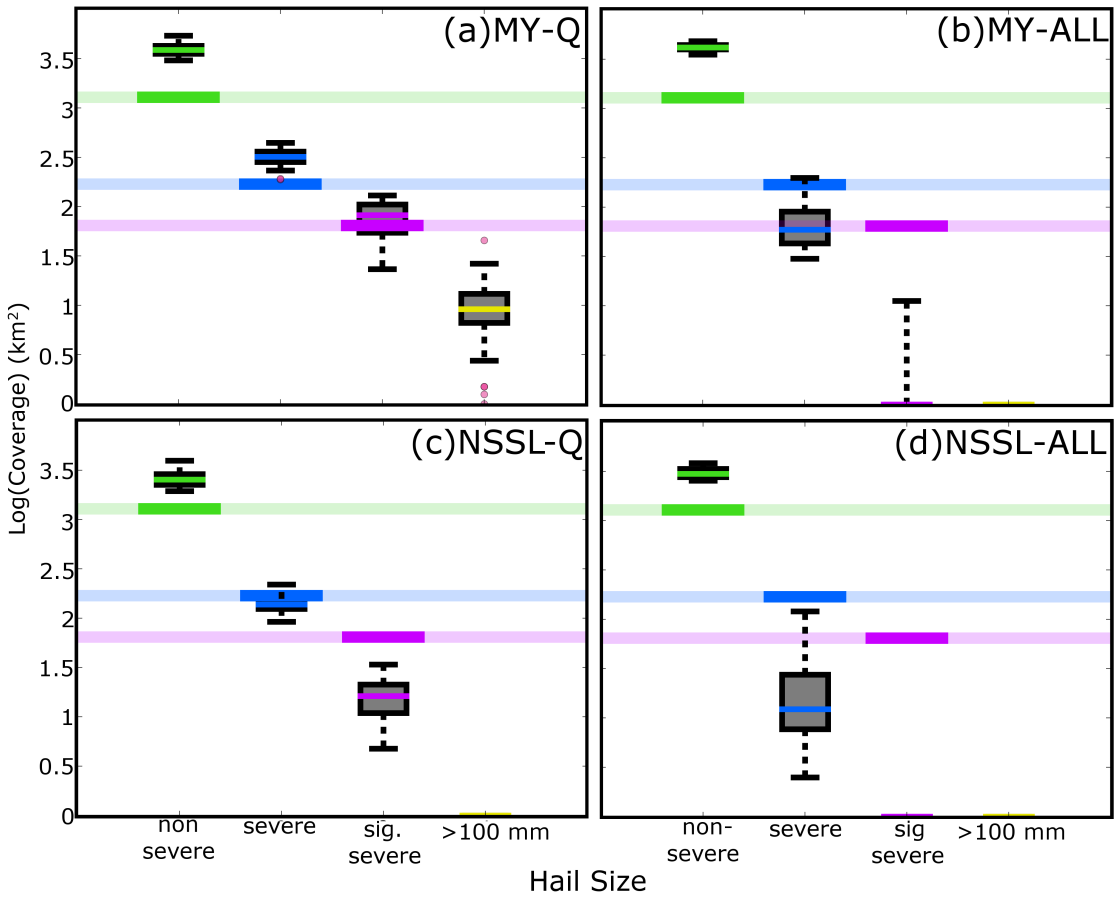

Fig. 3.8. Analysis estimated coverage of non-severe (green), severe (blue), and significant severe (purple) hail according to D_{max} between 2000 – 2040 UTC for ensembles MY-Q (a), MY-ALL (b), NSSL-Q (c), and NSSL-ALL (d). Hail coverage is only considered within the domain shown in Fig. 3.1 and where land elevation is less than 2.5 km above mean sea level. Actual coverage of non-severe, severe, and significant severe hail based upon HCA output (Fig. 3.1c) is marked with a thick horizontal line of the corresponding color. Although the largest observed hailstone on this day was approximately 70 mm, the coverage of D_{max} > 100 mm (yellow) is included to identify hail size overestimation biases.

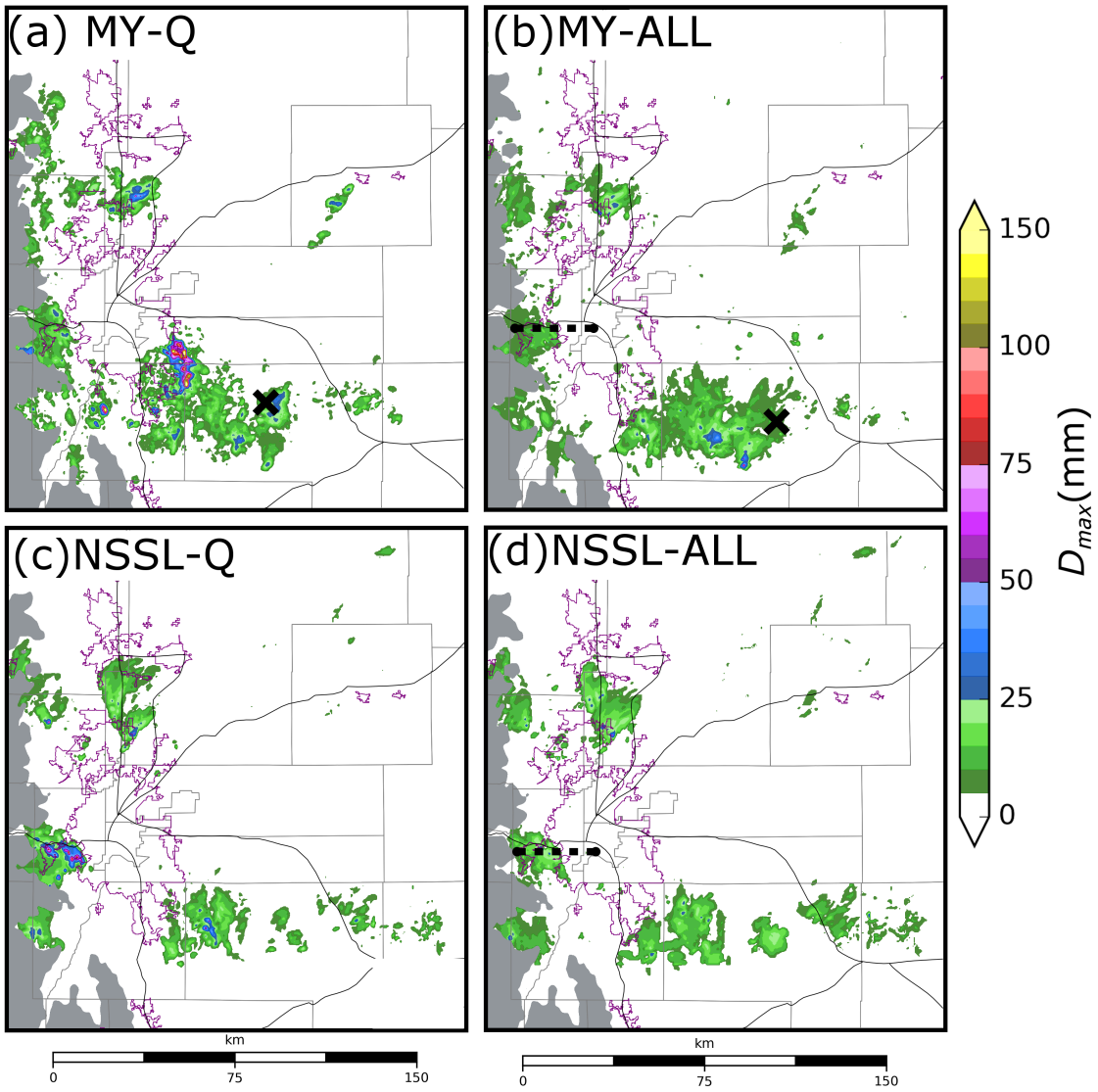

Fig. 3.9. The probability match mean of D_{max} for ensemble analyses MY-Q (a), MY-ALL (b), NSSL-Q (c), and NSSL-ALL (d) between 2000 – 2040 UTC. A black "x" in (a) and (b) marks the location where hail PSDs are sampled in Fig. 3.10. Horizontal dashed lines in (b) and (d) mark the locations where vertical cross-sections are taken in Figs. 3.11 – 3.15.

All ensembles overestimate the spatial coverage of non-severe hail (Fig. 3.8); in addition, MY-Q overestimates the coverage of severe and significant severe hail (Fig. 3.8a). In the probability-matched mean of the ensemble analyses, D_{max} in MY-Q (Fig. 3.9a) predicts the largest hail (D_{max} > 200 mm) in spurious convection located to the south and east of Denver. HCA output

is unable to identify the maximum hail size because all hail greater than 50 mm in diameter is classified as "significant severe"; however, the largest hail report recorded by the SPC for this event is 70 mm, suggesting that MY-Q substantially overestimates surface hail size. Further, almost all MY-Q estimated severe and significant severe hail coverage occurs in spurious convection away from where HCA output indicates large hail to occur (Fig. 3.1c). Although NSSL-Q underestimates the coverage of significant severe hail (Fig. 3.8c), the analyses (Fig. 3.9c) exhibit some qualitative skill and much of the largest hail occurs in regions where significant severe hail is detected in HCA output (Fig. 3.1c). MY-ALL (Fig. 3.8b) and NSSL-ALL (Fig. 3.8d) underestimate the spatial coverage of severe and significant severe hail, and their probability-matched means (Fig. 3.9b,d) suggest the analyses rarely indicate hail exceeding 25 mm in diameter.

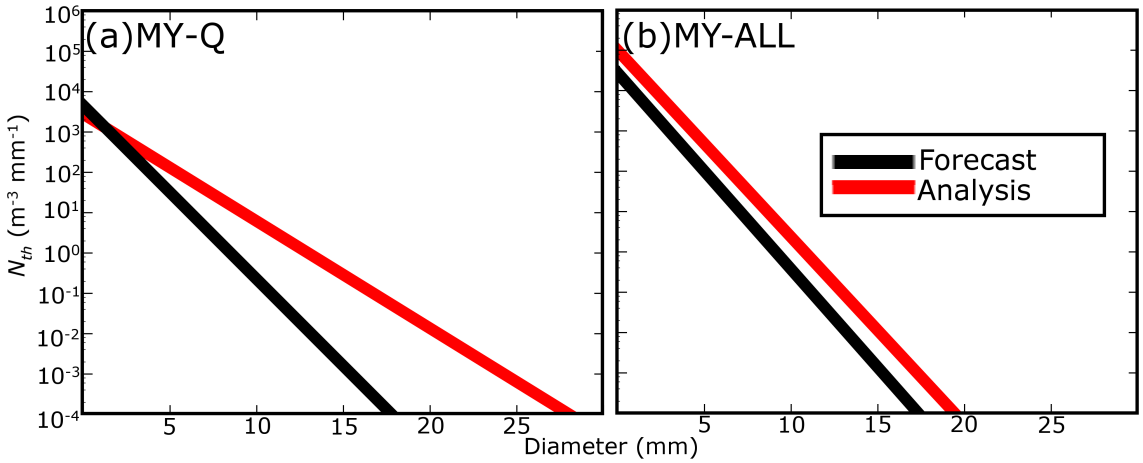

Fig. 3.10. Hail PSDs diagnosed from model level 15 (1.96 km above ground level) of MY-Q (a) and MY-ALL (b) ensemble mean forecasts and analyses at 2040 UTC. The grid point where the MY-Q and MY-ALL hail PSDs are diagnosed is marked with an "X" in Fig. 3.9a and Fig. 3.9b, respectively. The hail variables used to diagnose the PSDs are provided in Table 3.1. PSDs were selected to highlight examples of how the EnKF updates a hail PSD during assimilation.

	MY-Q Forecast	MY-Q Analysis	MY-ALL Forecast	MY-ALL Analysis
q_h (g kg^{-1})	1.64×10^{-2}	6.94×10^{-2}	6.02×10^{-2}	2.68×10^{-1}
N_{th} (m^{-3})	4.55	4.55	24.19	92.32
D_{max} (mm)	17	28	17	19

Table 3.2. MY-Q and MY-ALL forecast and analysis mean q_h , N_{th}, and D_{max} at 2040 UTC. The locations where the MY-Q and MY-ALL hail variables are sampled is marked with an "X" in Fig. 3.9a and Fig. 3.9b, respectively. Hail state variables diagnose the hail PSDs shown in Fig. 3.10.

Hail size changes as the DA system modifies q_h and N_{th}. The MY-Q filter, which does not modify N_{th}, frequently estimates large hailstones (Fig. 3.9a), primarily because the EnKF makes large adjustments to q_h in regions where relatively little hail is present. When large quantities of added ice mass are shared between a relatively small number of hailstones, the average hailstone diameter greatly increases. To demonstrate this behavior a PSD is sampled from a hailstorm southeast of Denver ("x" in Fig. 3.9a) using the MY-Q ensemble mean both before and after assimilation at 2140 UTC. At the MY-Q sampled grid point q_h increases by 5.3×10^{-2} g kg^{-1} (Table 3.1) during DA. Because q_h increases but N_{th} remains constant, the slope of the analysis hail PSD becomes more shallow than the forecast (Fig. 3.10a) and D_{max} increases by approximately 11 mm (Table 3.1). In extreme instances DA causes D_{max} to exceed 200 mm in MY-Q (Fig. 3.9a). NSSL-Q estimates hail sizes (Fig. 3.9c) that more closely resemble HCA output (Fig. 3.1c). A more reliable multivariate covariance, due in part to fewer spurious storms in the NSSL-Q background forecasts compared to the MY-Q background forecasts (not pictured), prevents large updates to q_h and limits the most extreme hail sizes.

Hail PSDs sampled from the MY-ALL ensemble forecast and analysis mean at 2140 UTC in a hailstorm to southeast of Denver ("x" in Fig. 3.9b) demonstrate how the EnKF adjusts hail

size through changes to both q_h and N_{th} (Fig. 3.10b). During DA, the MY-ALL EnKF increases q_h by approximately 2.08×10^{-1} g kg^{-1} and nearly triples N_{th} (Table 3.1). Although q_h in MY-ALL increases more than in MY-Q (Table 3.1), D_{max} remains relatively unchanged because of the increase in N_{th}. The slope of the MY-ALL analysis hail PSD is similar to that of the forecast PSD (Fig. 3.10b) and subsequently D_{max} increases by only 2 mm (Table 3.1). For this analyzed grid point (Fig. 3.10b) D_{max} is classified as non-severe, despite the HCA output (Fig. 3.1c) detecting significant severe hail in the surrounding region. Microphysical variables, such as N_{th}, need additional constraint to accurately estimate hail size; previous work by XJZ10 suggests assimilating additional data improves initial condition estimates by further constraining microphysical variables.

Although MY-ALL and NSSL-ALL analyses estimate microphysical properties, such as Z with a moderate to high level of skill (Fig. 3.3a,b), there are an insufficient number of observations to constrain all microphysical variables associated with multi-moment microphysics schemes and, as a result, the performance of hail size forecast is limited. To the west of Denver where the HCA detects significant severe hail (Fig. 3.1c), MY-ALL (Fig. 3.9b) and NSSL-ALL (Fig. 3.9d) analyses underestimate hail size in part because the EnKF modifies moments of the hail PSD so that the slope of the PSD remains relatively constant (e.g., Fig. 3.10b). This causes hail mass to be split amongst more hailstones, limiting the ability to increase the mean hailstone diameter.

3.3.3 Ensemble correlation analysis

Differences in the microphysical parameterizations cause hydrometeor properties (e.g., size, density, fall speed) to differ substantially between MP schemes. For example, Johnson et al. (2016) note that hail behavior differs between the MY and NSSL schemes primarily due to

differences in hail production processes. Typically, the MY scheme produces many small hail particles aloft because the MY hail category is primarily composed of small, frozen raindrops that have been converted into hail particles (Johnson et al. 2016). The NSSL scheme produces fewer and larger hailstones because hail is produced from dense graupel that has undergone wet growth (Mansell et al. 2010). Simulated Z is intrinsically related to such microphysical assumptions (e.g., Jung et al. 2008b, 2010a, 2012; Dawson et al. 2014; Putnam et al. 2014, 2017b, Johnson et al. 2016, 2018); these assumptions also determine the forecast error covariance (TX05, XJZ10).

The ensemble correlation between observation prior Z and model state variables (i.e., w, θ, q_h, N_{th}, q_r, N_{tr}) illustrates how different hail treatments in MP schemes impact the analysis increments. Ensemble correlation analyses are performed on the hailstorm located to the west of Denver over Interstate-70 (Fig. 3.1) prior to the final DA cycle at 2040 UTC. Vertical cross-sections are taken through the updraft of the hailstorm (locations shown in Fig. 3.9b, d) and coincide with where most hail growth processes occur. The correlation analysis is performed in the vertical to capture the variation of hail particle behavior both above and below the 0 °C isotherm.

When an assumed observation is taken from the ice-phase dominant region of the MY-ALL updraft, w and Z become negatively correlated (Fig. 3.11a). This correlation pattern was first observed in XJZ10 and was attributed to microphysical assumptions made in the MY scheme. Stronger updrafts loft more rain and cloud water aloft, leading to a positive correlation between w and rainwater throughout much of the updraft above the 0 °C isotherm (i.e., supercooled water) (Fig. 3.12b,d). Supercooled raindrops are converted into hail within the MY scheme, hence updraft intensity is also positively correlated with N_{th} (Fig. 3.12e) throughout much of the storm. Because raindrops are relatively small, the mean mass diameter of hail decreases in strong updrafts (e.g.,

Johnson et al. 2016, XJZ10), and thus Z is negatively correlated with w in MY-ALL forecasts. Weaker updrafts advect less supercooled liquid above the 0 °C isotherm and lower the production rate of small hailstones; additionally larger hailstones aloft fall through the weaker updrafts to enhance Z.

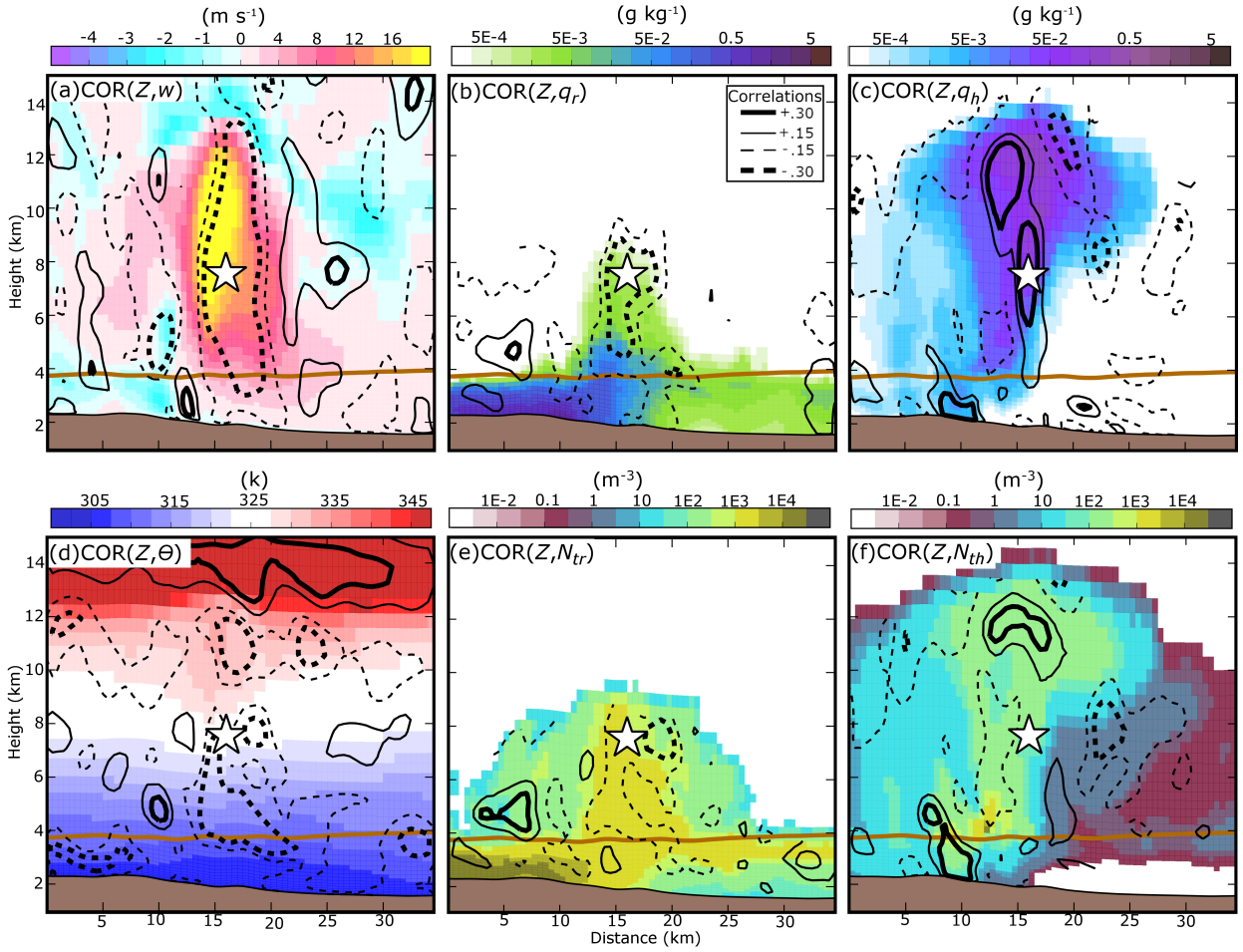

Fig. 3.11. Vertical cross-sections of MY-ALL ensemble forecast mean w (a), q_r (b), q_h (c), θ (d), N_{tr} (e), and N_{th} (f) prior to the final assimilation cycle at 2040 UTC. Cross-sections are taken through the Denver hailstorm, denoted as east-west line in Fig. 3.9b. Forecast error correlations between an assumed Z observation taken from within the hail growth zone (white star) and state variables are plotted. The same plotting convention is used to contour correlation as in Fig. 3.6. The horizontal brown line represents the 0 °C isotherm.

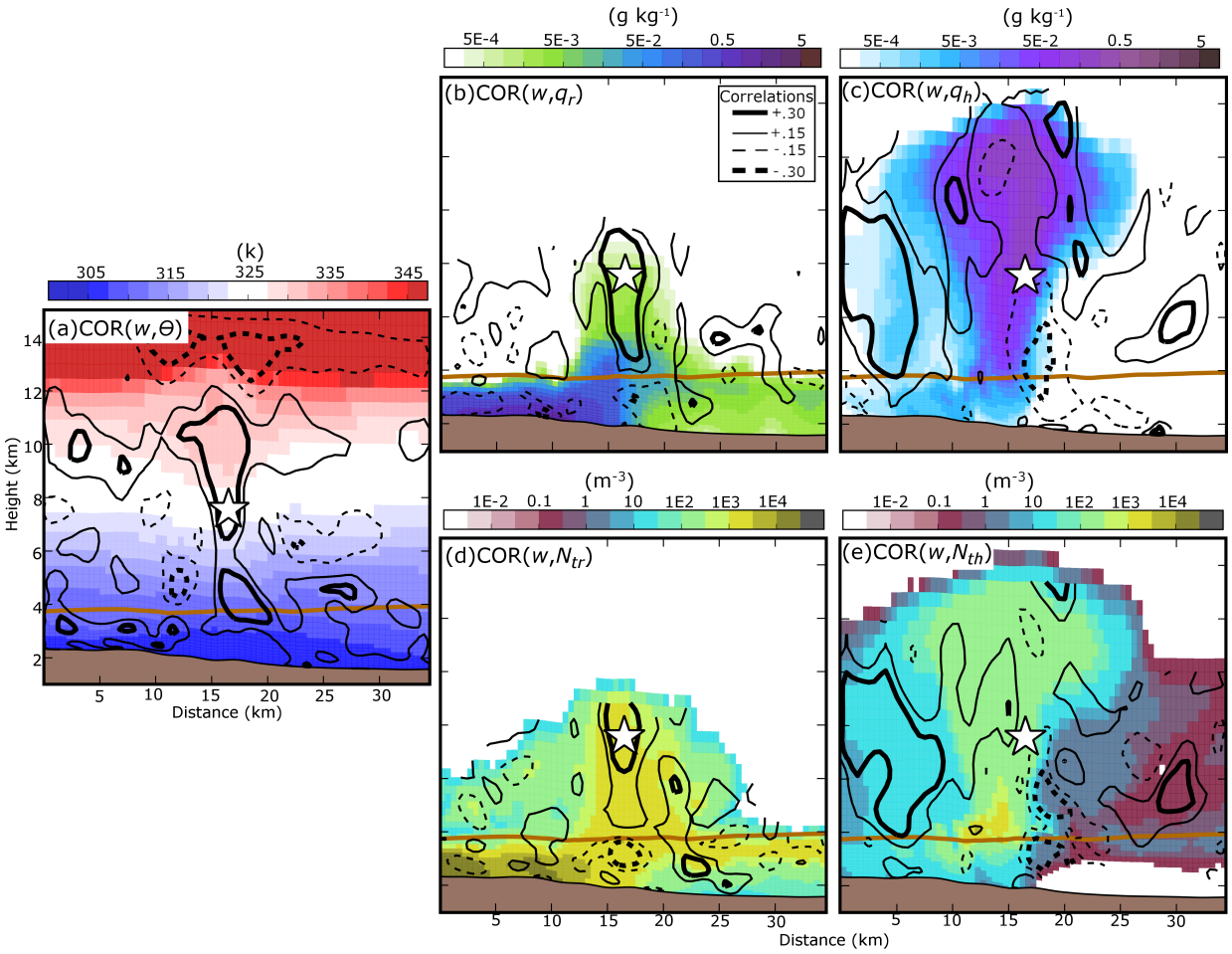

Fig. 3.12. Same as Fig. 3.11, but for θ (a), q_r (b), q_h (c), N_{tr} (d), and N_{th} (e). Plotted correlations are between an assumed w observation in the hail growth zone and the plotted model state variables.

MY-ALL correlations determine how model state variables are modified during reflectivity assimilation. Similar to XJZ10, Z at the observation location is negatively correlated with θ in the updraft region (Fig. 3.11d) because w and θ are positively correlated in the updraft region (Fig. 3.12a). Hydrometeor state variables: q_r (Fig. 3.11b), N_{tr} (Fig. 3.11e), and N_{th} (Fig. 3.11f) are negatively correlated with Z between 4 – 8 km above MSL; this is where frozen raindrops are converted to hail. Unlike with N_{th}, Z exhibits mostly positive correlations with q_h (Fig. 3.11c) above the 0 °C isotherm, suggesting the complex relationship between Z and the hail size spectrum.

Unlike MY-ALL, Z at the observation location is positively correlated with w (Fig. 3.13a)

throughout much of the NSSL-ALL storm updraft; however similar to MY-ALL, w is positively correlated with both rain and hail state variables (Fig. 3.14b-e). Strong updrafts in the NSSL-ALL storm loft liquid water above the 0°C isotherm. Instead of being converted into small hail particles, the supercooled liquid is accreted by graupel and hail, and causes rimed ice particles to increase in size (Labriola et al. 2019a). It is also noted that Z and θ are positively correlated throughout the updraft in the NSSL-ALL storm (Fig. 3.13d) due to the same processes in MY-ALL.

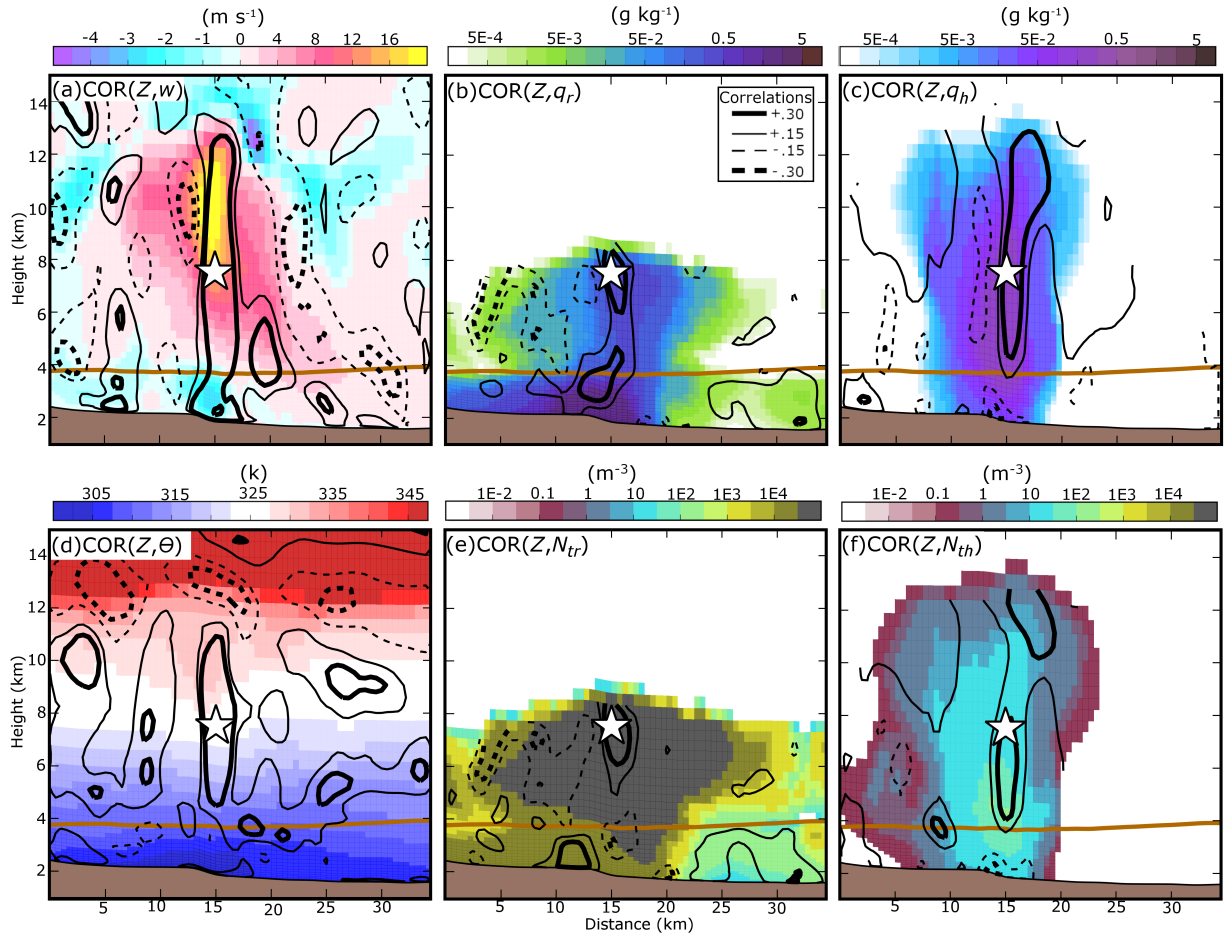

Fig. 3.13. Same as Fig. 3.11, except for NSSL-ALL w (a), q_r (b), q_h (c), θ (d), N_{tr} (e), N_{th} (f). Plotted correlations are between an assumed Z observation in the hail growth zone and the plotted model state variables. Cross-sections are taken through the Denver hailstorm, denoted as east-west line in Fig. 3.9d.

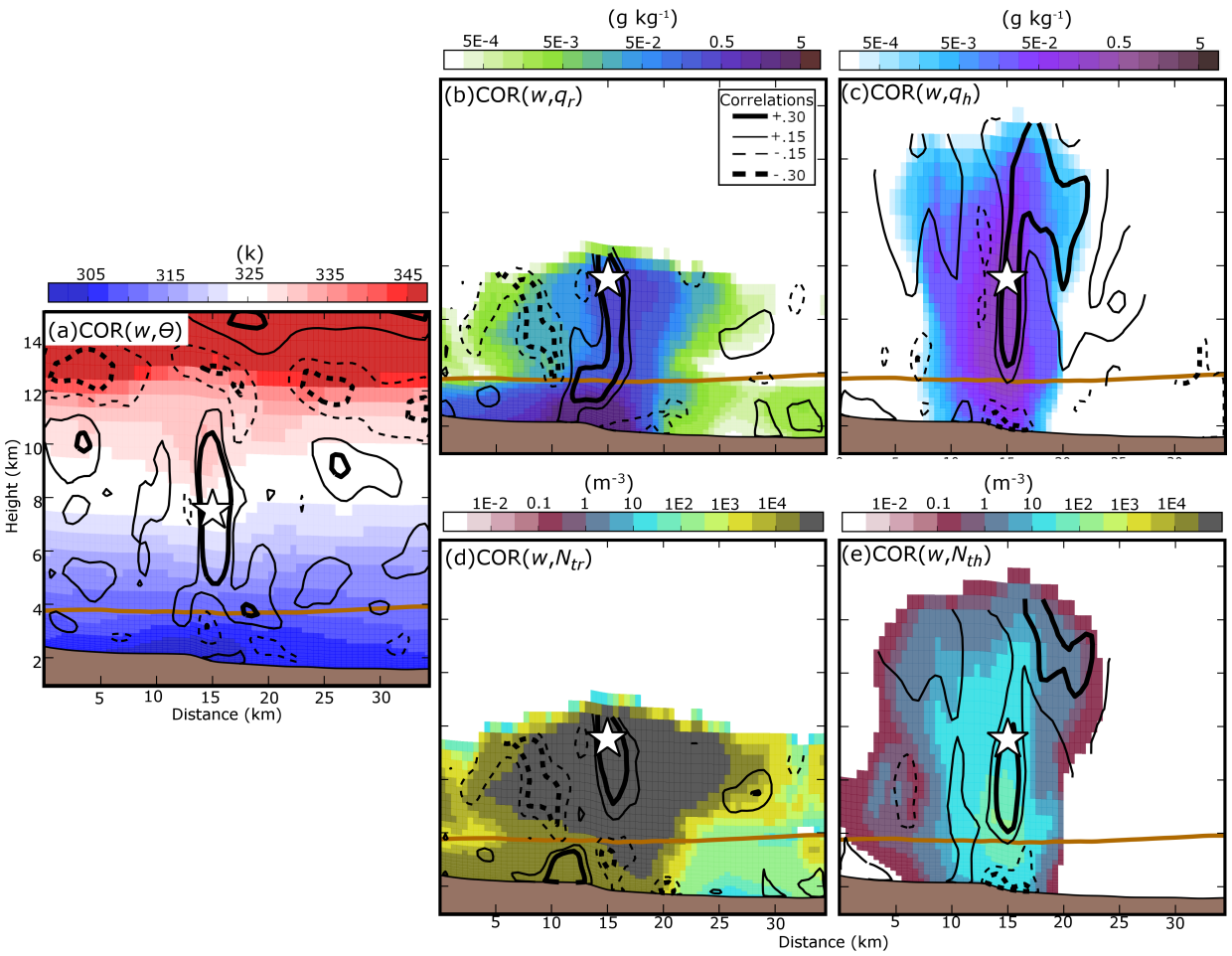

Fig. 3.14. Same as Fig. 3.13, except for θ (a), q_r (b), q_h (c), N_{tr} (d), N_{th} (e). Plotted correlations are between an assumed w observation in the hail growth zone and the plotted model state variables.

Prognostic graupel and hail volume (v_g and v_h, respectively) equations allow the NSSL scheme to vary rimed ice density during the forecast and modify both particle sedimentation and hail production processes. In the storm updraft region, Z and v_h are positively correlated (Fig. 3.15a). Hail accreting more liquid water in the stronger updraft increases hail volume and Z in the updraft (Fig. 3.15a, b). It is noted that hail density is a non-unique solution, and that an infinite combination of q_h, v_h pairs can produce the same density value; correlations are therefore not indicative of modifications to hail density.

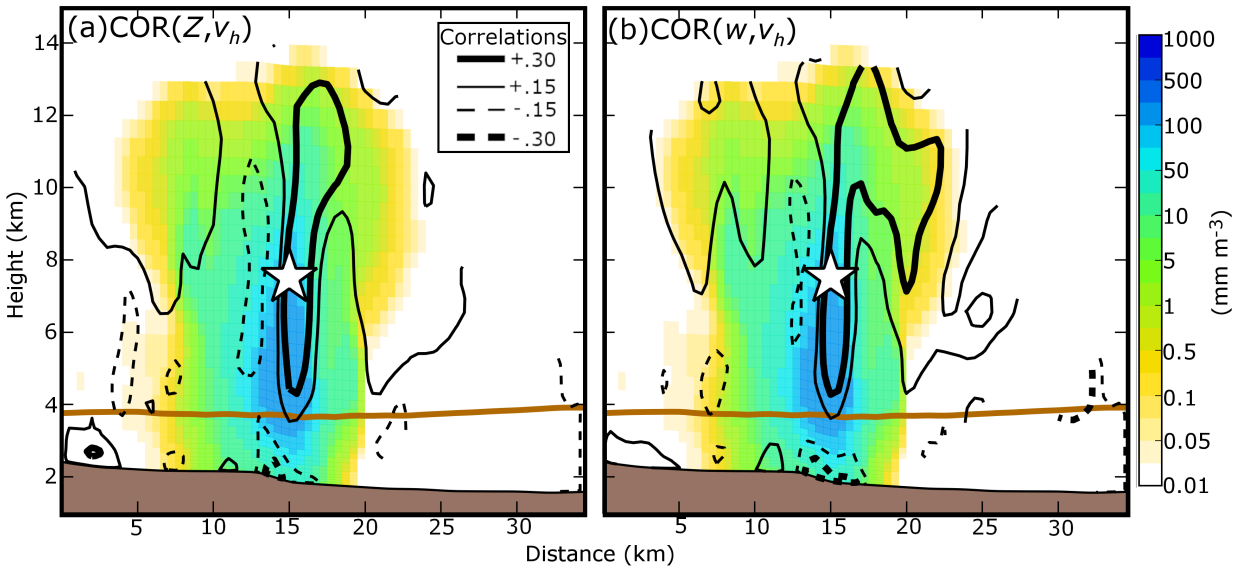

Fig. 3.15. Same as Fig. 3.13, but for v_h. Plotted correlations are between v_h and an assumed Z observation (a) or w observation (b) in the hail growth zone.

While this study primarily evaluates the ensemble background error correlation fields within strong thunderstorms at 2040 UTC, it is noted the background error covariance evolves with time through the DA cycles. Correlations between observation prior Z and w taken from within the updraft of a storm to the east of Denver are evaluated throughout the DA window (Fig. 3.16). This storm is selected because it is present in observations throughout the DA window. Prior to the first DA cycle both MY-ALL and NSSL-ALL are not confident in the location of organized convection, which causes correlations between w and Z to be noisy (Fig. 3.17a,d). Both ensembles predict storm updrafts to be weak ($w < 15$ m s^{-1}) at 2020 UTC (Fig. 3.17b,e). Because of that, the MY scheme predicts the storm to produce little hail. Instead, the weak updrafts loft precipitation above the 0 ºC isotherm and enhance Z. w and Z become negatively correlated with time as the storm intensifies and strong updrafts increase the number of small hail stones aloft; these correlation fields are similar to those observed in the Denver hail storm (Fig. 3.11a). Unlike MY-ALL, w and Z are positively correlated in NSSL-ALL between 2020 and 2040 UTC (Fig. 3.17f,i)

because the updrafts advect more precipitation above the 0 ºC isotherm and enhance the production of large hailstones. The evolving correlation fields demonstrate the sensitivity of the background error covariance to the microphysical assumptions made within MP schemes as well as other factors including the quality of the state estimation. In this paper, we focus more on the later time (2040 UTC) when the state estimation becomes better.

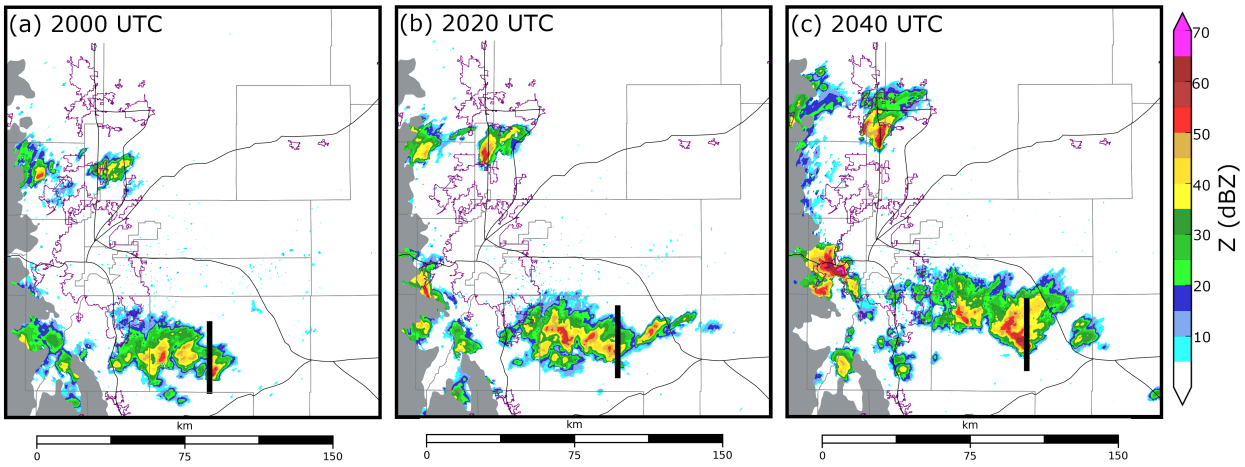

Fig. 3.16. KFTG observed Z at the lowest radar tilt (0.5º). Thick black lines denote the approximate location of where vertical cross-sections are taken in Fig. 3.17.

Correlations between model state variables (i.e., w and θ) and Z have important consequences on the representation of analyzed storms. The correlation patterns observed in this study were also observed in other strong storms at 2040 UTC (e.g., Fig. 3.17c, f) and during earlier DA cycles (not shown). When MY-ALL underestimates (overestimates) Z, the filter will decrease (increase) w (Fig. 3.11a) and make the updraft air temperature cooler (warmer) (Fig. 3.11d). In effect the EnKF in MY-ALL weakens the analyzed storm. Both w (Fig. 3.13a) and θ (Fig. 3.13d) are positively correlated with Z in NSSL-ALL and cause the filter to intensify the storm under similar circumstances. Z is generally a function of the diameter of a hailstone to the sixth power although it becomes more complex for a large or wet hailstone due to the Mie scattering effect. In

addition, numerous observational studies (e.g., Heymsfield 1983; Nelson 1983; Ziegler et al. 1983; Foote 1984; Dennis and Kumjian 2017) have suggested that hail size is not a monotonic function of updraft strength, but is also influenced by vertical wind shear, environmental moisture, and updraft volume. More observations are needed to determine which analysis increment that more closely reflects reality.

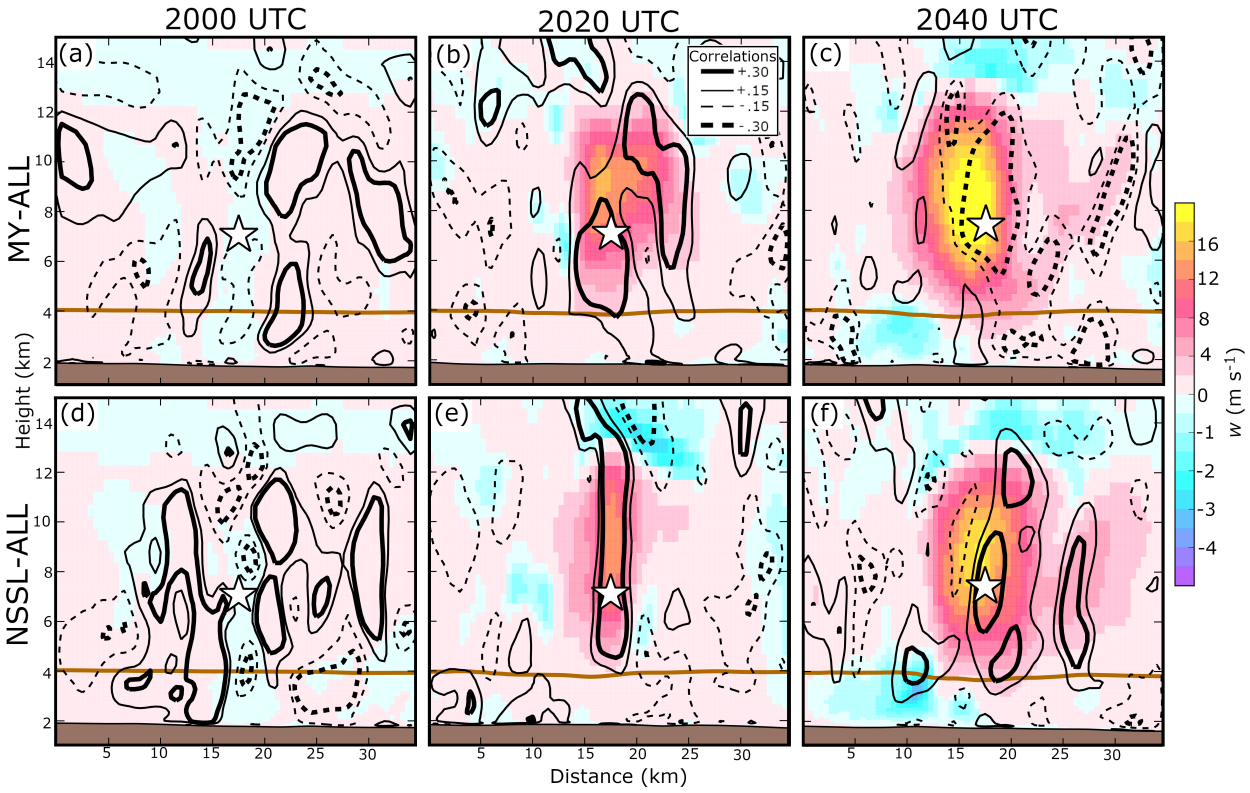

Fig. 3.17. Vertical cross-section of the MY-ALL (a-c) and NSSL-ALL (d-f) ensemble mean *w* at (a, d) 2000 (b, e) 2020, and (c, f) 2040 UTC. Cross-sections are taken from within a storm to the east of Denver. Forecast error correlations between an assumed *Z* observation taken from within the hail growth zone (white star) and *w* are plotted. The same plotting convention is used to contour correlation as in Fig. 3.6. The horizontal brown line represents the 0 °C isotherm.

3.4 Summary and further discussion

Newer multi-moment bulk microphysics schemes (MP schemes) are increasing in complexity and predicting more state variables in order to improve the representation of

microphysical processes (e.g. riming, sedimentation). As the number of degrees of freedom within a MP scheme increases, initial state estimation using an EnKF becomes more challenging, in part because the large number of microphysical state variables predicted by a MP scheme are insufficiently constrained by the limited number observations that can infer microphysical properties, e.g., radar reflectivity (Z).

In this study a cycled EnKF framework is used to update the microphysical properties of hail-producing storms for the 8 May 2017 Colorado severe hail event. Four ensemble forecast experiments are conducted using either the Milbrandt and Yau (2005a) double-moment (MY) scheme or the NSSL double-moment variable density rimed ice scheme (Mansell et al. 2010). An EnKF is used to update either only hydrometeor mixing ratios (MY-Q, NSSL-Q) or all microphysical state variables (MY-ALL, NSSL-ALL); in addition to dynamic and thermodynamic information. While most previous studies use observed Z to update all microphysical variables, this study examines whether updating a limited number of particle size distribution moments (i.e., mass mixing ratio) provides sufficient constraint for an ensemble system run using a multi-moment scheme. The model forecasts and analyses are evaluated against observed Z in addition to output from a hydrometeor classification algorithm (HCA) to determine which configuration produces the more realistic state variable estimates related to hail.

For ensembles that update only hydrometeor mixing ratio, in particular MY-Q, the forecast error covariance is often unreliable and limits the accuracy of state variable estimates. We suspect that the forecast error covariance is unreliable because updating a limited number of microphysical variables introduces large imbalances into the ensemble prediction system during assimilation. For example, negative correlations between Z and hydrometeor mixing ratios in experiment MY-Q cause spurious radar echoes to be enhanced during assimilation, even if clear-air reflectivity data

are assimilated. Further, due to the poor multivariate error covariance structure, the EnKF in MY-Q was unable to replicate Z intensity or structure within mature storms. Generally, the forecast error covariance of MY-ALL and the NSSL ensembles (NSSL-Q and NSSL-ALL) is more reliable than that of MY-Q, allowing the ensemble analyses to estimate Z and radial velocity (V_r) with less error.

Comparison (and verification) of analyzed surface hail size against HCA output provides insight into how the different EnKF configurations update hail particle size distributions. Generally, ensembles that update only mixing ratio (MY-Q, NSSL-Q) estimate large surface hail sizes because large quantities of hail mass can be distributed amongst a relatively small number of hailstones. Ensembles that update all microphysical variables (MY-ALL and NSSL-ALL) tend to underestimate surface hail size because these ensembles typically predict larger hail number concentrations and cause hail mass to be split between many small hailstones. Results suggest the need to assimilate additional observations of microphysical relevance in order to better constrain the increased number of state variables (e.g., number concentration).

Hail production and growth assumptions made by MP schemes substantially influence the forecast error covariance. The MY scheme generates hail primarily from small frozen raindrops; this process increases the number of small hailstones above the 0 °C isotherm and causes Z to be negatively correlated with air temperature and updraft strength. Due to differences in hail production processes, the opposite correlation patterns are observed in the NSSL scheme forecasts for intense hail storms. Hail growth assumptions also influence correlation patterns for hydrometeor state variables. For this study the NSSL-ALL experiment favors more positive correlations between Z and hail variables (mass and number) than the MY-ALL experiment. While this study provides insight into the complexities of updating microphysical variables via an EnKF,

it is noted that only a limited number of the available schemes are evaluated. A large number of multi-moment schemes (e.g., Thompson et al. 2008; Morrison et al. 2005, 2009; Lim and Hong 2010; Morrison and Milbrandt 2015; Morrison et al. 2015; Milbrandt and Morrison 2016) are used in weather prediction systems; however, relatively few studies have analyzed how underlying microphysical assumptions made within these schemes impact multivariate ensemble background error covariances and state variable updates within ensemble DA. We have also shown that the multivariate ensemble covariances can be sensitive to the quality of storm analysis and possibly also the storm intensity and morphology.

Due to the many possible combinations of variables that are able to produce a given Z value, microphysical state variables will remain insufficiently constrained by observed reflectivity. The assimilation of polarimetric observations is shown to provide additional constraint in observation system simulation experiments (OSSEs) conducted by Jung et al. (2008b, 2010b) and a real case study conducted by Putnam et al. (2019), however assimilating polarimetric variables remains non-trivial. Although MP schemes can replicate basic polarimetric signatures (e.g., Johnson et al. 2016, 2019; Putnam et al. 2017a, b, 2019), the NWP output from these schemes is often biased in intensity and coverage. To benefit from the assimilation of polarimetric observations, more effort is needed to improve the representation of microphysical processes that generate polarimetric signatures and to find the optimal configurations that can maximize the impact of polarimetric data. Improving the representation of microphysical processes has the potential to not only improve EnKF estimates, but also mitigate misrepresentations of sub-grid scale processes and reduce model errors.

Chapter 4 Evaluating Sensitivity to the GSI EnKF Data Assimilation Configuration and Forecast Performance with Experiments for the 28 – 29 May 2017 Mesoscale Convective System Case

4.1 Introduction

Severe storms have caused approximately $200 billion in property damage (NOAA 2019) since the beginning of the millennium. Most severe weather hazards occur with only tens of minutes of advance warning (average tornado warning lead time ≈ 17 minutes) because the National Weather Service issues warnings based upon the detection of a hazard, either from surface reports, radar detections, or an imminent threat as determined by the forecaster (Stensrud et al. 2013). Based on the current warning paradigm (i.e., warn-on-detection), the average warning lead time for detected tornadoes has remained relatively constant since 1986 (Stensrud et al. 2013) and is unlikely to increase without degrading warning skill (e.g., increasing the number of false alarms) (Brooks 2004). To extend warning lead time, the NWS is working on adopting the Warn-on-Forecast paradigm (Stensrud et al. 2009, 2013) where frequently updated, convection-allowing model (CAM) ensemble forecast guidance is incorporated into the warning process. Due to the rapid error growth of convective scale forecasts, in addition to limiting sources of forecast model error, one key component of a skilled forecast system is to limit initial condition errors by developing a skilled data assimilation (DA) system.

Various Warn-on-Forecast prototype systems (e.g., Yussouf et al. 2013; Wheatley et al. 2015; Jones et al. 2016; Snook et al. 2016; Johnson et al. 2017; Lawson et al. 2018; Labriola et al. 2019a; Stratman et al. 2020) use an EnKF (Evensen 1994, 2003) to assimilate observations and initialize CAM forecasts. This technique is particularly attractive at the convective scale because flow-dependent error covariance statistics derived from the forecast ensemble allow the filter to

update unobserved model state variables using often a limited number of observed parameters. This is the case with radar observations where typically only radar reflectivity (Z) and radial velocity (V_r) are available at the convective scale. Many observation system simulation experiments (OSSEs) and real-data experiments have demonstrated that assimilating Z and V_r observations improves estimated thermodynamic, dynamic, and microphysical state variables (e.g., Snyder and Zhang 2003; Zhang et al. 2004; Dowell et al. 2004; Caya et al. 2005; Tong and Xue 2005; Xue et al. 2010; Dawson et al. 2012; Jung et al. 2012; Johnson et al. 2015; Snook et al. 2015; Supinie et al. 2016; Wang and Wang 2017; Putnam et al. 2019). While EnKF systems have demonstrated their potential, performance of the DA system and skill of subsequent forecasts is contingent upon the optimization of the DA configurations. Each DA system contains many tuning parameters and further, the optimization is sensitive to the model configurations and observations assimilated (e.g., Dowell et al. 2004; Tong and Xue 2005; Sobash and Stensrud 2013; Johnson and Wang 2017).

Previous studies (e.g., Zhang et al. 2004; Dowell and Wicker 2009; Johnson and Wang 2017) have found ensemble forecasts and DA system configurations can have a substantial impact on the subsequent forecast skill. The remainder of this introduction discusses many commonly used techniques that improve the skill of EnKF initial condition estimates.

EnKF systems tend to underestimate the ensemble error covariance derived from the ensemble forecasts (e.g., Houtekamer and Mitchell 1998). This is partly because of the ensemble size is much smaller than the degrees of freedom of the system and partly because model errors and other sources of uncertainties remain unaccounted (e.g., Romine et al. 2014). Various methods are employed to better account for the sources of uncertainties in the ensemble so that the ensemble spread better represents the forecast error. The ensemble forecasting system should be initialized

with suitable initial perturbations while lateral boundary condition perturbations should be included in limited area ensembles. Their setup and configurations often require tuning. To initialize the ensemble when starting EnKF DA cycling, random perturbations with or without spatial correlations are often used for idealized storm simulations (e.g., Snyder and Zhang 2003; Dowell et al. 2004; Tong and Xue 2005) while real-data experiments often start from perturbations derived from an operational ensemble (e.g, Kong et al. 2009) or a combination of the above methods are used (e.g., Snook et al. 2016). Due to the sensitivity of CAM forecasts to model physics (e.g., Putnam et al. 2014; Cohen et al. 2015; Morrison et al. 2015; Johnson et al. 2016; Cohen et al. 2017; Johnson et al. 2019; Labriola et al. 2019b), varying ensemble member physics (e.g., microphysics, planetary boundary layer, surface) can increase ensemble variance even more than initial condition perturbations (e.g., Stensrud et al. 2000; Gallus and Bresch 2006). When properly configured, multi-physics ensembles (e.g., Schwartz et al. 2010; Duda et al. 2014; Johnson and Wang 2017) exhibit improved diversity and skill when predicting convection (Johnson and Wang 2017), but can also produce larger biases than a single-physics ensemble (Romine et al. 2013). Even with these spread enhancement techniques, ensembles tend to remain spread deficient and additional treatments are required during DA to further increase variance.

Underestimating the background error variance under-realizes the influence of observations during DA. Repeated assimilation of dense observations can reduce ensemble spread so much that the EnKF can no longer effectively assimilate observations because it becomes over-confident about the ensemble forecast, leading to filter divergence (Jazwinski 1970; Anderson and Anderson 1999). Spread maintenance algorithms artificially increase ensemble variance to better represent the true uncertainty of the atmospheric state during DA (Anderson 2001; Hamill et al. 2001). Some spread inflation algorithms add noise to the posterior ensemble (e.g., Dowell and

Wicker 2009; Sobash and Wicker 2015) while some, such as the relaxation-to-prior-spread (RTPS) algorithm (Whitaker and Hamill 2012), increase the spread of posterior ensemble to a fraction of the spread of the prior ensemble. All methods are met with varying degrees of success between experiments, and often a combination of these methods is used to optimize EnKF performance (Jung et al. 2012).

The quality of background error covariance is often poor because it is computed from an insufficient number of ensemble members. Such sampling errors cause grid points to become spuriously correlated with distant observations, and consequently the filter erroneously updates the background forecasts. A distance-based Gaussian weighting function (Gaspari and Cohn 1999) is often used to localize the influence of observations; the optimal cutoff radius is contingent upon many factors including ensemble size, observation type, density, location, computational cost, model resolution, and correlation length scale from model dynamics (Sobash and Stensrud 2013; Ying et al. 2018). Most studies (e.g., Snyder and Zhang 2003; Zhang et al. 2004; Dowell et al. 2004; Caya et al. 2005; Tong and Xue 2005; Aksoy et al. 2009; Jung et al. 2012; Sobash and Stensrud 2013; Johnson et al. 2015; Wheatley et al. 2015) use a relatively small cutoff radius for radar observations (4 – 18 km horizontal, 4 – 8 km vertical) because they are relatively dense and the convective scale flows have small spatial correlation length scales.

Radar data require pre-processing prior to assimilation because observations are provided on radar polar coordinates at resolutions finer than most forecast models. Most EnKF experiments assimilate coarsened radar data that are interpolated to a regular grid or the model grid itself. For example, Xue et al. (2006) assimilates radar observations that are interpolated to the horizontal model grid but kept the radar elevation levels in the vertical. After interpolation, some experiments assimilate a fraction of the available radar observations to decrease the computational expense

(Gao and Xue 2008) or mitigate the effects of spatially correlated observation errors by removing neighboring observations (Chang et al. 2014). Thinning assimilated observations can also decrease the ensemble spread reduction during EnKF DA, though this does not necessarily improve forecast skill (Aksoy et al. 2012). Radar data are most commonly thinned at regular user-specified intervals throughout the model domain.. Despite potential benefits, thinning observations can remove important in-storm observational information during DA and limit the analysis skill.

Errors on the convective scale are poorly understood and thus represent a challenge when assimilating radar observations. While unrealistic, most EnKF-based studies assume radar observation errors are Gaussian in nature, constant in standard deviation, and spatially uncorrelated. Well-calibrated WSR-88D radars have observational errors of approximately 1 dBZ (Z) and 1 m s^{-1} (V_r) (Doviak and Zrnić 1993; Ryzhkov et al. 2005), but DA experiments typically assume larger errors to account for representativeness errors and other uncertainties (such as those with observation operators). Increasing observation errors also alleviates ensemble under-dispersion by decreasing spread reduction during DA and is shown to improve performance of ensemble analyses (e.g., Dowell et al. 2004; Snook et al. 2013).

To obtain accurate initial conditions for convective storms, radar observations are often assimilated at high frequencies. Real-time systems often assimilate radar observations every 15 minutes (e.g., Wheatley et al. 2015; Johnson et al. 2017; Snook et al. 2019), although improvements to computational infrastructure and new weather radar technology such as phased-array radars (Weber et al. 2007; Zrnic et al. 2007; Heinselman and Torres 2011; Curtis and Torres 2011) provide the opportunity to assimilate observations more frequently in the future. OSSEs that assimilate rapid-scanning radar information note a decrease the spin-up time for convection and a reduction in errors in both observed and unobserved variables (Zhang et al. 2004; Xue et al. 2006;

Yussouf and Stensrud 2010). Real-data cases have successfully assimilated radar information at relatively high frequencies (≤ 5 minutes) (Snook et al. 2011, 2016; Jung et al. 2012; Supinie et al. 2017; Labriola et al. 2017; Stratman et al. 2020); however, rapid updates can introduce imbalances that are unable to adjust to the model before the next DA cycle and can degrade both analysis and forecast skill (Wang et al. 2013; Johnson and Wang 2017) unless extra care is taken.

Most CAM studies evaluate forecast and analysis sensitivity to only one or two EnKF parameters. This study evaluates short-term (0 - 6 hour) forecasts for a mesoscale convective system (MCS) event on 28 – 29 May 2017 that are initialized using an EnKF system based on the NCEP operational gridpoint statistical interpolation (GSI) framework, using a DA configuration that is based upon the Center for Analysis and Prediction of Storms (CAPS) Storm Scale Ensemble Forecast (SSEF) run during the 2019 Hazardous Weather Testbed Spring Forecast Experiment (Clark et al. 2019). Experiments also investigate forecast sensitivity to many important aspects of the EnKF system design including the methods used to generate initial ensemble perturbations and enhance spread as well as covariance localization, data thinning, observation error, and DA frequency for radar observations. Knowledge gained from this study identifies which aspects of the EnKF configuration forecasts are most sensitive to and will guide the design of future data assimilation systems.

The rest of the chapter is organized as follows: section 4.2 provides a brief overview of the case study, a description of the experiments, and verification procedures. Experiment results including objective and subjective forecasts evaluations are discussed in section 4.3 and section 4.4 summarizes the results of this study and discusses potential future directions for research.

4.2 Event overview, experiment configuration, and verification methodology

4.2.1 Case overview

The focus of this study is to forecast the evolution of an MCS and nearby isolated convective storms that produced strong winds, hail, and tornadoes in Texas, Louisiana, and Mississippi on 28 – 29 May 2017 (Fig. 4.1). Upper-level wind patterns were favorable for convective development during this event; a trough was located over the Central United States and Texas, and was to the right of a jet entrance region. At approximately 2000 UTC on 28 May multiple thunderstorms initiated along a frontal boundary extending from Arkansas towards the Texas-Mexico border. Thunderstorms that initiated over eastern Texas quickly grew in scale between 2200 and 0000 UTC as they ingested unstable air (CAPE > 2200 j kg^{-1}). The storms eventually merged to form a squall line that impacted Louisiana and Mississippi between 0000 and 0600 UTC on 29 May (Fig. 4.2). Storms embedded within the line produced several weak tornadoes and multiple wind reports extending from Shreveport to Jackson (Fig. 4.1). Between 0000 and 0200 UTC isolated thunderstorms in Texas (Fig. 4.2a-b) produced multiple hail and wind reports (Fig. 4.1), but by 0300 UTC (Fig. 4.2c) many of these storms weakened and formed a large region of mostly stratiform precipitation. With the exception of convection near Houston and the MCS located east of Jackson, precipitation remains stratiform through 0600 UTC (Fig. 4.2d). The presence of active MCS convection within the DA period before the forecast initial condition time, and during several hours of forecast make this case suitable for investigating impact of DA and its configurations on subsequent storm forecasts. For forecasts, we will focus on the Texas, Louisiana, and Mississippi regions.

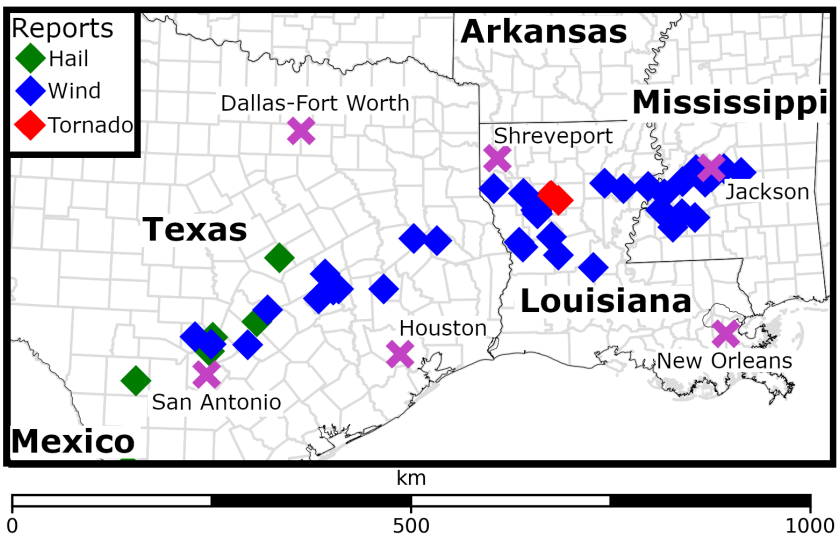

Fig. 4.1. A diagram of the 28-29 May 2017 forecast domain, states and countries are labeled in bold. A legend for hail, wind, and tornado SPC storm reports is provided in the upper left-hand corner. Cities referred to during the study are marked with a fuchsia "X".

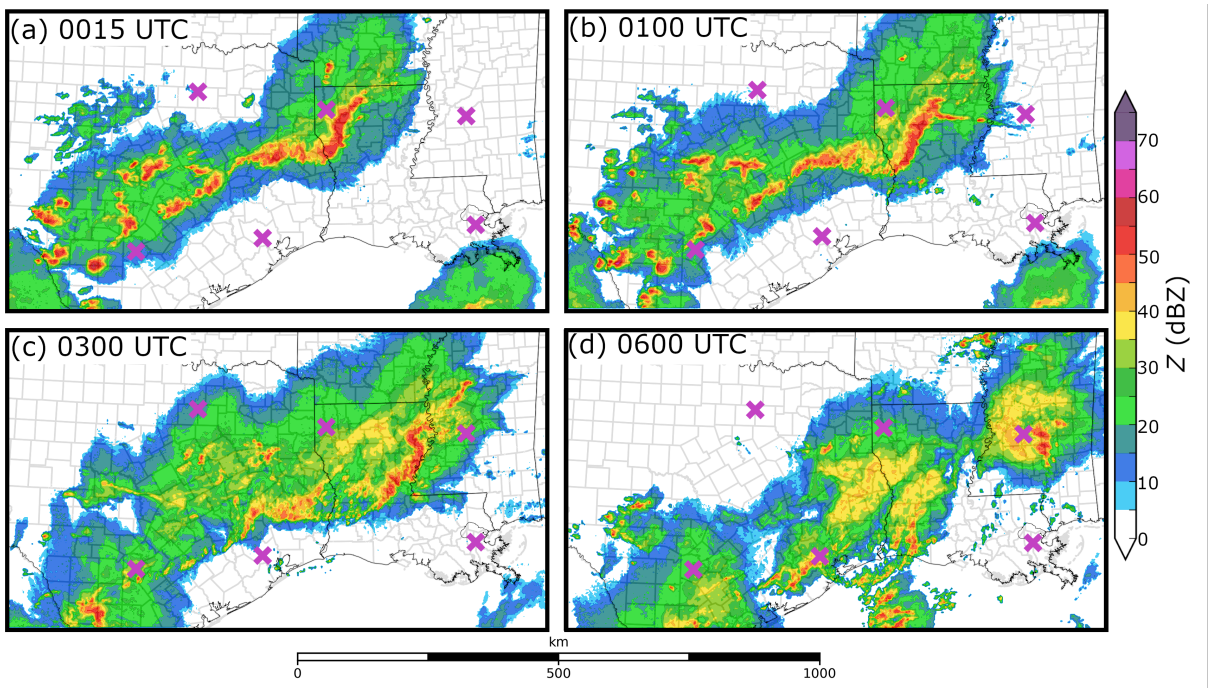

Fig. 4.2. MRMS observed column maximum *Z* valid at (a) 0015, (b) 0100, (c) 0300 UTC, and (d) 0600 UTC. Cities referred to during the study are marked with a fuchsia "X" and are the same as Fig. 4.1.

4.2.2 Control experiment DA system settings

This subsection describes the control experiment (hereafter referred to at CTRL) DA configuration (Table 1), modified DA configuration parameters are discussed in section 4.2.4. This study uses the GSI-based EnKF system for DA. The GSI system performs observation quality control (QC) and applies forward observation operators to the model background to generate observation priors. A version of the ensemble square root filter (Whitaker and Hamill 2002) that pre-calculates observation priors then updates them together with the state variables within the filter (Anderson and Collins 2007) is used. Recently, the GSI EnKF system was updated to assimilate radar observations; radar forward operators were added to the GSI suite and the EnKF system was given the capability to update hydrometeor variables (e.g., Johnson et al. 2015; Jones et al. 2018a; Tong et al. 2020). CAPS recently added radar reflectivity forward operator consistent with the Thompson microphysics scheme. The new operator follows Jung et al. (2008a), which uses the T-matrix method (Vivekanandan et al. 1991; Bringi and Chandrasekar 2001) for raindrops and Rayleigh scattering approximation for ice hydrometeor species (i.e., snow, hail, graupel) to calculate scattering amplitudes. In this study, radar observations undergo automatic QC including velocity dealiasing, using a procedure developed by CAPS (Brewster et al. 2005). These enhancements allow the GSI EnKF system to be used in several storm-scale modeling studies (e.g., Johnson et al. 2015; Johnson and Wang 2017; Johnson et al. 2017; Jones et al. 2018b; Jung et al. 2018a,b; Chipilski et al. 2020).

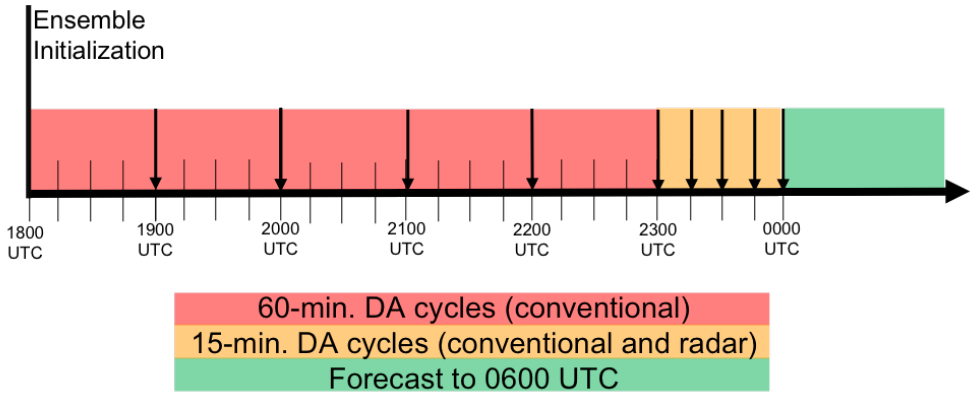

Fig. 4.3. Flow diagram detailing the CTRL DA experiment configuration. A bold vertical line at 1800 UTC marks when the ensemble of forecasts is first initialized. Downward pointing arrows mark when DA occurs. The final free forecasts start at 0000 UTC, after 6 hours of DA cycles.

CTRL (Table 1) closely resembles the configuration of the 2019 CAPS storm-scale ensemble based on GSI EnKF initial conditions as part of the Community-Leveraged Unified Ensemble (CLUE) (Clark et al. 2019). A schematic on the setup of DA cycles is shown in Fig. 4.3. To create the initial ensemble, the 1800 UTC North American Mesoscale (NAM) Model analysis is first interpolated to the model grid to serve as the initial ensemble mean. 3-hour forecast perturbations from the 1500 UTC operational Short-Range Ensemble Forecast (SREF) as well as their negative versions are added to the interpolated NAM analysis to generate a 40-member ensemble. Corresponding SREF forecasts also provide lateral boundary conditions. Conventional observations (e.g., surface stations, buoys, soundings) are assimilated hourly between 1900 – 0000 UTC, and radar (Z and V_r) observations are assimilated every 15 minutes in the last hour. Covariance localization uses the Gaspari and Cohn (1999) weighting function with user-specified cutoff radii. In this study the horizontal cutoff radius is 300 km for conventional observations and 12 km for radar observations. The vertical cutoff radius for all observations is 0.7 scale height (natural log of pressure). The RTPS covariance inflation algorithm is used that restores the spread of the analysis to 99% of the background spread. At 0000 UTC an ensemble of 10 free forecasts

are initialized from the final analyses and run to 0600 UTC; the forecast configuration is discussed in the following subsection. Ensemble members selected to initialize forecasts employ a diverse suite of model physics options, and are chosen to enhance forecast diversity.

Initial Conditions	1800 UTC NAM analysis with 1500 UTC SREF perturbations
Assimilation Frequency	60 Minutes 1800 – 2300 UTC (conventional only) 15 Minutes 2300 – 0000 UTC (radar and conventional when available)
Radar Localization radius (Horizontal / Vertical)	12 km / 0.7 scale height
Conventional Localization Radius (Horizontal / Vertical)	300 km / 0.7 scale height
Observation errors (reflectivity / radial velocity)	6 dBZ / 3 m s^{-1}
Inflation	99% RTPS
Radar data thinning interval (horizontal / vertical)	6 km / 1 km

Table 4.1. CTRL experiment configuration

For radar DA, Z observations are from the Multi-Radar Multi-Sensor (MRMS; Smith et al. 2016) products. The MRMS system performs QC on 140 WSR-88D radars and generates a mosaic of the observations on a three-dimensional grid with a horizontal resolution of 0.01° latitude × 0.01° longitude and 33 vertical levels. V_r observations are processed by ourselves using CAPS's software package that includes QC. The data are interpolated to the model grid horizontally and kept on radar elevation levels in the vertical for each radar site (Xue et al. 2006). Z and V_r observations are thinned horizontally to every 6 km during assimilation. Z observations

are also thinned vertically to every 1 km in radar echoes (Z > -90 dBZ) and to every 2 km in clear air regions (Z < -90 dBZ). Such thinning is done to fit the GSI EnKF analyses into available computer memory, which is necessary when a large continent-size domain is used during, e.g., the CLUE experiment (the GSI memory usage is inefficient). The V_r and Z observation errors are assumed to be 3 m s^{-1} and 6 dBZ, respectively, in the DA.

4.2.3 Prediction model settings

Aside from the smaller forecast domain focused on the MCS impacted region, the grid specifications and model physics largely follow the CAPS GSI EnKF-initialized ensemble (Jung et al. 2018b) that is part of the CLUE (Clark et al. 2018). Forecasts are run using the advanced research Weather Research and Forecast Model (WRF-ARW; Skamarock et al. 2008) version 3.8.1. The forecast domain spans much of Texas, Louisiana, and Mississippi (Fig. 4.1) with 433 × 241 grid points in the horizontal and 51 vertical levels in sigma-pressure coordinates. The horizontal grid spacing is 3 km. The vertical grid follows the 2017 CLUE configuration (Clark et al. 2018) where the finest vertical grid spacing is located near the surface. All ensemble members use the Rapid Radiative Transfer Model (RRTM; Mlawer et al. 1997) for general circulation model (RRTMG: Iacono et al. 2008) to represent short- and longwave radiation. During DA (1800 – 0000 UTC), forecasts are run with the aerosol aware Thompson microphysics scheme (Thompson et al. 2008; Thompson and Eidhammer 2014) but with varying graupel density across the members; 0000 UTC EnKF-initialized forecasts vary microphysics schemes (Thompson [Thompson et al. 2008], Morrison [Morrison and Grabowski 2008], Milbrandt and Yau [MY; Milbrandt and Yau 2005], National Severe Storms Laboratory [NSSL; Mansell et al. 2010], and Predicted Particle Properties [P3; Morrison and Milbrandt 2015]) between ensemble members. Ensemble members

also employ different planetary boundary layer physics parameterizations (Mellor-Yamada-Janjić [MYJ; Janjić 1990, 1996, 2001], Yonsei University [YSU; Hong et al. 2006], and Mellor-Yamada-Nakanishi-Niino [MYNN; Nakanishi and Niino 2009]) and the Noah land surface model (Chen and Dudhia 2001) both during the DA and forecast periods. An in-depth overview of ensemble member physics options is provided in CAPS (2017).

4.2.4 DA configuration experiments

Sensitivity experiments are designed to evaluate the impact of EnKF DA configurations on the forecast skill. Each experiment repeats the CTRL DA procedure (Table 1) except one aspect of the DA configuration (e.g., DA frequency, covariance localization radius) is modified. Experiment results are broken into three categories that analyze forecast sensitivity to 1) ensemble initialization, 2) radar DA parameters (i.e., data thinning, covariance localization, covariance inflation, and observation errors), and 3) how frequently radar observations are assimilated.

To start the EnKF DA cycles, an ensemble of forecasts needs to be initialized. The ensemble is often started from an ensemble of another forecasting system (e.g., Wheatley et al. 2015; Schwartz 2019), by adding perturbations to a control initial condition (e.g., Snyder and Zhang 2003; Dowell et al. 2011; Johnson et al. 2015), or by combing the two perturbations (e.g., Snook et al. 2016; Jung et al. 2018b). One experiment in this study investigates the impact of different methods of the initial ensemble initialization. CTRL is initialized from the 1800 UTC NAM analysis plus perturbations derived from 3-hour forecasts of the 1500 UTC cycle SREF (Table 1). The SREF ensemble has 24 members, allowing for the construction of 12 pairs of independent positive and negative perturbations. The first member of CNTL used NAM analysis without SREF perturbation. The remaining 15 members reuse some of the perturbations out of the

first 24 members. The second experiment directly interpolates initial conditions from the SREF 3-hour forecasts without using the NAM analysis (hereafter NO_NAM). To initialize the 40-member NO_NAM ensemble, 16 of the 24 SREF forecasts provide initial conditions twice. Corresponding SREF forecasts used to generate initial conditions also provide lateral boundary conditions in CTRL and NO_NAM. The boundary conditions of CTRL member 1 is provided by the NAM forecast. Because of the use of different physics options, no two members have the same configuration, even though share the same initial and boundary conditions. Ensemble initialization experiments provide insight into forecast sensitivity to the initial ensemble initialization, even with many DA cycles performed.

The second set of experiments evaluates the impact of different radar DA parameters including data thinning, covariance localization radius, assumed observation errors, and covariance inflation. Data thinning experiments assimilate radar observations thinned over increasingly large intervals either horizontally (3 km, 6 km, 9 km) or vertically (500m, 1 km, 2 km) and are labeled by the direction observations are thinned in the horizontal and vertical directions (ThinH3V1, ThinH6V1, ThinH9V1, ThinH3V0.5, ThinH3V2). Following CTRL, all experiments thin Z observations over twice as large of a vertical interval in clear air regions. V_r observations are not further thinned during this study because the data are preserved on about a dozen radar elevations and thus already sparser than MRMS Z. The covariance localization experiments vary the localization radius for radar observations either horizontally (6 km, 12 km, 18 km) or vertically (0.4 scalar height, 0.7 scalar height, 1.0 scalar height) during DA and are referred to by the length of the applied radius in the horizontal and vertical directions (CovH18V0.7, CovH12V0.7, CovH6V0.7, CovH12V0.4, CovH18V1.0). The covariance inflation experiments vary the percentage (80%, 90%, 99%, 110%) that posterior ensemble spread is relaxed

to that of the prior ensemble via the RTPS algorithm when radar observations are assimilated (2300 – 0000 UTC). Experiments are referred to by the inflation factor applied (Inf80, Inf90, Inf99, Inf110). Observation error experiments vary observation errors for V_r (3 m s^{-1}, 6 m s^{-1}) and Z (6 dBZ, 9dBZ) and are referred to by the assumed errors (3ms6dBZ, 6ms9dBZ).

Experiments that assimilate radar data at higher frequencies have demonstrated mixed success in previous studies. Stratman et al. (2020) suggest that assimilating observed Z more frequently can cause predicted storms to spin up more quickly and better suppress spurious convection. Frequently assimilating observations can also introduce imbalances into the ensemble that propagate with time to degrade forecast skill (Hu and Xue 2007; Johnson and Wang 2017). To determine the impact of radar DA frequency, Z and V_r observations are assimilated at 5, 10 and 15 minute intervals during the final hour of DA (2300 – 0000 UTC). These intervals roughly correspond to the frequency that WSR-88D radars sample the atmosphere (~ 5 minutes) and the frequency some current real-time systems assimilate radar observations (e.g., Wheatley et al. 2015; Jung et al. 2018a). DA frequency experiments are referred to by how frequently observations are assimilated (5Min, 10Min, 15Min).

4.2.5 Forecast evaluation

The 0000 – 0600 UTC forecast Z is subjectively and objectively verified against observations to evaluate the predicted evolution of storm structure. This forecast evaluation period corresponds to when the impact of assimilated radar observations is most prominent (Kain et al. 2010a), and when the Warn-on-Forecast paradigm is expected to offer the most benefit to operational forecasters (e.g., Stensrud et al. 2009).

During forecast evaluation small errors in storm placement can substantially degrade

objective performance by double penalizing a forecast i.e., adjoining grid points may receive a false negative and false positive. This problem is exacerbated when verifying localized events (e.g., convective storms) but can be ameliorated by verifying the occurrence of an event within a prescribed radius. The neighborhood maximum ensemble probability (NMEP; Schwartz et al. 2010; Schwartz and Sobash 2017) method is used to generate probabilistic forecasts. This study verifies the probability *Z* exceeding 40 dBZ (P[Z > 40 dBZ]) within a 12-km neighborhood. This neighborhood radius reduces impact of small forecast displacement errors but ensures the short-term forecast remains precise enough to detect local impacts. To smooth probabilistic forecasts a Gaussian filter with a smoothing length scale of 12-km is applied to the output. It is noted this study verifies a relatively large Z threshold; this is done to verify the location of predicted storm cores and to eliminate regions of stratiform precipitation from the statistics.

Probabilistic forecast skill is objectively evaluated using the Brier Skill Score (BSS; Brier 1950), which can be decomposed into three distinct components (Murphy 1973): reliability, resolution, and uncertainty (Table 4.2). Reliability is the difference between predicted probability and observed frequency, forecast skill improves when this difference is minimized. Resolution, which should be maximized to improve ensemble performance, is the difference between the climatological probability and the observed relative frequency for a given probability threshold. Unlike the other two components, forecast uncertainty cannot be changed through calibration and is a function of the climatological probability. Reliability diagrams provide a visual representation of the BSS by plotting forecast probability against observed frequency over increasingly large thresholds. In an unbiased system forecast probability is equal to observed frequency and the reliability curve falls along the one-to-one line. If forecast probability is larger (smaller) than observed frequency then the curve falls below (above) the one-to-one line and the forecast is high

(low) biased. Reliability diagrams also provide the frequency occurrence of each probability to evaluate forecast sharpness.

Metric	Formula
Reliability	$\frac{1}{N}\sum_{k-1}^{K} n_k(p_k - \bar{o}_k)^2$
Resolution	$\frac{1}{N}\sum_{k-1}^{K} n_k(\bar{o}_k - \bar{o})^2$
Uncertainty	$\bar{o}(1-\bar{o})$
Brier Skill Score	$\frac{Resolution - Reliability}{Uncertainity}$

Table 4.2. Scores used to evaluate forecasts in the study. The number of forecasts is N, the number of forecasts for a given k probability threshold is n_k, $\bar{o}$ is the observed climatology, and $\bar{o}_k$ is the observed relative frequency, and pk is the forecast probability.

4.3 Results

4.3.1 CTRL evaluation

Shortly after the start of forecast at 0015 UTC (Fig. 4.4a) CTRL predicts the P(Z > 40 dBZ) to be relatively high (> 0.90) within the confines of observed storms including the MCS located near Shreveport, Louisiana. Although CTRL predicts the P(Z > 40 dBZ) to exceed 0.7 for isolated storms located along the Mexico-Texas border at 0005 UTC (not shown), probabilities quickly diminish to approximately 0.4 by 0015 UTC (Fig. 4.4a). This region is cooler and drier than locations near the coast, suggesting that the environment is less conducive for convective development. Several forecasts in the CTRL ensemble (Fig. 4.4a) also predict spurious storms to initiate to the north and east of San Antonio. Reliability diagrams show probabilistic 0015 UTC forecasts exhibit a slight overprediction bias for both low and moderate probability threshold

events (P[Z > 40 dBZ] < 0.7) (Fig. 4.5a) primarily due to the spurious convection in Texas. Despite this bias, the BSS at 0015 UTC is relatively high (0.39) because CTRL is unbiased for high-probability events (P[*Z* > 40 dBZ] > 0.8).

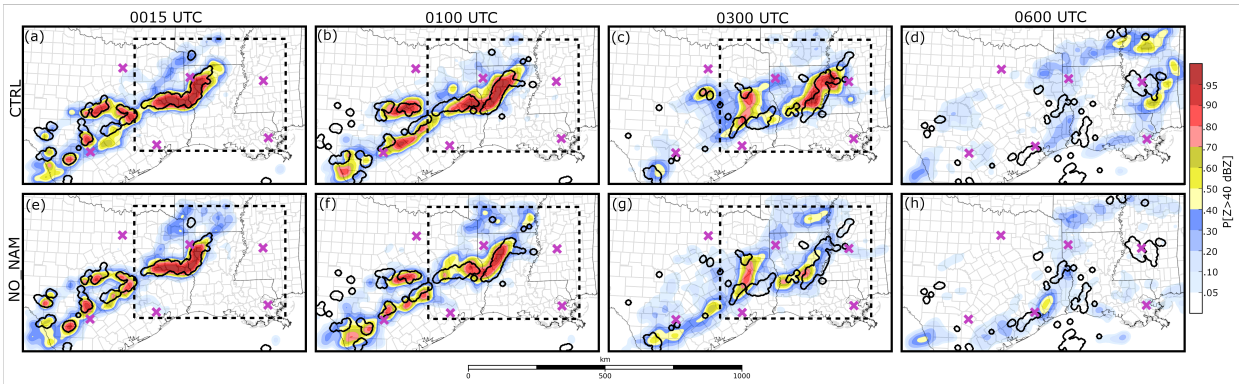

Fig. 4.4. The P(*Z* > 40 dBZ) predicted by (a-d) CTRL and (e-h) NO_NAM valid at the labeled times. Thick black contours represent locations where observed *Z* > 40 dBZ. Dashed squares in (a-c) and (e-g) mark the subdomains plotted in Fig. 4.6. Background maps are the same as Fig. 4.2.

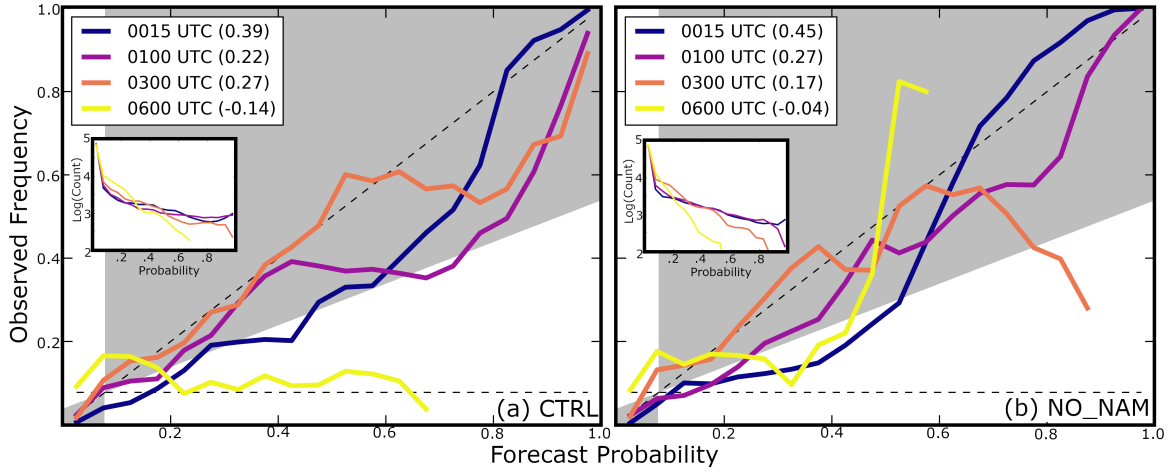

Fig. 4.5. Reliability diagrams for the probabilistic forecasts shown in (a) Fig. 4.4a-d and (b) Fig. 4.4e-h. Line colors correspond to when the forecasts are valid, the BSS for each forecast time is included in the legend.

As the MCS moves eastward through Louisiana (0100 – 0300 UTC), CTRL predicts the P(*Z* > 40 dBZ) to remain relatively high (> 0.7) near the observed storms (Fig. 4.4b-c). CTRL

predicts the isolated convection near San Antonio to grow in scale and form a line of storms (Fig. 4.4b) that spans Central Texas by 0100 UTC. This line of storms is spurious; the observed storms south and west of Houston at 0100 UTC are isolated (Fig. 4.2b) and the observed MCS is further east. By 0300 UTC, most of the observed storms in Central Texas weaken (Fig. 4.2c) and leave behind a large swath of stratiform precipitation and some weak storms. CTRL also predicts the storms in Central Texas to weaken around 0300 UTC, which causes forecast probabilities near San Antonio to decrease (Fig. 4.4c). The BSS at 0100 UTC is lower than at 0300 UTC (Fig. 4.5a) because more spurious storms are predicted at the earlier time. Spurious storms also cause the 0100 UTC reliability curve (Fig. 4.5a) to become high-biased for moderate and high probability events (P[Z > 40 dBZ] > 0.4). By 0300 UTC, when the spurious storms begin to weaken, the reliability curve (Fig. 4.5a) becomes less biased and more closely follows the one-to-one line.

CTRL predicts storms to move too quickly, and by 0600 UTC the predicted MCS (Fig. 4.4d) is east of the observed storms. Storm motion biases, such as this, are commonly observed in CAM forecasts and are often a consequence of model errors (e.g., Yussouf et al. 2016). Displacement errors cause the BSS at 0600 UTC to become negative and the reliability curve to become high-biased (Fig. 4.5a), suggesting CTRL exhibits no objective skill at this time. It is noted that the P(Z > 40 dBZ) exceeds 0.4 to the east of the observed MCS (Fig. 4.4d). Increasing the neighborhood radius when calculating the NMEP could improve forecast performance but would make forecasts less precise, thus this is not performed. Despite displacement errors, CTRL demonstrates some qualitative skill and predicts the MCS to begin weakening between 0500 – 0600 UTC, which is approximately the same time as observations (Fig. 4.2d).

4.3.2 Ensemble initialization

Forecast skill is highly sensitive to the fields used to initialize the initial ensemble. CTRL and NO_NAM *Z* forecasts (Fig. 4.4a, e) are qualitatively similar at 0015 UTC because both experiments are recently initialized using the same GSI EnKF configuration. As the forecasts continue the experiments become less similar; at 0300 UTC NO_NAM (Fig. 4.4g) predicts the P(*Z* > 40 dBZ) to be lower near the observed MCS (on the Louisiana-Mississippi border) and to be higher between San Antonio and Houston. Many of the NO_NAM-predicted storms in Central Texas (Fig. 4.4g) produce spuriously intense *Z* but correspond with weak storms observed between Houston and San Antonio (Fig. 4.2c). Unlike CTRL, which predicts the MCS to remain organized through 0600 UTC (Fig. 4.4d), NO_NAM predicts the MCS to weaken faster than observations. This causes NO_NAM to predict the P(*Z* > 40 dBZ) to be low (0 – 0.05) near the observed MCS in Mississippi (Fig. 4.4h).

Although CTRL qualitatively predicts the evolution of the MCS with more skill than NO_NAM, objective verification measures provide mixed results. During the first forecast hour NO_NAM (Fig. 4.4e-f) predicts the P(*Z* > 40 dBZ) to be lower than CTRL (Fig. 4.4a-b) for spurious convection located near San Antonio. Due to these spurious storms, the NO_NAM reliability curve at 0100 UTC (Fig. 4.5b) exhibits less of an overprediction bias than CTRL (Fig. 4.5a) and its BSS is larger. By 0300 UTC the NO_NAM BSS (Fig. 4.5b) is less than CTRL. Although the NO_NAM 0300 UTC reliability curve (Fig. 4.5b) is relatively unbiased for low and moderate probability threshold events (P[*Z* > 40 dBZ] < 0.6), the ensemble lacks sharpness (i.e., does not predict the P[Z > 40 dBZ] > 0.9) and overpredicts high probability threshold events (> 0.7). Both experiments score negative BSSs at 0600 UTC (Fig. 4.5); however, CTRL (Fig. 4.4d) is qualitatively more skilled because NO_NAM predicts no MCS in Mississippi (Fig. 4.4h).

Differences in forecast performance are attributed to the methods used to initialize each respective ensemble.

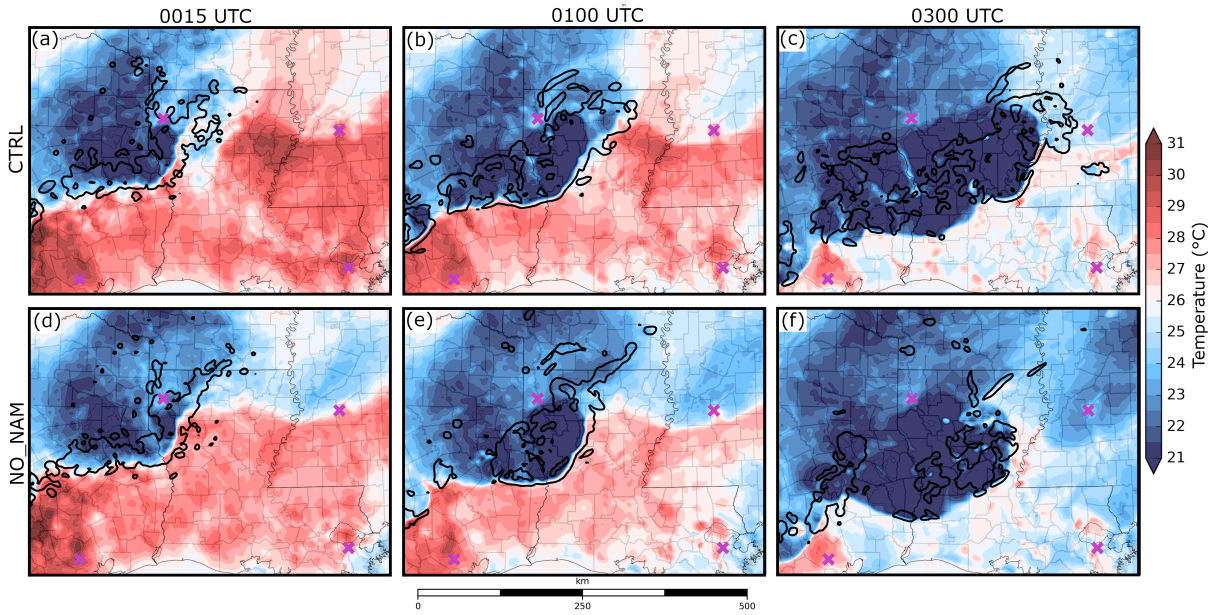

Fig. 4.6. (a-c) CTRL and (d-f) NO_NAM predicted air temperature at the lowest model grid level above the surface valid at the labeled times. Thick black contours mark where the column maximum updraft exceeds 1 m s^{-1}. A fuschia "x" marks the location of mentioned cities.

The predicted evolution of convection is highly sensitive to small-scale errors (e.g., Zhang et al. 2006; Hawblitzel et al. 2007; Melhauser and Zhang 2012). Deterministic forecasts initialized from the 0000 UTC CTRL and NO_NAM ensemble mean analyses and run with the same physics configuration evaluate how differences between the two initial conditions evolve to impact the predicted MCS. Both experiments predict a weak cold front moving south through Louisiana and Mississippi. At 0015 UTC CTRL (Fig. 4.6a) predicts air near the front to be approximately 2 °C warmer than NO_NAM (Fig. 4.6d). The relatively warm surface air is believed to enhance atmospheric instability and cause the MCS to intensify and form a robust updraft that extends along the Louisiana-Mississippi border by 0300 UTC (Fig. 4.6a-c). In contrast, NO_NAM predicts storm updrafts to be weaker and more fragmented at 0300 UTC because the MCS ingests cooler

and more stable air (Fig. 4.6d-f).

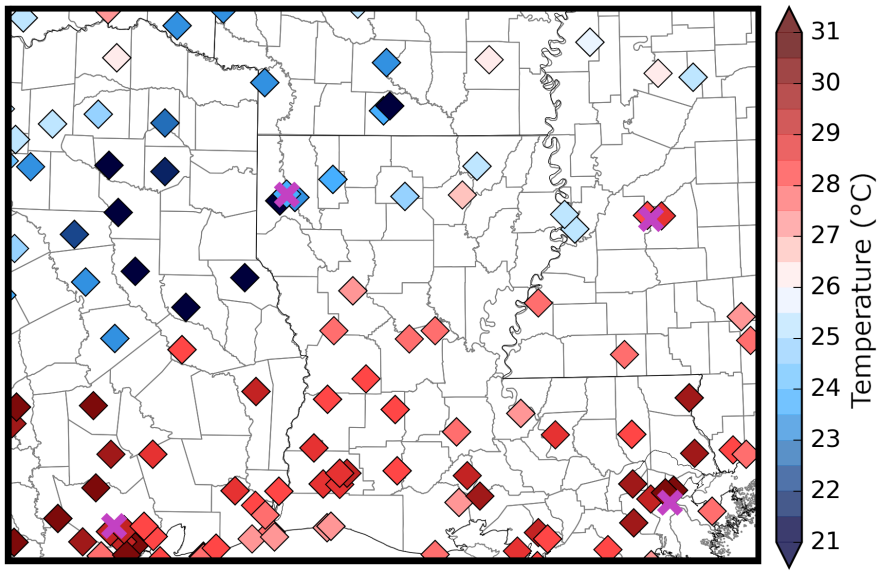

Fig. 4.7. Surface air temperature observations at 0000 UTC.

In a well observed environment, the GSI EnKF system should correct surface air temperature errors; however, Northern Louisiana is relatively rural and surface observing stations in this region are sparse (Fig. 4.7). Due to a limited number of surface observations, the EnKF more heavily weights observations taken outside this region and the ensemble background when generating the initial condition. This potentially causes analyses to become susceptible to errors originating from the initial ensemble at 1800 UTC. For example, CTRL predicts surface air temperature with a smaller root mean square innovation than NO_NAM (2.13 °C and 2.46 °C, respectively) prior to the first DA cycle at 1900 UTC. This is likely because the CTRL ensemble mean at 1800 UTC is from the 1800 UTC NAM analysis, which assimilates more recent observations than the 1500 UTC cycle SREF. These results demonstrate the importance of reducing initial condition errors both when initializing an experiment and during DA, for the current experiment setup at least.

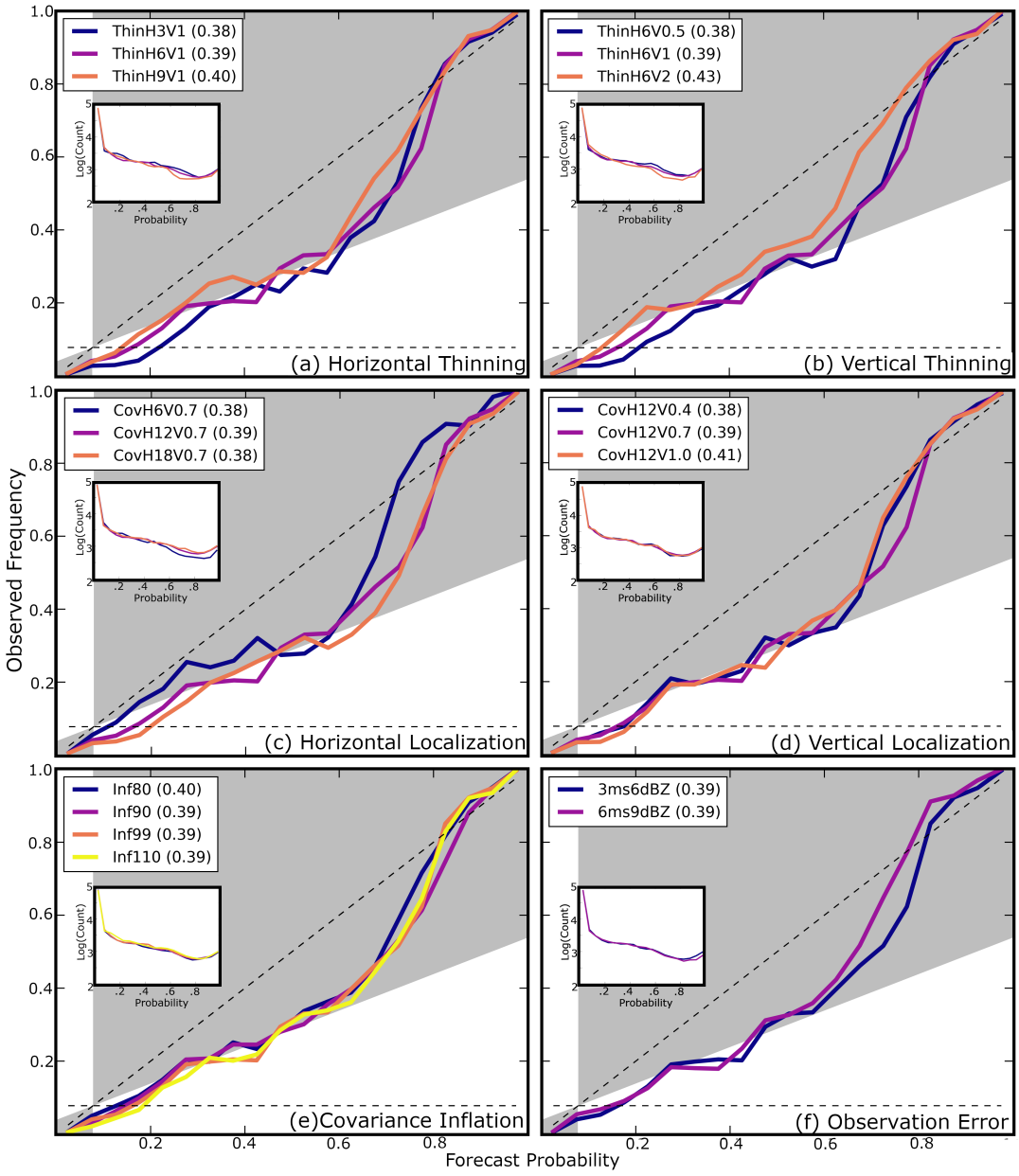

Fig. 4.8. Reliability diagrams for the forecast P(Z > 40 dBZ) valid at 0015 UTC for the labeled experiments including (a-b) radar data thinning, (c-d) covariance localization radius, (e) covariance inflation, and (f) observation errors. Experiments correspond to different colors, the BSS for each experiment is included in the legend.

4.3.3 Radar DA parameters

Many studies show radar observations provide important storm scale information, and when assimilated by an optimally configured EnKF, improve analyzed storm structure (e.g.,

Snyder and Zhang 2003; Dowell et al. 2004; Tong and Xue 2005) and subsequent forecasts (e.g., Snook et al. 2012, 2015). Modifying parameters directly related to radar DA, including data thinning, covariance localization radius and inflation, and observation errors initially impacts storm structure and the near-storm environment. The differences between experiments are most prominent during the first forecast hour because small-scale errors grow quickly (e.g., Melhauser and Zhang 2012) and degrade the initial benefits of radar DA. At 0015 UTC reliability curves (and BSSs) are already similar between all experiments (Fig. 4.8). Despite considerable overlap between the reliability curves, experiments that assimilate more radar observations (i.e., smaller data thinning intervals; Fig. 4.8a-b) or employ a larger horizontal covariance radius (Fig. 4.8c) have larger overprediction biases. Impacts of radar DA are often localized and not readily obvious in objective verification metrics that consider the full model domain and multiple storm systems. Subjectively evaluating probabilistic forecasts identifies how these parameters impact Z forecasts for individual storm systems (i.e., isolated convection and MCS).

Thinning radar observations potentially removes important fine-scale details on the storms. For small storms, such as those near San Antonio, thinning radar observations removes many if not all available in-storm observations. ThinH6V0.5 (Fig. 4.9d) predicts the P(Z > 40 dBZ) to be larger than ThinH6V2 (Fig. 4.9e), particularly for small storms located near San Antonio, because the experiment assimilates more in-storm observations. Assimilating more radar observations also causes ThinH6V0.5 (Fig. 4.9d) to predict more spurious convection in Central Texas, which contributes to the overprediction bias observed in the reliability diagram for low to moderate probability thresholds (P[Z > 40 dBZ] < 0.7) (Fig. 4.8b). The predicted MCS is less sensitive to data thinning (Fig. 4.9) because the storm system is much larger in scale and therefore all data thinning experiments assimilate a large number of observations from within the storm. Although

probabilistic forecasts of the vertical data thinning experiments are primarily discussed, similar but less extreme trends are observed in ThinH3V1 (Fig. 4.9b), ThinH6V1 (Fig. 4.9a), and ThinH9V1 (Fig. 4.9c). This is to be expected because the radar data are thinned over different intervals in the horizontal and vertical directions.

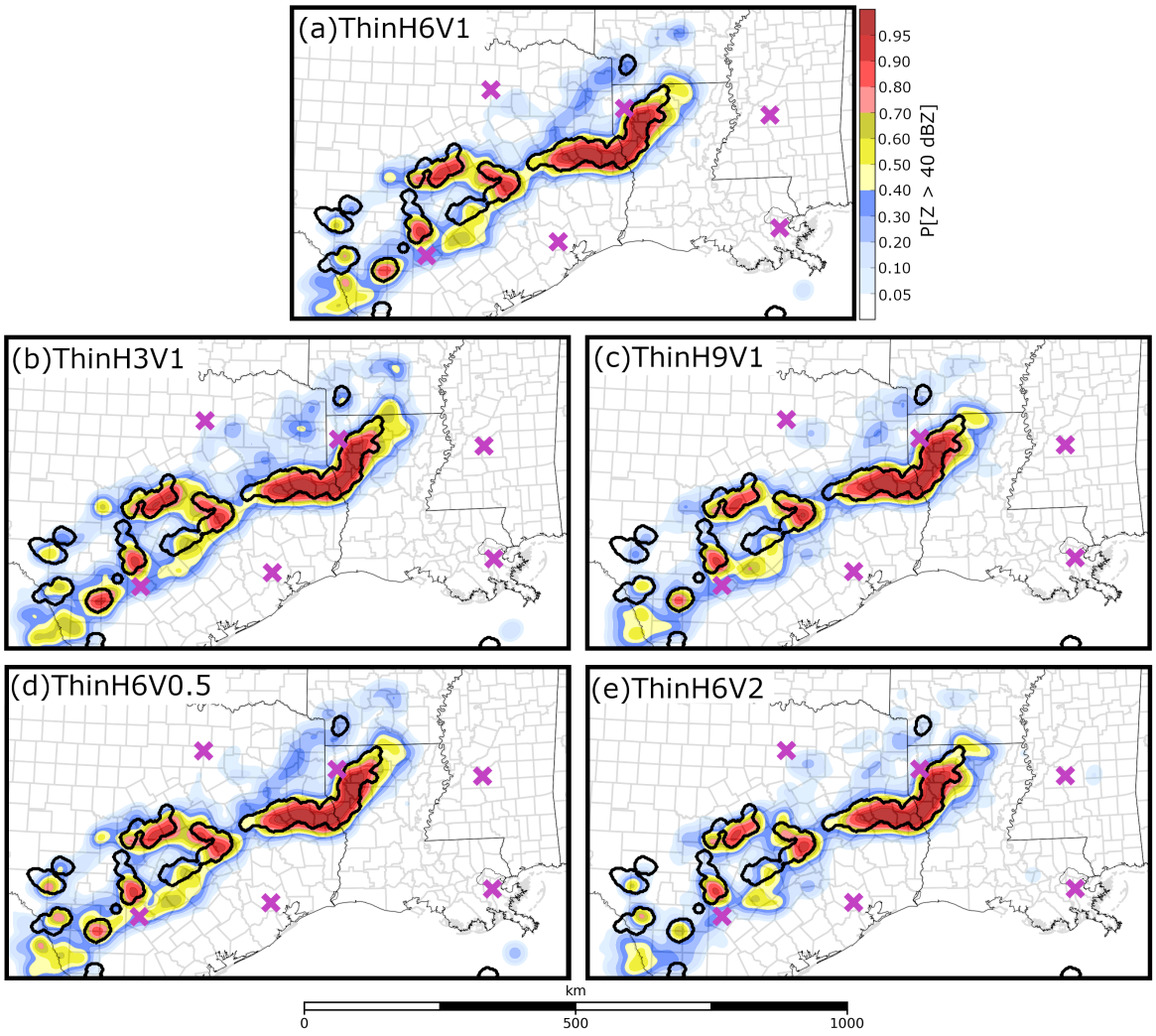

Fig. 4.9. The P(Z > 40 dBZ) predicted by (a) CTRL and the (b-c) horizontal and (d -e) vertical radar data thinning experiments valid at 0015 UTC. Thick black contours represent locations where observed Z > 40 dBZ. Background maps are the same as Fig. 4.2.

Many of the overprediction biases observed in experiments that assimilate near full-resolution radar observations (i.e., ThinH3V1, ThinH6V0.5) can be mitigated by reducing the

covariance localization radius for radar observations. When the most dense radar observations in this study are assimilated (ThinH3V0.5), the P(Z > 40 dBZ) increases outside of observed storm cores when the covariance localization radius is increased from 6 km (Fig. 4.10a) to 12 km (Fig. 4.10b). The reliability curve for the ThinH3V0.5 6 km horizontal localization radius experiment exhibits a smaller overprediction bias and larger BSS than the 12 km experiment (Fig. 4.10c) because the experiment predicts fewer spurious storms. A smaller localization radius is likely appropriate for experiments that assimilate more dense observations because the radar information does not need to be spread over large regions to fill data gaps.

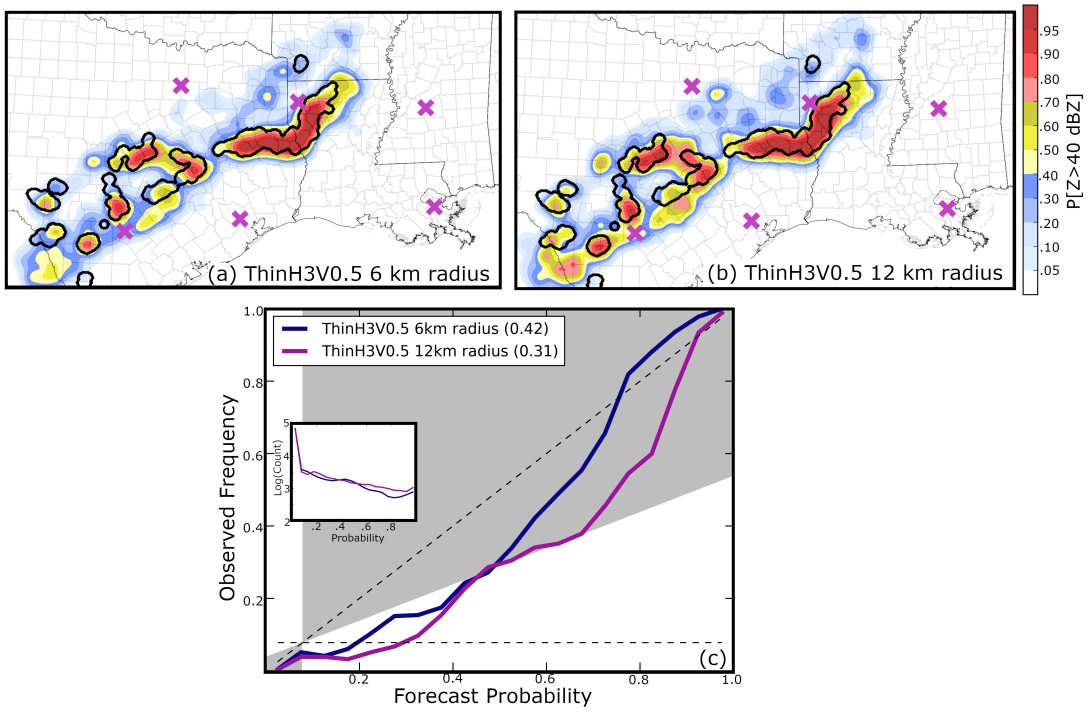

Fig. 4.10. (a-b) The P(Z > 40 dBZ) predicted by ThinH3V0.5 with a horizontal covariance localization radius of (a) 6 km and (b) 12 km during data assimilation. Forecasts are valid at 0015 UTC. Thick black contours represent locations where observed Z > 40 dBZ. Background maps are the same as Fig. 4.2. (c) The reliability diagram for the two probabilistic forecasts.

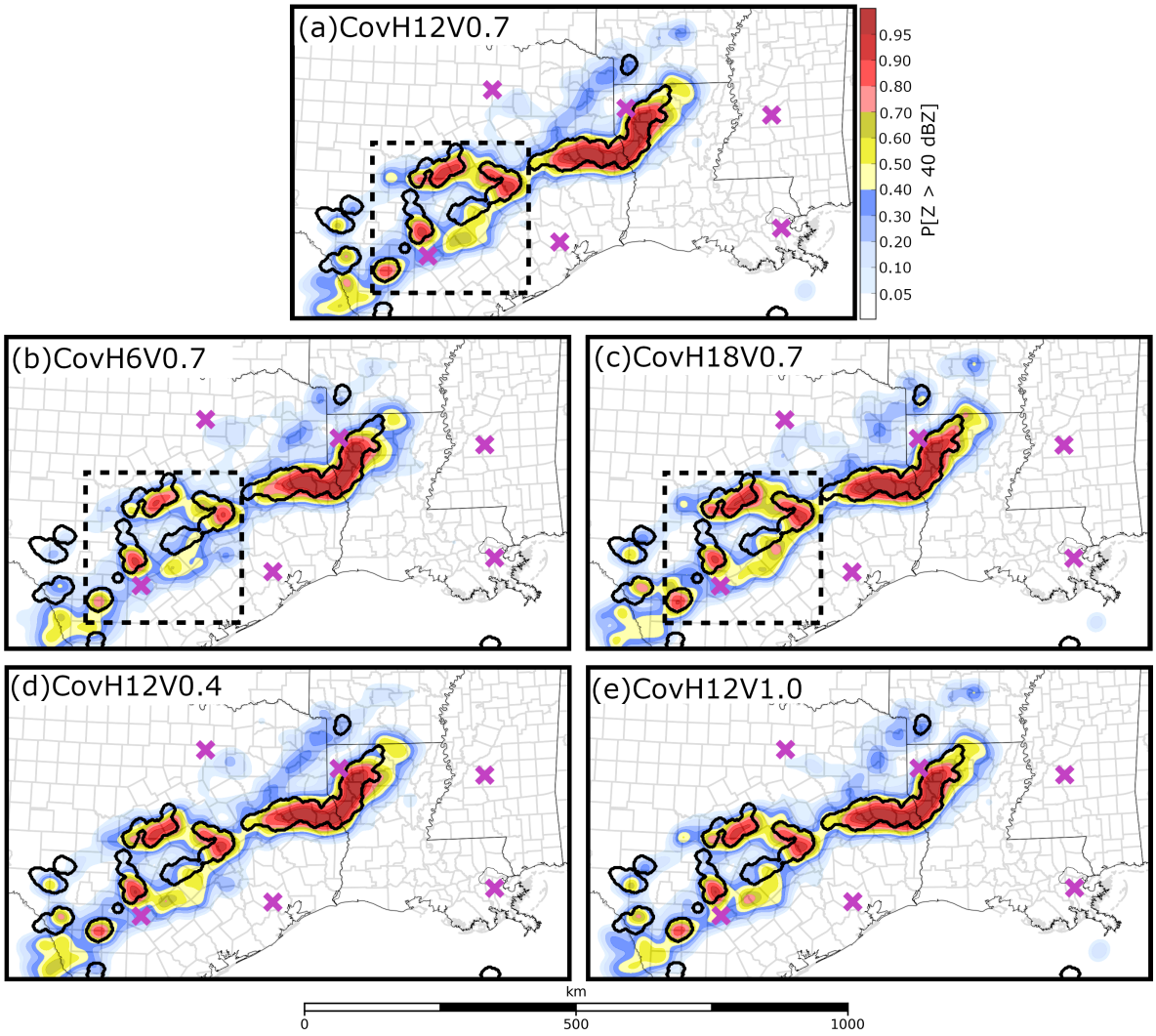

Fig. 4.11. The P(Z > 40 dBZ) predicted by (a) CTRL and the (b-c) horizontal and (d -e) vertical covariance localization experiments valid at 0015 UTC. Thick black contours represent locations where observed Z > 40 dBZ. Dashed squares in (a-c) mark the boundaries of the Fig. 4.12 domain. Background maps are the same as Fig. 4.2.

Increasing the horizontal covariance localization radius for radar observations causes ensembles to become more confident in high Z values located near observed storms for short-term forecasts. CovH6V0.7 (Fig. 4.11b) predicts the P(Z > 40 dBZ) to be smaller than either CovH12V0.7 (Fig. 4.11a) or CovH18V0.7 (Fig. 4.11c) for the isolated storms located near the Mexico-Texas border. CovH18V0.7 also predicts more spurious storms to be located between San

Antonio and Houston (Fig. 4.11c) than CovH6V0.7 (Fig. 4.11b), which increases the coverage of moderate forecast probabilities (> 0.4) outside observed storm cores. The CovH18V0.7 reliability curve is more high-biased than CovH6V0.7 at 0015 UTC (Fig. 4.8c) because the ensemble predicts more spurious convection in Texas, but BSSs are relatively similar because CovH18V0.7 predicts high forecast probabilities near observed storms and has improved resolution. Increasing vertical localization radius modestly increases the BSS (Fig. 4.8d) but probabilistic forecasts appear to be insensitive to this EnKF parameter (Fig. 4.11d-e). Although *Z* forecast skill is relatively insensitive to the vertical covariance localization radius in this study, Sobash and Stensrud (2013) suggest ensembles that employ a smaller vertical localization radius produce analyses with smaller root-mean square errors.

CovH18V0.7 predicts more spurious convection to develop because the ensemble is initialized at 0000 UTC (Fig. 4.12c) with more spurious, albeit weak, storm updrafts than the CovH6V0.7 (Fig. 4.12a) or CovH12V0.7 (Fig. 4.12b). Further, CovH18V0.7 (Fig. 4.12c) predicts the storm inflow region (southeast corner of the subdomain) to be moister than CovH6V0.7 (Fig. 4.12a) or CovH12V0.7 (Fig. 4.12b) because the EnKF makes larger adjustments to the background field (Fig. 4.12d-f). Although analyzed surface temperatures do not substantially change between the experiments (not pictured), the enhanced surface moisture makes the analyzed environment in CovH18V0.7 less stable than CovH6V0.7, which causes many of the weak storm updrafts to intensify shortly after DA. The localization radius employed by CTRL (CovH12V0.7) offers a compromise between CovH6V0.7 and CovH18V0.7. CovH12V0.7 (Fig. 4.11a) predicts the P(*Z* > 40 dBZ) to be larger than CovH6V0.7 (Fig. 4.11b) near observed storms but also predicts fewer spurious storms than CovH18V0.7 (Fig. 4.11c). This horizontal radius is smaller than the optimal radius determined by Sobash and Stensrud (2013) (18 km) but larger than what was used in Tong

and Xue (2005) (6 km). Discrepancies between these studies are attributed to differences in ensemble configuration (e.g., grid spacing, radar data density, ensemble size, DA system).

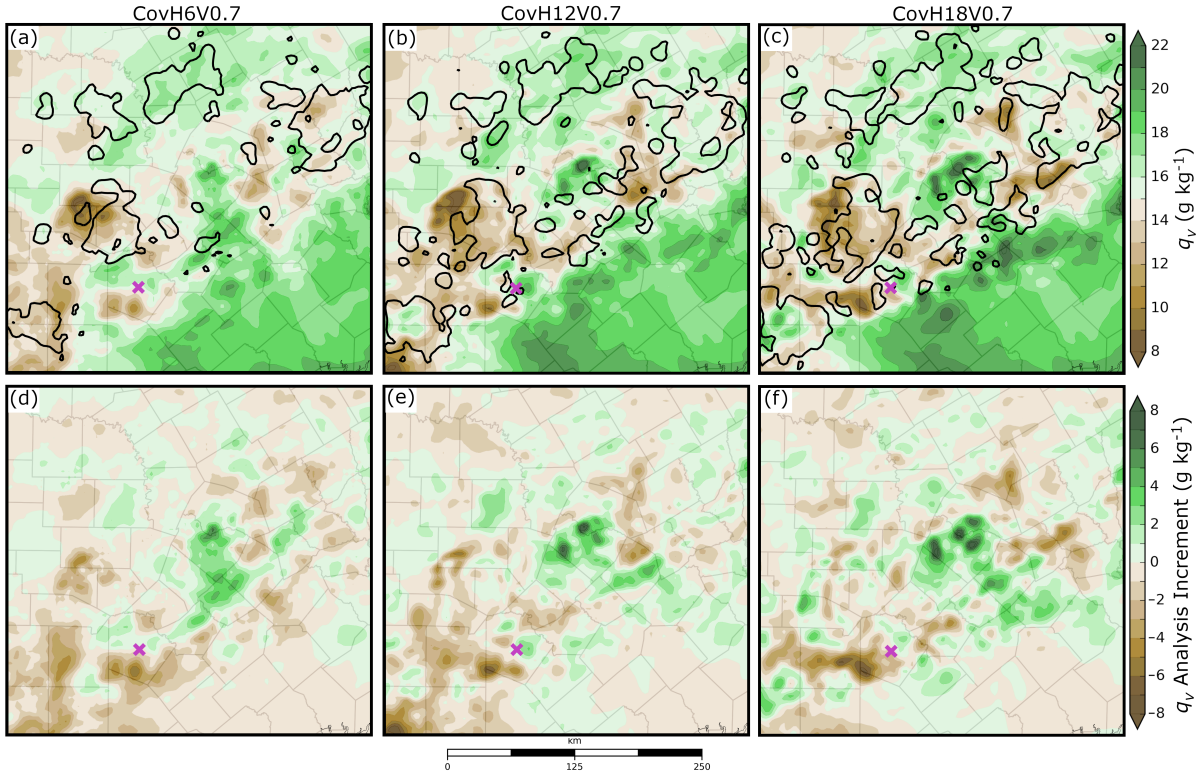

Fig. 4.12. (a-c) 0000 UTC analysis mean water vapor mixing ratio (q_v) at the lowest model level above the surface for the horizontal localization radius experiments. (d-f) q_v analysis increment at the same time and location. Thick black contours mark where the column maximum updraft exceeds 1 m s^{-1} in (a-c), the fuschia "x" marks the location of San Antonio.

The Z forecast skill does not substantially change between the covariance inflation experiments conducted, which is demonstrated by the reliability curves that exhibit considerable overlap (Fig. 4.8e). Although *Z* forecast skill is insensitive to the RTPS inflation factor, it remains important to perform ensemble covariance inflation to prevent collapse of ensemble spread and potential filter divergence when assimilating a dense network of observations. With some minor differences, 3ms6dBZ and 6ms9dBZ reliability curves are also quite similar at 0015 UTC (Fig. 4.8f) and consequently both ensembles have the same BSS. Because there are only minor

differences between probabilistic forecasts (not pictured), both sets of experiments will not be further discussed.

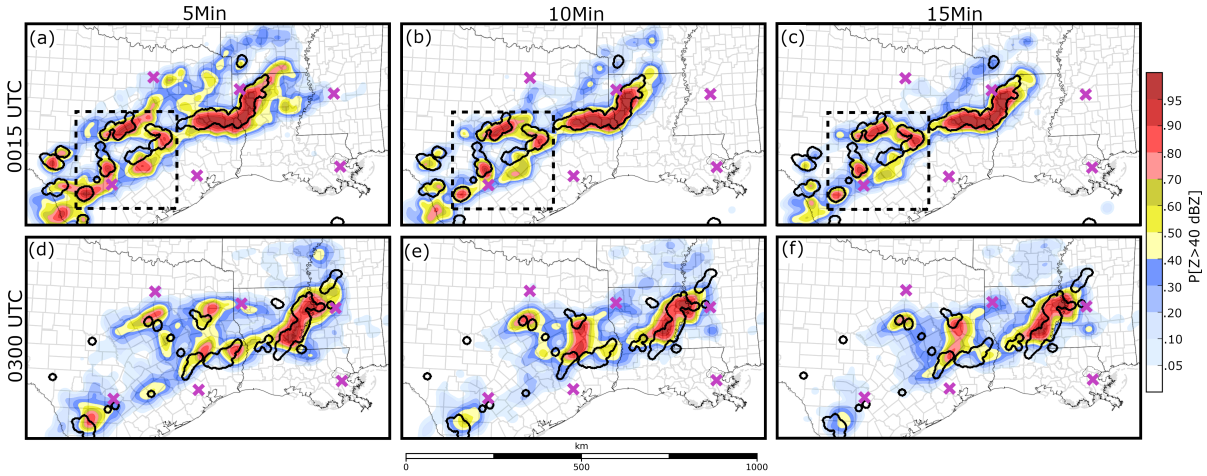

Fig. 4.13. The P(Z > 40 dBZ) predicted by (a,d) 5Min, (b,e) 10Min, and (c,f) 15Min valid at the labeled times. Thick black contours represent locations where observed Z > 40 dBZ. Dashed squares in (a-c) mark the boundaries of the Fig. 4.14 domain. Background maps are the same as Fig. 4.2.

4.3.4 DA frequency

Z forecast skill is highly sensitive to how frequently radar observations are assimilated by the GSI EnKF system. 5Min (Fig. 4.13a) predicts the P(Z > 40 dBZ) to be larger than either 10Min (Fig. 4.13b) or 15Min (Fig. 4.13c) for isolated storms that occur along the Mexico-Texas border at 0015 UTC. Forecast probabilities are larger because 5Min assimilates many more radar observations than any other experiment. 5Min (Fig. 4.13a) also predicts more spurious storms to develop and predicts more widespread moderate and high forecast probabilities (P[Z > 40 dBZ] > 0.4) than either 10Min (Fig. 4.13b) or 15Min (Fig. 4.13c). 5Min predicts additional spurious convection to develop because ensemble analyses (Fig. 4.14a) contain more storm updrafts than either 10Min (Fig. 4.14b) or 15Min (Fig. 4.14c). Many of the spurious storms diminish in intensity by 0300 UTC, which causes the three DA frequency experiments to become more similar (Fig.

4.13d-f), though 5Min (Fig. 4.13d) predicts higher forecast probabilities near observed storms in Texas at this time.

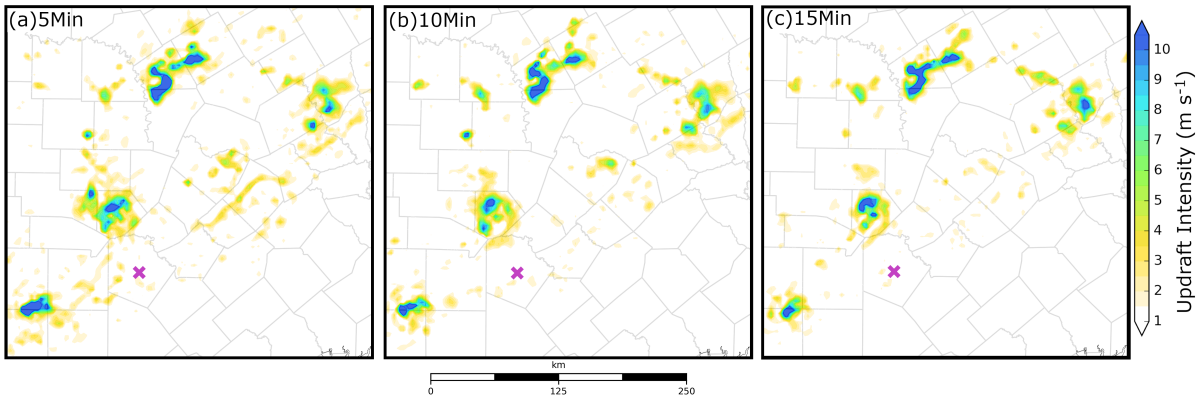

Fig. 4.14. Analysis mean column maximum updraft at 0000 UTC for the data assimilation frequency experiments. The fuschia "x" marks the location of San Antonio.

The 5Min reliability curve (Fig. 4.15a) exhibits a large overprediction bias at 0015 UTC because the ensemble is initialized with a large number of spurious storms. The overprediction bias is large enough to eliminate almost all *Z* forecast skill (5Min BSS ≈ 0). 10Min and 15Min reliability curves (Fig. 4.15a) also have an overprediction bias at 0015 UTC, but it is smaller in magnitude and has less of an impact on *Z* forecast skill (BSS > 0.35). Although the 5Min reliability curve remains high-biased at 0300 UTC (Fig. 4.15b), the BSS increases to be more similar to 15Min because many of the spurious storms weaken.

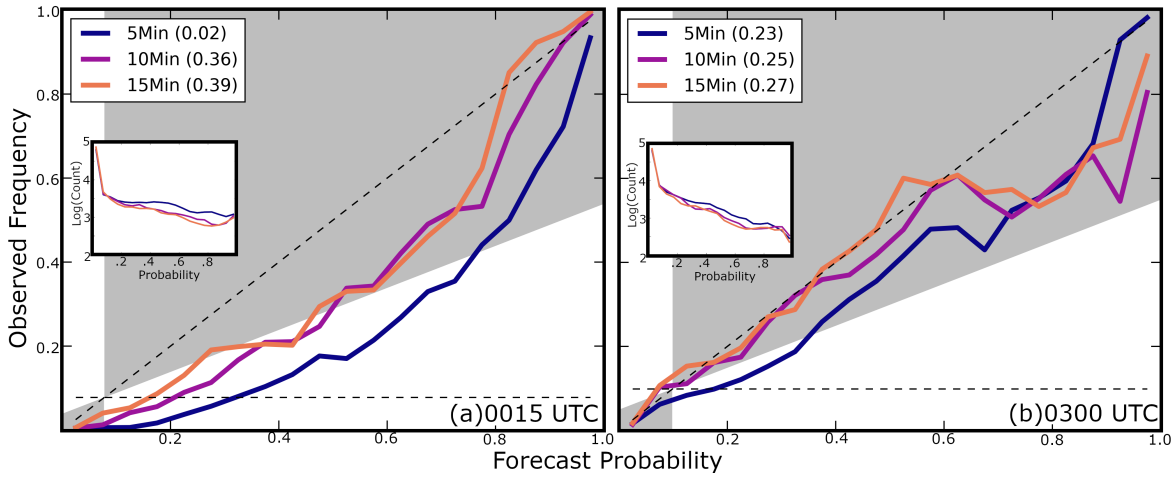

Fig. 4.15. Reliability diagrams for the probabilistic forecasts shown in (a) Fig. 4.13a-c and (b) Fig. 4.13d-f. DA frequency experiments correspond to different colors, the BSS for each experiment is included in the legend.

The balance of model states for each DA frequency experiment is evaluated using the absolute mean surface pressure tendency (*N*) (Lynch and Huang 1992; Hu and Xue 2007; Pan and Wang 2019). *N*, which is averaged over the domain and ensemble members, is defined by:

$$\mathrm{N} = \left(\frac{1}{n_x \times n_u \times n_{mem}}\right) \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \sum_{k=1}^{n_{mem}} \left|\frac{\partial p_s}{\partial t}\right|_{i,j} \qquad (1)$$

where n_x and n_y are the dimensions of the domain, n_{mem} is the number of ensemble members, and $\left|\frac{\partial p_s}{\partial t}\right|$ is the absolute change in air pressure at the lowest grid point above the surface. *N* values are calculated every 5 minutes during the 6-hour forecast period. For all experiments *N* values are initially large (~40 – 55 pa min^{-1}) but decrease with time to approximately 5 pa min^{-1} as the model adjusts to any imbalances that originate from DA (Fig. 4.16). To a large extent this is done through acoustic waves and hydrostatic adjustment. Many of the storms also weaken during the forecast period, further minimizing changes in pressure. 5Min produces the largest *N* values throughout the forecast period (Fig. 4.16), while 10Min and 15Min *N* values are smaller and more similar in magnitude. Large pressure tendencies suggest that the more frequent DA cycling causes 5Min

forecasts to become less balanced. Since 10Min and 15Min *N* values are more similar in magnitude, small increases in the DA frequency (i.e., 10Min) does not appear to cause the ensemble to become as unbalanced.

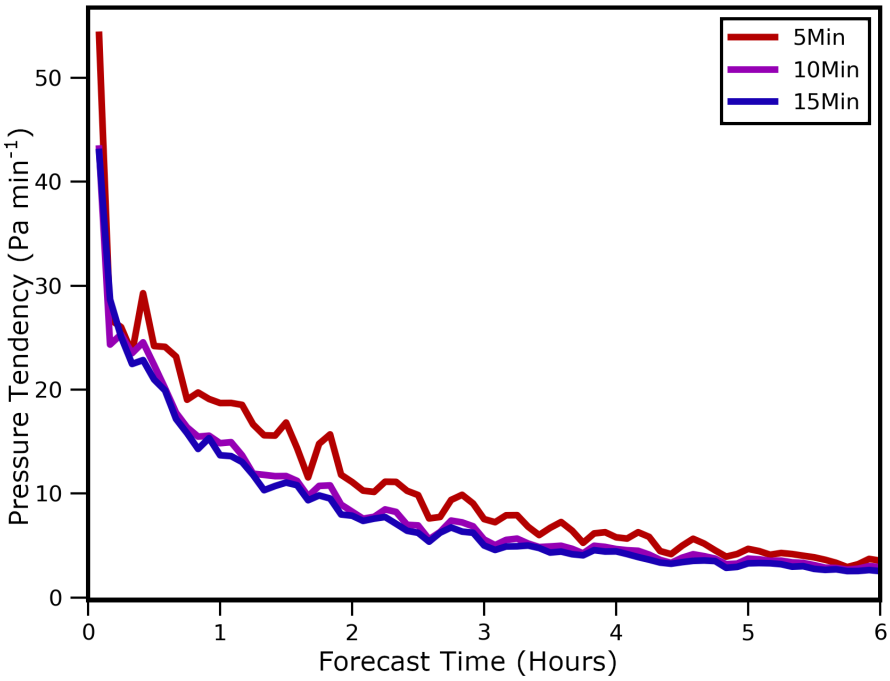

Fig. 4.16. The ensemble mean absolute pressure tendency during the first six forecast hours for the three DA frequency experiments.

Results of the DA frequency experiments concur with previous studies (e.g., Wang et al. 2013; Johnson et al. 2017; Pan and Wang 2019). Johnson and Wang (2017) hypothesize that imbalances introduced during DA do not have enough time to adjust to the model before the next assimilation cycle. It is noted that many other studies (e.g., Aksoy et al. 2009; Jung et al. 2012; Dawson et al. 2012) have successfully assimilated radar observations much more frequently (≤ 5 minute intervals). Stratman et al. (2020), who assimilates phased array radar observations every minute, suggests frequent DA cycling can quickly spin up thunderstorms and suppress spurious convection. Given that most of the adjustment occurs within the first few forecast minutes so that

all experiments have similar levels of surface pressure noise after 5 minutes (Fig. 4.16), the negative impact of assimilating 5 minute data is likely insignificant given the potential benefit quickly spinning up observed storms within the model. It is likely 5Min skill can be further improved by more thoroughly tuning the GSI EnKF system. For realtime forecasting systems where computational cost is a significant issue, assimilating radar observations at 10 to 15 min intervals yields good results is encouraging.

4.4 Summary and discussion

This study evaluates short-term (0 – 6 hour) convection-allowing model (CAM) forecasts initialized using the GSI EnKF system enhanced with radar data assimilation capabilities. Forecasts are run for a mesoscale convective system (MCS) event that occurred in the Southern United States on 28 – 29 May 2017 and produced multiple tornado, wind, and hail reports. The control configuration of the GSI EnKF system resembles the Center for the Analysis and Prediction of Storms (CAPS) storm scale ensemble forecast run during the 2019 Hazardous Weather Testbed Spring Forecast Experiment as part of the CLUE. In addition to verifying forecasts run using this configuration, sensitivity experiments are run to evaluate the impact of GSI EnKF configurations, including the initial ensemble initialization, as well as data thinning, covariance localization and inflation, observation error specification, and DA frequency for radar observations. Ten-member multi-physics ensemble forecasts and deterministic forecasts from the final EnKF ensemble and mean analyses, respectively, are run for the assessment of DA impacts on forecasts.

The CTRL DA configuration creates the most skilled forecasts in this study while remaining viable for realtime use. The initial ensemble of CNTL is centered on the 1800 UTC NAM analysis with perturbations derived from 3 hour SREF forecasts. The DA system assimilates

thinned radar observations (6 km horizontally, 1 km vertically for in-storm regions, and 2 km vertically for clear air regions) every 15 minutes to remain computationally efficient for experiment run over the CONUS domain. Although forecast skill can be further improved when assimilating more observations (i.e., less data thinning, increased DA frequency) the experiment becomes prohibitively expensive for realtime use when run over the CONUS domain. The covariance localization radius for Z and V_r is set to 12 km in the horizontal and 0.7 scalar height in the vertical to update unobserved regions. CTRL skillfully predicts the evolution of the MCS during the first three forecast hours; however, the predicted storm becomes displaced from observations at later times because it moves too quickly. Despite displacement errors that limit objective forecast skill, CTRL predicts the MCS to weaken during the final forecast hour, which is approximately the same time as observations. CTRL predicts isolated convection with less skill. The ensemble predicts small storms located near the domain boundary with less confidence and predicts nearby spurious convection.

Other experiments modify the CTRL configuration to determine which DA parameters the forecast is most sensitive to or how they impact the forecast skill. To evaluate the impact of initial ensemble initialization at the beginning of DA window, two experiments are conducted. CTRL is initialized from the 1800 UTC NAM analysis plus perturbations derived from 3-hour SREF forecasts, and the second experiment (NO_NAM) is initialized by directly interpolating from the 3-hour SREF forecasts valid at 1800 UTC. NO_NAM forecasts exhibit less skill than CTRL because the ensemble predicts the atmosphere to be cooler and more stable, which causes the predicted MCS to weaken too quickly. EnKF temperature updates are limited because the region ahead of the MCS is relatively rural and surface observing stations are sparse. To generate an analysis the EnKF relies mostly on distant observations and the ensemble background, causing

analyses to be more susceptible to errors that originate from ensemble initialization. CTRL initial conditions are generally more skilled because of the additional observations contained in the 1800 UTC NAM analysis.

Modifying the radar DA parameters, including data thinning, covariance localization radii and inflation, and observation errors, has a relatively small impact on forecast skill. Further, the radar DA parameter experiment differences are most prominent during the first forecast hour because small-scale errors grow rapidly to degrade the benefits of radar DA that mainly improve storm-scale structures. Results show increasing the horizontal covariance localization radius or assimilating more radar observations (i.e., less data thinning) increases forecast confidence in high reflectivity values near observed storms but also modestly increases the coverage of spurious convection. Small, isolated convection is most sensitive to changes in these parameters. In contrast, the predicted MCS is relatively insensitive to the covariance localization radius or data thinning because the storm is larger in scale so many more in-storm observations are available for assimilation. Changing the assumed radial velocity and reflectivity observation errors (3 - 6 m s^{-1} and 6 – 9 dBZ, respectively) does not substantially impact forecast skill in this study. Forecast skill is also relatively insensitive to the range of relaxation-to-prior-spread inflation factors (80% - 110%) employed in this study. Still, covariance inflation is an important aspect of EnKF DA and prevents the collapse of ensemble spread when assimilating a dense network of observations (e.g., radar observations).

DA frequency experiments demonstrate that assimilating radar observations more frequently (i.e., every 5 minutes) increases forecast confidence for observed storms, but substantially reduces forecast skill initially because the analysis produces too many spurious storms. Results of this study concur with Johnson and Wang (2017), which hypothesizes

assimilating observations too frequently degrades forecast skill because the DA-induced imbalances are unable to adjust to the model before the next cycle. Although assimilating observation less frequently (e.g., every 15 minutes) improves objective forecast skill in this study, Stratman et al. (2020) note more frequent DA cycling can reduce the spin up time for storms and suppress spurious convection. It is believed the GSI EnKF system requires additional tuning to assimilate radar observations more frequently and produce more skilled analyses and forecasts.

Although this study evaluates several key aspects of the experimental design, many other factors that impact forecast performance should be investigated. For example, future studies should compare forecast differences between mixed physics and single physics ensembles. CTRL is a mixed physics ensemble which can enhance forecast diversity (Snook et al. 2012; Johnson and Wang 2017) but can also introduce more systematic biases into the forecast system than a single physics ensemble (e.g., Romine et al. 2013). Future studies should also evaluate forecast sensitivity to conventional observation (ground-based observations, soundings) DA parameters because these observations modify atmospheric environmental conditions and can substantially impact storm evolution (e.g., Sobash and Stensrud 2015; Snook et al. 2015). It is noted the conclusions of this work is limited to a single case study, additional experiments are required to ensure results are robust for a variety of cases. Even with these limitations, results provide useful insights into the development of future ensemble DA systems.

Chapter 5 Summary and Future Work

5.1 Dissertation summary

Designing a convection-allowing model (CAM) ensemble forecast system that skillfully predicts the evolution of convective storms represents a substantial challenge because the non-linear growth of small initial condition and model errors alters thunderstorm evolution and degrades forecast skill. To improve forecast skill, sources of model error must be identified and minimized. For example, microphysics (MP) schemes are a major source of model error because they simulate poorly understood microphysical processes and make simplifying assumptions to remain computationally feasible. Initial condition errors can also limit forecast skill. Data assimilation systems can improve initial condition estimates; however, their skill is dependent upon many factors including experimental design. This dissertation evaluates the skill of forecasts run using different MP schemes and data assimilation configurations and provides insights relevant to the development of a skilled CAM ensemble forecast system.

Results evaluate the skill of three case studies including supercell thunderstorms near Oklahoma City on 19 May 2013, mountain-initiated hailstorms that impact Denver on 8 May 2017, and a mesoscale convective system (MCS) event on 28 – 29 May 2017 that occurred in the Southern United States. All experiments assimilate radar and convectional observations using an ensemble Kalman filter (EnKF) framework and the forecast model employs advanced microphysics schemes including the NSSL double-moment variable density rimed ice scheme (NSSL) as well as the Milbrandt and Yau double-moment (MY2) and triple-moment (MY3) schemes. Forecast and analysis skill is evaluated using simulated radar variables, maximum estimated hail size (D_{max}) diagnosed from rimed ice particle size distributions (PSDs), microphysical process budgets, and ensemble correlation analyses.

An ensemble of high-resolution hail size forecasts run using the MY2, MY3, and NSSL schemes are verified against hydrometeor classification output for a supercell hailstorm event that occurred near Oklahoma City on 19 May 2013. Probabilistic forecasts show ensembles run using all three MP schemes skillfully predict the coverage of severe hail. Although reliability diagrams show the MY2 and MY3 schemes skillfully predict the coverage of significant severe hail (D_{max}> 50 mm), both ensembles predict hail to be too much larger (D_{max} > 150 mm) than surface-based reports (~ 75 mm). NSSL deterministic hail size forecasts qualitatively resemble HCA output more closely than the other ensembles, but forecasts underpredict the coverage of significant severe hail (D_{max} > 50 mm).

Hail production tendencies are evaluated during the forecast period for each MP scheme to determine the hail growth and decay mechanisms that cause surface hail size forecast biases. Both MY schemes predict storms to produce many small hailstones above the 0°C isotherm that are created from frozen raindrops. These storms produce large plumes of hail aloft because the small hailstones are advected more than 40 km downstream of the updraft. The NSSL scheme predicts storms to produce fewer, but larger hail particles above the 0 °C isotherm because the hail category consists of dense graupel that has undergone wet growth. Beneath the 0 °C isotherm, the MY2 and MY3 schemes predict hail to accrete water, which causes model predicted hail to exceed 150 mm in diameter near the surface. The MY3 scheme can increase the hail shape parameter, this causes the PSD to narrow and often limits the most extreme hail sizes (D_{max} > 150 mm). Extreme D_{max} values decrease in both MY2 and MY3 forecasts when the schemes no longer allow hail to accrete rainwater beneath the 0 °C isotherm. The NSSL scheme predicts all accreted water to be shed beneath the 0 °C isotherm; however, this is not realistic when compared to

observational studies (e.g., Rasmussen et al. 1984) and potentially limits hail growth due to the collection of water (Lesins and List 1986).

In the third chapter of this dissertation, a cycled EnKF framework updates hailstorm microphysical properties for the 8 May 2017 Colorado severe hail event. Four experiments are conducted. Each is run using the MY2 or NSSL schemes, where assimilated Z observations update either only hydrometeor mixing ratio (MY_Q, NSSL_Q) or all microphysical state variables (MY_ALL, NSSL_ALL), in addition to thermodynamic and dynamic information. Analysis and forecast Z and D_{max} are verified against observed Z and HCA output, respectively, to determine which configuration produces the most realistic state variable estimates.

Experiments that update only hydrometeor mixing ratio (i.e., MY-Q), often cause the error covariance to become unreliable and limit analysis skill. For example, negative correlations between the observation prior Z and rain mixing ratios causes the EnKF to intensify spurious Z. The forecast error covariance is more reliable for MY_ALL and the NSSL ensembles during data assimilation, all have relatively small Z and radial velocity root mean square innovations. Ensembles that update only hydrometeor mixing ratio (MY_Q, NSSL_Q) estimate large surface hail because the EnKF can distribute large quantities of hail mass amongst a relatively small number of hailstones. MY_ALL and NSSL_ALL underestimate hail size because the EnKF updates the hail PSD so that mass is distributed amongst many more hailstones. Results suggest additional observations of microphysical relevance are needed to constrain the microphysical state variables.

The cross-correlation between observation priors (i.e., Z and updraft intensity) at an assumed observation location and model state variables is analyzed. These correlation analyses show the background error covariance is sensitive to the rimed ice growth assumptions made

within each MP scheme. Hail predicted by the MY2 scheme is composed of small frozen raindrops. This process causes the mean size of hail above the 0 °C isotherm to decrease in strong storms and causes Z to become negative correlated with both updraft intensity and air temperature. Since the NSSL scheme predicts strong storms to produce large hail above the 0 °C isotherm, the opposite correlation patterns are observed. Hail growth assumptions also influence how the EnKF updates hydrometeor state variables. Generally, correlations between Z and hail state variables are more positive for the NSSL experiment than the MY2 experiment. Differences in the multivariate covariance show that under similar circumstances model state variables can be updated in different directions if different MP schemes are used.

Finally, in the fourth chapter of this dissertation, the 2019 CAPS storm-scale ensemble forecast gridpoint statistical interpolation (GSI) EnKF system is used to assimilate observations (conventional and radar) and initialize short-term (0 – 6 hour) CAM forecasts for a mesoscale convective system (MCS) event that occurred in the Southern United States on 28 – 29 May 2017. Forecasts predict the MCS with skill during the first three forecast hours but exhibit less skill later because the predicted storm system moves too quickly. Forecasts predict small, isolated convection located near the MCS with less skill because the storms near the domain boundary weaken too quickly, and nearby spurious convection is predicted.

Experiments also evaluate forecast sensitivity to the configuration of GSI EnKF system. Analyzed parameters include ensemble initialization and spread inflation as well as radar observation data thinning, covariance localization radius, observation errors, and data assimilation frequency. Results show errors that originate from ensemble initialization can alter EnKF analyses and degrade forecast skill. When the ensemble is initialized from 1500 UTC SREF, instead of 1800 UTC NAM with SREF perturbations (default configuration), the predicted MCS weakens

too quickly because the storm inflow environment is relatively cool and more stable. The NAM initialized ensemble predicts the MCS to be stronger and more closely resemble observations Modifying the radar data assimilation parameters has a relatively small impact on forecast skill during the first forecast hour. Ensembles that assimilate more radar observations (i.e., less thinning of data) or increase the horizontal covariance localization radius for radar observations predict higher *Z* values near observed storms but also predict more spurious convection. Both ensembles predict more spurious convection occurs because the ensemble analyses contain more storm updrafts that quickly intensify because the storm inflow environment is relatively moist and unstable. Similar trends are also observed when radar observations are assimilated more frequently. Experiments that assimilate radar observations more frequently (i.e., 5 minutes) predict high *Z* values near observed storms but exhibit a substantial decrease in skill for the first three forecast hours because of widespread spurious convection. Forecast skill is relatively insensitive to the range of ensemble spread inflation factors and observation errors that were tested during this study.

5.2 Future work

The studies in this dissertation demonstrate that no single experiment configuration can produce a skilled CAM ensemble forecast system for all event types because forecast skill is dependent upon many factors including horizontal grid spacing, other employed physics options, event type, and data assimilation system. Although no optimal configuration exists, forecast skill can be improved by choosing model physics/microphysics options (e.g., MP schemes) that better represent observed sub-grid scale processes and a data assimilation system that uses reliable multivariate covariances to update the many unobserved model state variables.

Due to the sensitivity of forecast skill to MP schemes, model physics must be improved to better represent physical processes. For example, hail size forecasts produced using the NSSL MP scheme qualitatively exhibit more skill than those run using the MY2 and MY3 schemes because the NSSL scheme prognoses total volume for two rimed ice categories. Having the ability to diagnose hydrometeor volume for two rimed ice categories allows the NSSL scheme to better represent the growth of large and dense hail through the wet growth of graupel. Since the hail size forecasts produced using the NSSL scheme exhibits more skill than the MY3 scheme, it can also be inferred predicting hydrometeor characteristics is equally important as predicting additional moments of the hydrometeor PSD.

Future studies should evaluate MP schemes that prognose hydrometeor characteristics to determine which methods improve forecast skill and guide future parameterization development. For example, the predicted particle properties (P3) scheme (Morrison and Milbrandt 2015) predicts rime fraction and rime volume in addition to ice mass and number concentration. Although the scheme represents multiple hydrometeor types within a single category, it rarely predicts large fully-rimed ice (i.e., hail) because the scheme employs a maximum ice number-weighted mean diameter (Johnson et al. 2019) and the ice category is often dominated by small partially rimed ice (Labriola et al. 2019b). This demonstrates that predicting additional hydrometeor characteristics does not necessarily improve skill and that careful calibration is required so that MP schemes can accurately represent diverse event types (e.g., hailstorms, MCS, stratiform precipitation). Other recently developed MP schemes that should be investigated under a diverse number of cases predict hydrometeor characteristics such as water fraction (Cholette et al. 2019) and axis ratio (Jensen et al. 2017).

Although the case studies in this dissertation provide insight into hail growth and decay mechanisms, additional microphysically relevant observations are needed to verify these simulated processes. Few case studies directly observe in-storm hydrometeors (e.g., Waugh et al. 2015) because in-storm observations are too costly and dangerous to collect. Remotely sensed radar observations are instead used to verify forecasts. Z provides important information on the largest hydrometeors present within a volume; however, analyses are limited because the signature can be recreated by a non-unique combination of hydrometeor moments. Future studies must rely more heavily upon simulated polarimetric radar products to infer more hydrometeor properties including shape, size, orientation, and phase (Doviak et al. 2000; Kumjian and Ryzhkov 2008). Recent studies (e.g., Jung et al. 2012; Johnson et al. 2016; Putnam et al. 2017b) evaluate MP scheme rimed ice treatments using simulated polarimetric fields such as differential reflectivity (Z_{dr}) and cross-correlation coefficient (ρ_{HV}). Aside from verification, these fields should be used to correct for biases within schemes (e.g., Johnson et al. 2019). Such as done in this study with verifying hail size forecasts, additional novel methods of forecast verification should be investigated to gain additional insights into forecast biases (e.g., hail pads, lightning, satellite observations).

As MP schemes grow in complexity to represent the diverse range of hydrometeor types found in convective storms, initializing CAM forecasts becomes more difficult. Results of this work show assimilating observed Z can improve initial condition estimates; however, hydrometeor state variables remain under constrained which can prevent hydrometeor PSDs from being properly initialized (e.g., Xue et al. 2010). Additional microphysically relevant observations are needed to correctly initialize model state variables. Polarimetric radar observations can potentially be assimilated via an EnKF; Putnam et al. (2019) successfully assimilated Z_{dr} in the lowest 2 km of the atmosphere. Although assimilation of Z_{dr} improves microphysical moment estimation, large

uncertainties in the treatment of mixed-phase and frozen hydrometeors in the radar emulator prevents the assimilation of observations further aloft. To get the full benefit of polarimetric radar data assimilation, the treatment of melting must be better understood and simulated, such as developing MP schemes that explicitly predict the water fraction (e.g., Ferrier 1994; Cholette et al. 2019).

While improvements can be made to the weather forecast systems, operational CAM ensemble forecasts will continue to be run at horizontal grid spacings that are unable to resolve severe weather hazards (e.g., tornado, severe wind, maximum hail size). Post-processing techniques must be improved to extract as much useful information as possible from model output to determine where potential severe weather hazards are located. In this dissertation hail size forecasts are diagnosed from the model particle size distributions using a method similar to the Thompson hail method (Gagne et al. 2019) and convective storms are verified by identifying regions of high Z (> 40 dBZ). Although these method produces skilled forecasts, they use limited information and are sensitive to arbitrary thresholds. Similar problems are observed when using updraft helicity (Kain et al. 2008). Although this storm-surrogate variable highlights regions of potential severe weather (e.g., Sobash et al. 2011), the technique is subject to high false alarm rates, favors rotating thunderstorms, and relies upon subjective thresholds to delineate between non-severe and severe events (e.g., Gagne et al. 2017). New methods such as machine learning models can be used to infer trends from model output and determine the probability of a severe event occurrence while correcting biases (e.g., intensity, location) that degrade forecast skill. Previous studies have shown machine learning models can be designed to skillfully predict severe weather hazards such as strong winds, heavy rainfall, and hail (Gagne et al. 2014; Gagne et al. 2017; Lagerquist et al. 2017; Burke et al. 2020).

This dissertation presents the results of only three case studies. While the results provide useful insight into the performance of the different microphysics schemes and data assimilation configurations, additional work is needed to determine the results are robust and remain the same under a wide variety of atmospheric conditions. Evaluating forecasts that predict many different event types (e.g., supercell, disorganized storm cells, MCS) that occur in different environments will provide more insight into forecast sensitivities.

References

Adams-Selin, R. D., and C. L. Ziegler, 2016: Forecasting Hail Using a One-Dimensional Hail Growth Model within WRF. *Mon. Weather Rev.*, **144**, 4919–4939, https://doi.org/10.1175/MWR-D-16-0027.1.

Aksoy, A., F. Zhang, and J. W. Nielsen-gammon, 2006: Ensemble-based simultaneous state and parameter estimation with MM5. *Geophys. Res. Lett.*, **33**, L12801, https://doi.org/10.1029/2006GL026186.

——, D. C. Dowell, and C. Snyder, 2009: A multicase comparative assessment of the ensemble Kalman filter for assimilation of radar observations. Part I: Storm-scale analyses. *Mon. Weather Rev.*, **137**, 1805–1824, https://doi.org/10.1175/2008MWR2691.1.

——, S. Lorsolo, T. Vukicevic, K. J. Sellwood, S. D. Aberson, and F. Zhang, 2012: The HWRF Hurricane Ensemble Data Assimilation System (HEDAS) for high-resolution data: The impact of airborne doppler radar observations in an OSSE. *Mon. Weather Rev.*, **140**, 1843–1862, https://doi.org/10.1175/MWR-D-11-00212.1.

Allen, J. T., and M. K. Tippett, 2015: The Characteristics of United States Hail Reports : 1955 – 2014. *Electron. J. Sev. Storms Meteorol.*, **10**, 1–31.

Anderson, J. L., 2001: An Ensemble Adjustment Kalman Filter for Data Assimilation. *Mon. Weather Rev.*, **129**, 2884–2903, https://doi.org/10.1175/1520-0493(2001)129<2884:AEAKFF>2.0.CO;2.

——, and S. L. Anderson, 1999: A Monte Carlo Implementation of the Nonlinear Filtering Problem to Produce Ensemble Assimilations and Forecasts. *Mon. Weather Rev.*, **127**, 2741–2758.

——, and N. Collins, 2007: Scalable implementations of ensemble filter algorithms for data assimilation. *J. Atmos. Ocean. Technol.*, **24**, 1452–1463, https://doi.org/10.1175/JTECH2049.1.

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

Brewster, K. A., M. Hu, M. Xue, and J. Gao, 2005: Efficient Assimilation of Radar Data at High Resolution for Short-Range Numerical Weather Prediction. *Int. Symp. on Nowcasting Very Short Range Forecasting*, Toulouse, France, WMO/World Weather Research Program, 3.06.

Brier, G. W., 1950: Verification of forecasts expersses in terms of probaility. *Mon. Weather Rev.*, **78**, 1–3, https://doi.org/10.1126/science.27.693.594.

Bringi, V. N., and V. Chandrasekar, 2001: *Polarimetric Doppler Weather Radar*. Cambridge, 636 pp.

Brooks, H. E., 2004: Tornado-warning performance in the past and future. *Bull. Am. Meteorol. Soc.*, **85**, 837–843, https://doi.org/10.1175/BAMS-85-6-837.

Buizza, R., A. Hollingsworth, F. Lalaurette, and A. Ghelli, 1999: Probabilistic Predictions of Precipitation Using the ECMWF Ensemble Prediction System. *Weather Forecast.*, **14**, 168–189.

Bunkers, M. J., B. A. Kilmowski, J. W. Zeitler, R. L. Thompson, and M. L. Weisman, 2000: Predicting Supercell Motion Using a New Hodograph Technique. *Weather Forecast.*, **15**, 61–79.

Burke, A., N. Snook, D. J. Gagne II, S. McCorkle, and A. McGovern, 2020: Calibration of

Machine Learning–Based Probabilistic Hail Predictions for Operational Forecasting. *Weather Forecast.*, **35**, 149–168, https://doi.org/10.1175/waf-d-19-0105.1.

CAPS, 2013: *2013 CAPS Spring Forecast Experiment Program Plan*. 24 pp. http://forecast.caps.ou.edu/SpringProgram2013_Plan-CAPS.pdf.

——, 2017: 2017 CAPS Spring Forecast Experiment Program Plan. Center for The Analysis and Prediction of Storms. http://www.caps.ou.edu/~fkong/sub_atm/spring17.html.

Carlin, J. T., J. Gao, J. C. Snyder, and A. V. Ryzhkov, 2017: Assimilation of ZDR Columns for Improving the Spin-Up and Forecast of Convective Storms in Storm-Scale Models: Proof-of-Concept Experiments. *Mon. Weather Rev.*, 5033–5057, https://doi.org/10.1175/MWR-D-17-0103.1.

Caya, A., J. Sun, and C. Snyder, 2005: A Comparison between the 4DVAR and the Ensemble Kalman Filter Techniques for. *Mon. Weather Rev.*, **133**, 3081–3094.

Chang, W., K.-S. Chung, L. Fillion, and S.-J. Baek, 2014: Radar data assimilation in the Canadian high-resolution ensemble Kalman filter system: Performance and verification with real summer cases. *Mon. Weather Rev.*, **142**, 2118–2138, https://doi.org/10.1175/MWR-D-13-00291.1.

Changnon, S. A., 2009: Increasing major hail losses in the U.S. *Clim. Change*, **9**, 161–166, https://doi.org/10.1007/s10584-009-9597-z.

Changnon, S. A., D. Changnon, and S. D. Hilberg, 2009: *Hailstorms Across the Nation An Atlast about Hail and Its Damages*.

Chen, F., and J. Dudhia, 2001: Coupling an Advanced Land Surface- Hydrology Model with the Penn State NCAR MM5 Modeling System. Part I: Model Implementation and Sensitivity.

Mon. Weather Rev., **129**, 569–585.

Chipilski, H. G., X. Wang, and D. B. Parsons, 2020: Impact of Assimilating PECAN Profilers on the Prediction of Bore-Driven Nocturnal Convection: A Multiscale Forecast Evaluation for the 6 July 2015 Case Study. *Mon. Weather Rev.*, **148**, 1147–1175, https://doi.org/10.1175/mwr-d-19-0171.1.

Cholette, M., H. Morrison, J. A. Milbrandt, and J. M. Thériault, 2019: Parameterization of the bulk liquid fraction on mixed-phase particles in the predicted particle properties (P3) Scheme: Description and idealized simulations. *J. Atmos. Sci.*, **76**, 561–582, https://doi.org/10.1175/JAS-D-18-0278.1.

Chong, S.-L., and C. S. Chen, 1974: Water Shells on Ice Pellets and Hailstones. *J. Atmos. Sci.*, **31**, 1384–1391.

Chou, M.-D., 1990: Parameterizations for the Absorption of Solar Radiation by O2 and CO2 with Application to Climate Studies. *J. Clim.*, **3**, 209–217, https://doi.org/10.1175/1520-0442(1990)003<0209:pftaos>2.0.co;2.

——, 1992: A Solar Radiation Model For Use in Climate Studies. *J. Atmos. Sci.*, **49**, 762–772.

——, and M. Suarez, 1994: *An effcient thermal infrared radiation parameterization for use in general circulation models*. 85 pp.

Cintineo, J. L., T. M. Smith, V. Lakshmanan, H. E. Brooks, and K. L. Ortega, 2012: An Objective High-Resolution Hail Climatology of the Contiguous United States. *Weather Forecast.*, **27**, 1235–1248, https://doi.org/10.1175/WAF-D-11-00151.1.

Clark, A. J., J. S. Kain, P. T. Marsh, J. Correia, M. Xue, and F. Kong, 2012: Forecasting Tornado Pathlengths Using a Three-Dimensional Object Identification Algorithm Applied to

Convection-Allowing Forecasts. *Weather Forecast.*, **27**, 1090–1113, https://doi.org/10.1175/WAF-D-11-00147.1.

——, and Coauthors, 2018: The community leveraged unified ensemble (CLUE) in the 2016 NOAA/hazardous weather testbed spring forecasting experiment. *Bull. Am. Meteorol. Soc.*, **99**, 1433–1448, https://doi.org/10.1175/BAMS-D-16-0309.1.

——, and Coauthors, 2019: *SPRING FORECASTING EXPERIMENT 2019*. 1–39 pp. https://hwt.nssl.noaa.gov/sfe/2019/docs/HWT_SFE2019_operations_plan.pdf.

Cohen, A. E., S. M. Cavallo, M. C. Coniglio, and H. E. Brooks, 2015: A review of planetary boundary layer parameterization schemes and their sensitivity in simulating southeastern U.S. cold season severe weather environments. *Weather Forecast.*, **30**, 591–612, https://doi.org/10.1175/WAF-D-14-00105.1.

——, ——, ——, ——, and I. L. Jirak, 2017: Evaluation of multiple planetary boundary layer parameterization schemes in southeast U.S. cold season severe thunderstorm environments. *Weather Forecast.*, **32**, 1857–1884, https://doi.org/10.1175/WAF-D-16-0193.1.

Crum, T. D., R. L. Alberty, and D. W. Burgess, 1993: Recording, archiving, and using WSR-88D data. *Bull. Am. Meteorol. Soc.*, **74**, 645–653, https://doi.org/10.1175/1520-0477(1993)074<0645:raauwd>2.0.co;2.

Curtis, C. D., and S. M. Torres, 2011: Adaptive range oversampling to achieve faster scanning on the National Weather Radar Testbed phased-array radar. *J. Atmos. Ocean. Technol.*, **28**, 1581–1597, https://doi.org/10.1175/JTECH-D-10-05042.1.

Dawson, D. T. I., M. Xue, J. A. Milbrandt, and M. K. Yau, 2010: Comparison of Evaporation and Cold Pool Development between Single-Moment and Multimoment Bulk Microphysics

Schemes in Idealized Simulations of Tornadic Thunderstorms. *Mon. Weather Rev.*, **138**, 1152–1171, https://doi.org/10.1175/2009MWR2956.1.

——, L. J. Wicker, E. R. Mansell, and R. L. Tanamachi, 2012: Impact of the Environmental Low-Level Wind Profile on Ensemble Forecasts of the 4 May 2007 Greensburg, Kansas, Tornadic Storm and Associated Mesocyclones. *Mon. Weather Rev.*, **140**, 696–716, https://doi.org/10.1175/MWR-D-11-00008.1.

——, E. R. Mansell, Y. Jung, L. J. Wicker, M. R. Kumjian, and M. Xue, 2014: Low-Level Z DR Signatures in Supercell Forward Flanks: The Role of Size Sorting and Melting of Hail. *J. Atmos. Sci.*, **71**, 276–299, https://doi.org/10.1175/JAS-D-13-0118.1.

Dawson, L. C., G. S. Romine, R. J. Trapp, and M. E. Baldwin, 2017: Verifying Supercellular Rotation in a Convection-Permitting Ensemble Forecasting System with Radar-Derived Rotation Track Data. *Weather Forecast.*, **32**, 781–795, https://doi.org/10.1175/waf-d-16-0121.1.

Deardorff, J. W., 1980: Stratocumulus-Capped Mixed Layers Derived From a Three-Dimensional Model. *Boundary-Layer Meteorol.*, **18**, 495–527, https://doi.org/10.1007/BF00119502.

Dennis, E. J., and M. R. Kumjian, 2017: The Impact of Vertical Wind Shear on Hail Growth in Simulated Supercells. *J. Atmos. Sci.*, **74**, 641–663, https://doi.org/10.1175/JAS-D-16-0066.1.

Doswell, C. A., H. E. Brooks, and M. P. Kay, 2005: Climatological Estimates of Daily Local Nontornadic Severe Thunderstorm Probability for the United States. *Weather Forecast.*, **20**, 577–595, https://doi.org/10.1175/WAF866.1.

Doviak, R., and D. S. Zrnić, 1993: *No Doppler Radar and Weather Observations*. 2nd ed. Academic Press, 562 pp.

Doviak, R. J., V. Bringi, A. Ryzhkov, A. Zahrai, and D. Zrnić, 2000: Considerations for polarimetric upgrades to operational WSR-88D radars. *J. Atmos. Ocean. Technol.*, **17**, 257–278, https://doi.org/10.1175/1520-0426(2000)017<0257:CFPUTO>2.0.CO;2.

Dowell, D., 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. *Mon. Weather Rev.*, **139**, 272–294, https://doi.org/10.1175/2010MWR3438.1.

Dowell, D. C., and L. J. Wicker, 2009: Additive noise for storm-scale ensemble data assimilation. *J. Atmos. Ocean. Technol.*, **26**, 911–927, https://doi.org/10.1175/2008JTECHA1156.1.

——, F. Zhang, L. J. Wicker, C. Snyder, and N. A. Crook, 2004: Wind and Temperature Retrievals in the 17 May 1981 Arcadia, Oklahoma, Supercell: Ensemble Kalman Filter Experiments. *Mon. Weather Rev.*, **132**, 1982–2005, https://doi.org/10.1175/1520-0493(2004)132<1982:WATRIT>2.0.CO;2.

Du, J., G. Dimego, B. Zhou, D. Jovic, B. Ferrier, and B. Yang, 2015: Regional ensemble forecast systems at NCEP. Preprints,. *27th Conf. on Weather Analysis and Forecasting/23rd Conf. on Numerical Weather Prediction*, Chicago, IL, Amer. Meteor. Soc., 2A.5, https://ams.confex.com/ams/27WAF23NWP/webprogram/Manuscript/Paper273421/NWP2015_NCEP_RegionalEnsembles_paper.pdf.

Duda, J. D., X. Wang, F. Kong, and M. Xue, 2014: Using varied microphysics to account for uncertainty in warm-season QPF in a convection-allowing ensemble. *Mon. Weather Rev.*, **142**, 2198–2219, https://doi.org/10.1175/MWR-D-13-00297.1.

Ebert, E. E., 2001: Ability of a Poor Man's Ensemble to Predict the Probability and Distribution

of Precipitation. *Mon. Weather Rev.*, **129**, 2461–2480, https://doi.org/10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2.

Edwards, R., and R. L. Thompson, 1998: Nationwide Comparisons of Hail Size with WSR-88D Vertically Integrated Liquid Water and Derived Thermodynamic Sounding Data. *Weather Forecast.*, **13**, 277–285.

Elmore, K. L., Z. L. Flamig, V. Lakshmanan, B. T. Kaney, V. Farmer, H. D. Reeves, and L. P. Rothfusz, 2014: MPING: Crowd-sourcing weather reports for research. *Bull. Am. Meteorol. Soc.*, **95**, 1335–1342, https://doi.org/10.1175/BAMS-D-13-00014.1.

Environmental Modeling Center, 2017: North American Mesoscale (NAM) Analysis and Forecast System Characteristics. 9 pp.

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

——, 2003: The Ensemble Kalman Filter : theoretical formulation and practical implementation. *Ocean Dyn.*, **53**, 343–367, https://doi.org/10.1007/s10236-003-0036-9.

Ferrier, B. S., 1994: A Double-Moment Multiple-Phase Four-Class Bulk Ice Scheme. Part I: Description. *J. Atmos. Sci.*, **51**, 249–280, https://doi.org/10.1175/1520-0469(1994)051<0249:ADMMPF>2.0.CO;2.

Foote, G. B., 1984: A Study of Hail Growth Utilizing Observed Storm Conditions. *J. Clim. Appl. Meteorol.*, **23**, 84–101.

Fritz, A., 2017: Remember that crazy hailstorm in Denver? It's going to be Colorado's costliest catastrophe. *Washington Post*,. https://www.washingtonpost.com/news/capital-weather-gang/wp/2017/05/25/remember-that-crazy-hail-storm-in-denver-its-going-to-be-colorados-

costliest-catastrophe/?utm_term=.1d3f823248f4 (Accessed May 25, 2017).

Gagne, D. J., A. Mcgovern, and M. Xue, 2014: Machine learning enhancement of storm-scale ensemble probabilistic quantitative precipitation forecasts. *Weather Forecast.*, **29**, 1024–1043, https://doi.org/10.1175/WAF-D-13-00108.1.

——, S. E. Haupt, D. W. Nychka, and G. Thompson, 2019: Interpretable Deep Learning for Spatial Analysis of Severe Hailstorms. *Mon. Weather Rev.*, **14**, 2827-, https://doi.org/10.1175/MWR-D-18-0316.1.

Gagne II, D. J., A. Mcgovern, J. Brotzge, M. Coniglio, J. Correia Jr., and M. Xue, 2015: Day-Ahead Hail Prediction Integrating Machine Learning with Storm-Scale Numerical Weather Models. *27th Conf. on Innovative Applications of Artificial Intelligence*, Austin, TX, 3954–3960.

——, ——, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017: Storm-Based Probabilistic Hail Forecasting with Machine Learning Applied to Convection-Allowing Ensembles. *Weather Forecast.*, **32**, 1819–1840, https://doi.org/10.1175/WAF-D-17-0010.1.

Gallo, B. T., A. J. Clark, B. T. Smith, R. L. Thompson, I. Jirak, and S. R. Dembek, 2019: Incorporating UH occurrence time to ensemble-derived tornado probabilities. *Weather Forecast.*, **34**, 151–164, https://doi.org/10.1175/WAF-D-18-0108.1.

Gallus, W. A., and J. F. Bresch, 2006: Comparison of impacts of WRF dynamic core, physics package, and initial conditions on warm season rainfall forecasts. *Mon. Weather Rev.*, **134**, 2632–2641, https://doi.org/10.1175/MWR3198.1.

Gao, J., and M. Xue, 2008: An efficient dual-resolution approach for ensemble data assimilation and tests with simulated Doppler radar data. *Mon. Weather Rev.*, **136**, 945–963,

https://doi.org/10.1175/2007MWR2120.1.

Gaspari, G., and S. E. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Q. J. R. Meteorol. Soc.*, 723–757.

Hamill, T. M., J. S. Whitaker, and C. Snyder, 2001: Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Weather Rev.*, **129**, 2776–2790, https://doi.org/10.1175/1520-0493(2001)129<2776:DDFOBE>2.0.CO;2.

Hawblitzel, D. P., F. Zhang, Z. Meng, and C. A. Davis, 2007: Probabilistics evaluation of the dynamics and predictability of the mesoscale convective vortex of 10-13 June 2003. *Mon. Weather Rev.*, **135**, 1544–1563, https://doi.org/10.1175/MWR3346.1.

Heinselman, P. L., and A. V. Ryzhkov, 2006: Validation of Polarimetric Hail Detection. *Weather Forecast.*, **21**, 839–850, https://doi.org/10.1175/WAF956.1.

——, and S. M. Torres, 2011: High-temporal-resolution capabilities of the National Weather Radar Testbed Phased-Array Radar. *J. Appl. Meteorol. Climatol.*, **50**, 579–593, https://doi.org/10.1175/2010JAMC2588.1.

Herman, G. R., and R. S. Schumacher, 2018a: Money doesn't grow on trees, but forecasts do: Forecasting extreme precipitation with random forests. *Mon. Weather Rev.*, **146**, 1571–1600, https://doi.org/10.1175/MWR-D-17-0250.1.

——, and ——, 2018b: "Dendrology" in numerical weather prediction: What random forests and logistic regression tell us about forecasting extreme precipitation. *Mon. Weather Rev.*, **146**, 1785–1812, https://doi.org/10.1175/MWR-D-17-0307.1.

Heymsfield, A. J., 1983: Case Study of a Halistorm in Colorado. Part IV: Graupel and Hail Growth Mechanisms Deduced through Particle Trajectory Calculations. *J. Atmos. Sci.*, **6**, 1482–1509,

https://doi.org/http://dx.doi.org/10.1175/1520-0469(1983)040%3C1482:CSOAHI%3E2.0.CO;2.

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

Houtekamer, P. L., and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Weather Rev.*, **126**, 796–811, https://doi.org/10.1175/1520-0493(1998)126<0796:DAUAEK>2.0.CO;2.

——, ——, G. Pellerin, M. Buehner, L. Spacek, and B. Hansen, 2005: Atmospheric Data Assimilation with an Ensemble Kalman Filter: Results with Real Observations. *Mon. Weather Rev.*, **133**, 604–620.

Hu, M., and M. Xue, 2007: Impact of configurations of rapid intermittent assimilation of WSR-88D radar data for the 8 May 2003 Oklahoma City tornadic thunderstorm case. *Mon. Weather Rev.*, **135**, 507–525, https://doi.org/10.1175/MWR3313.1.

Iacono, M. J., J. S. Delamere, E. J. Mlawer, M. W. Shephard, S. A. Clough, and W. D. Collins, 2008: Radiative forcing by long-lived greenhouse gases: Calculations with the AER radiative transfer models. *J. Geophys. Res. Atmos.*, **113**, 1–9, https://doi.org/10.1029/2008JD009944.

Janjić, Z. I., 1990: The step-mountain coordinate: physical package. *Mon. Weather Rev.*, **118**, 1429–1443, https://doi.org/10.1175/1520-0493(1990)118<1429:TSMCPP>2.0.CO;2.

——, 1996: The surface layer in the NCEP Eta Model. *11th Conf. on Numerical Weather Prediction*, Norfolk, VA, Amer. Meteor. Soc., 354–355.

——, 2001: *Nonsingular implementation of the Mellor–Yamada level 2.5 scheme in the NCEP*

Meso Model. 61 pp.

Jazwinski, A. H., 1970: *Stochastic Processes and Filtering Theory*. Academic Press, 376 pp.

Jensen, A. A., J. Y. Harrington, H. Morrison, and J. A. Milbrandt, 2017: Predicting Ice Shape Evolution in a Bulk Microphysics Model. *J. Atmos. Sci.*, **74**, 2081–2104, https://doi.org/10.1175/JAS-D-16-0350.1.

——, ——, and ——, 2018: Microphysical Characteristics of Squall-Line Stratiform Precipitation and Transition Zones Simulated Using an Ice Particle Property-Evolving Model. *Mon. Weather Rev.*, **146**, 723–743, https://doi.org/10.1175/MWR-D-17-0215.1.

Johnson, A., and X. Wang, 2017: Design and Implementation of a GSI-Based Convection-Allowing Ensemble Data Assimilation and Forecast System for the PECAN Field Experiment. Part I: Optimal Configurations for Nocturnal Convection Prediction Using Retrospective Cases. *Weather Forecast.*, **32**, 289–315, https://doi.org/10.1175/waf-d-16-0102.1.

——, ——, J. R. Carley, L. J. Wicker, and C. Karstens, 2015: A comparison of multiscale GSI-based EnKF and 3DVar data assimilation using radar and conventional observations for midlatitude convective-scale precipitation forecasts. *Mon. Weather Rev.*, **143**, 3087–3108, https://doi.org/10.1175/MWR-D-14-00345.1.

——, ——, and S. Degelia, 2017: Design and implementation of a GSI-based convection-allowing ensemble-based data assimilation and forecast system for the PECAN field experiment. Part II: Overview and evaluation of a real-time system. *Weather Forecast.*, **32**, 1227–1251, https://doi.org/10.1175/WAF-D-16-0201.1.

Johnson, M., Y. Jung, D. T. I. Dawson, and M. Xue, 2016: Comparison of Simulated Polarimetric

Signatures in Idealized Supercell Storms using Two-moment Bulk Microphysics Schemes in WRF. *Mon. Weather Rev.*, **144**, 971–996, https://doi.org/10.1175/MWR-D-15-0233.1.

——, ——, D. Dawson, T. Supinie, M. Xue, J. Park, and Y.-H. Lee, 2018: Evaluation of Unified Model Microphysics in High-resolution NWP Simulations Using Polarimetric Radar Observations: Evaluation of Unified Model Microphysics in High-resolution NWP Simulations Using Polarimetric Radar Observations. *Adv. Atmos. Sci*, **35**, 771–784, https://doi.org/10.1007/s00376-017-7177-0.

——, ——, J. A. Milbrandt, H. Morrison, and M. Xue, 2019: Effects of the Representation of Rimed Ice in Bulk Microphysics Schemes on Polarimetric Signatures. *Mon. Weather Rev.*, conditionally accepted.

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

Jones, T. A., K. H. Knopfmeier, D. M. Wheatley, G. J. 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.

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

——, X. Wang, P. Skinner, A. Johnson, and Y. Wang, 2018b: Assimilation of GOES-13 imager

clear-sky water vapor (6.5 μm) radiances into a warn-on-forecast system. *Mon. Weather Rev.*, **146**, 1077–1107, https://doi.org/10.1175/MWR-D-17-0280.1.

Jung, Y., M. Xue, G. Zhang, and J. M. Straka, 2008a: Assimilation of Simulated Polarimetric Radar Data for a Convective Storm Using the Ensemble Kalman Filter. Part I: Observation Operators for Reflectivity and Polarimetric Variables. *Mon. Weather Rev.*, **136**, 2228–2245, https://doi.org/10.1175/2007MWR2288.1.

——, ——, ——, and J. M. Straka, 2008b: Assimilation of Simulated Polarimetric Radar Data for a Convective Storm Using the Ensemble Kalman Filter. Part II: Impact of Polarimetric Data on Storm Analysis. *Mon. Weather Rev.*, **136**, 2246–2260, https://doi.org/10.1175/2007MWR2288.1.

——, ——, and ——, 2010a: Simulations of polarimetric radar signatures of a supercell storm using a two-moment bulk microphysics scheme. *J. Appl. Meteorol. Climatol.*, **49**, 146–163, https://doi.org/10.1175/2009JAMC2178.1.

——, ——, and ——, 2010b: Simultaneous estimation of microphysical parameters and the atmospheric state using simulated polarimetric radar data and an ensemble Kalman filter in the presence of an observation operator error. *Mon. Weather Rev.*, **138**, 539–562, https://doi.org/10.1175/2009MWR2748.1.

——, ——, and M. Tong, 2012: Ensemble Kalman Filter Analyses of the 29–30 May 2004 Oklahoma Tornadic Thunderstorm Using One- and Two-Moment Bulk Microphysics Schemes, with Verification against Polarimetric Radar Data. *Mon. Weather Rev.*, **140**, 1457–1475, https://doi.org/10.1175/MWR-D-11-00032.1.

——, ——, F. Kong, Y. Wang, K. W. Thomas, and F. Shen, 2014: CAPS real-time storm-scale

EnKF data assimilation and forecast system for the hazardous weather testbed spring experiment. *The 6th EnKF Workshop*, Buffalo, NY http://hfip.psu.edu/fuz4/EnKF2014/EnKF-day2/EnKF_2014_yjung.pdf.

——, and Coauthors, 2018a: Development of GSI-Based EnKF and Hybrid EnVar Data Assimilation Capabilities for Continental-Scale 3-km Convection-Permitting Ensemble Forecasting and Testing via NOAA Hazardous Weather Testbed Spring Forecasting Experiments. *29th Conference on Weather Analysis and Forecasting/25th Conference on Numerical Weather Prediction*, Denver, CO.

——, R. Kong, C. Liu, J. Luo, M. Xue, and G. Zhao, 2018b: CAPS Real-time Storm-Scale EnKF Data Assimilation and Forecasts for the NOAA Hazardous Weather Testbed Spring Forecasting Experiments: Towards the Goal of Operational Ensemble-Variational Data Assimilation for Convection Permitting Models. *The 8th EnKF Data Assimilation Workshop*, Sainte-Adèle, Québec, Canada.

Kain, J. S., and Coauthors, 2008: Some Practical Considerations Regarding Horizontal Resolution in the First Generation of Operational Convection-Allowing NWP. *Weather Forecast.*, **23**, 931–952, https://doi.org/10.1175/2008WAF2007106.1.

——, and Coauthors, 2010a: Assessing advances in the assimilation of radar data and other mesoscale observations within a collaborative forecasting-research environment. *Weather Forecast.*, **25**, 1510–1521, https://doi.org/10.1175/2010WAF2222405.1.

——, S. R. Dembek, S. J. Weiss, J. L. Case, J. J. Levit, and R. A. Sobash, 2010b: Extracting Unique Information from High-Resolution Forecast Models : Monitoring Selected Fields and Phenomena Every Time Step. *Weather Forecast.*, **25**, 1536–1542,

https://doi.org/10.1175/2010WAF2222430.1.

Klemp, J. B., and R. B. Wilhelmson, 1978: The Simulation of Three-Dimensional Convective Storm Dynamics. *J. Atmos. Sci.*, **35**, 1070–1096.

Kong, F., and Coauthors, 2009: A Real-Time Storm-Scale Ensemble Forecast System: 2009 Spring Experiment. *23rd Conf. on Weather Analysis and Forecasting/19th Conf. Num. Wea. Pred.*, Omaha, Nebraska, Amer. Meteor. Soc., 1–10.

Kumjian, M. R., and A. V. Ryzhkov, 2008: Polarimetric signatures in supercell thunderstorms. *J. Appl. Meteorol. Climatol.*, **47**, 1940–1961, https://doi.org/10.1175/2007JAMC1874.1.

——, and ——, 2012: The Impact of Size Sorting on the Polarimetric Radar Variables. *J. Atmos. Sci.*, **69**, 2042–2060, https://doi.org/10.1175/JAS-D-11-0125.1.

——, A. P. Khain, N. Benmoshe, E. Ilotoviz, A. V. Ryzhkov, and V. T. J. Phillips, 2014: The Anatomy and Physics of ZDR Columns: Investigating a Polarimetric Radar Signature with a Spectral Bin Microphysical Model. *J. Appl. Meteorol. Climatol.*, **53**, 1820–1843, https://doi.org/10.1175/JAMC-D-13-0354.1.

Labriola, J., N. A. Snook, Y. Jung, B. J. Putnam, and M. Xue, 2017: Ensemble Hail Prediction for the Storms of 10 May 2010 in South-Central Oklahoma Using Single- and Double-Moment Microphysical Schemes. *Mon. Weather Rev.*, https://doi.org/10.1175/MWR-D-17-0039.1.

——, N. Snook, Y. Jung, and M. Xue, 2019a: Explicit Ensemble Prediction of Hail in 19 May 2013 Oklahoma City Thunderstorms and Analysis of Hail Growth Pocesses with Several Multi-Moment Microphysics Schemes. *Mon. Weather Rev.*, **147**, 1193–1213, https://doi.org/10.1175/MWR-D-18-0266.1.

——, ——, M. Xue, and K. W. Thomas, 2019b: Forecasting the 8 May 2017 Severe Hail Storm

in Denver, Colorado, at a Convection-Allowing Resolution: Understanding Rimed Ice Treatments in Multimoment Microphysics Schemes and Their Effects on. *Mon. Weather Rev.*, **147**, 3045–3068.

Lagerquist, R., A. McGovern, and T. Smith, 2017: Machine learning for real-time prediction of damaging straight-line convective wind. *Weather Forecast.*, **32**, 2175–2193, https://doi.org/10.1175/WAF-D-17-0038.1.

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.

Lesins, G. B., and R. List, 1986: Sponginess and Drop Shedding of Gyrating Hailstones in a Pressure-Controll ed Icing Wind Tunnel. *J. Atmos. Sci.*, **43**, 2813–2825.

Lim, K.-S. S., and S.-Y. Hong, 2010: Development of an Effective Double-Moment Cloud Microphysics Scheme with Prognostic Cloud Condensation Nuclei (CCN) for Weather and Climate Models. *Mon. Weather Rev.*, **138**, 1587–1612, https://doi.org/10.1175/2009mwr2968.1.

Lin, Y.-L., R. D. Farley, and H. D. Orville, 1983: Bulk Parameterization of the Snow Field in a Cloud Model. *J. Clim. Appl. Meteorol.*, **22**, 1065–1092, https://doi.org/10.1175/1520-0450(1983)022<1065:BPOTSF>2.0.CO;2.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus A*, **1**.

Luo, L., M. Xue, K. Zhu, and B. Zhou, 2017: Explicit prediction of hail using multimoment microphysics schemes for a hailstorm of 19 March 2014 in eastern China. *J. Geophys. Res.*

Atmos., **122**, 7560–7581, https://doi.org/10.1002/2017JD026747.

——, ——, ——, and ——, 2018: Explicit Prediction of Hail in a Long-Lasting Multi-Cellular Convective System in Eastern China Using Multimoment Microphysics Schemes. *J. Atmos. Sci.*, **75**, 3115–3137.

Lynch, P., and X.-Y. Huang, 1992: Initialization of the HIRLAM model using a digital filter. *Mon. Weather Rev.*, **120**, 1019–1034, https://doi.org/10.1175/1520-0493(1992)120<1019:IOTHMU>2.0.CO;2.

Mansell, E. R., 2010: On Sedimentation and Advection in Multimoment Bulk Microphysics. *J. Atmos. Sci.*, **67**, 3084–3094, https://doi.org/10.1175/2010JAS3341.1.

——, C. Ziegler, and E. Bruning, 2010: Simulated Electrification of a Small Thunderstorm with Two-Moment Bulk Microphysics. *J. Atmos. Sci.*, **67**, 171–194, https://doi.org/10.1175/2009JAS2965.1.

Marsh, P. T., 2017: Mesoscale Discussion 664. *Storm Predict. Cent.*,. http://www.spc.noaa.gov/products/md/2017/md0664.html (Accessed August 5, 2017).

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

McGovern, A., K. L. Elmore, D. J. Gagne, S. E. Haupt, C. D. Karstens, R. Lagerquist, T. Smith, and J. K. Williams, 2017: Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bull. Am. Meteorol. Soc.*, **98**, 2073–2090, https://doi.org/10.1175/BAMS-D-16-0123.1.

Melhauser, C., and F. Zhang, 2012: Practical and intrinsic predictability of severe and convective weather at the mesoscales. *J. Atmos. Sci.*, **69**, 3350–3371, https://doi.org/10.1175/JAS-D-11-0315.1.

Milbrandt, J. A., and M. K. Yau, 2005a: A Multimoment Bulk Microphysics Parameterization. Part II: A Proposed Three-Moment Closure and Scheme Description. *J. Atmos. Sci.*, **62**, 3065–3081, https://doi.org/10.1175/JAS3535.1.

——, and ——, 2005b: A Multimoment Bulk Microphysics Parameterization. Part I: Analysis of the Role of the Spectral Shape Parameter. *J. Atmos. Sci.*, **62**, 3051–3064, https://doi.org/10.1175/JAS3534.1.

——, and ——, 2006: A Multimoment Bulk Microphysics Parameterization . Part III : Control Simulation. *J. Atmos. Sci.*, **63**, 3114–3136.

——, and H. Morrison, 2013: Prediction of Graupel Density in a Bulk Microphysics Scheme. *J. Atmos. Sci.*, **70**, 410–429, https://doi.org/10.1175/JAS-D-12-0204.1.

Milbrandt, J. A., and H. Morrison, 2016: Parameterization of Cloud Microphysics Based on the Prediction of Bulk Ice Particle Properties. Part III : Introduction of Multiple Free Categories. *J. Atmos. Sci.*, **73**, 975–995, https://doi.org/10.1175/JAS-D-15-0204.1.

Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *J. Geophys. Res. D Atmos.*, **102**, 16663–16682, https://doi.org/10.1029/97jd00237.

Moeng, C.-H., 1984: A Large-Eddy-Simulation Model for the Study of Planetary Boundary-Layer Turbulence. *J. Atmos. Sci.*, **41**, 2052–2062.

Morrison, H., and W. W. Grabowski, 2008: A Novel Approach for Representing Ice Microphysics in Models: Description and Tests Using a Kinematic Framework. *J. Atmos. Sci.*, **65**, 1528–1548, https://doi.org/10.1175/2007JAS2491.1.

——, and J. A. Milbrandt, 2015: Parameterization of Cloud Microphysics Based on the Prediction

of Bulk Ice Particle Properties. Part I: Scheme Description and Idealized Tests. *J. Atmos. Sci.*, **72**, 287–311, https://doi.org/10.1175/JAS-D-14-0065.1.

Morrison, H., J. a. Curry, M. D. Shupe, and P. Zuidema, 2005: A New Double-Moment Microphysics Parameterization for Application in Cloud and Climate Models. Part II: Single-Column Modeling of Arctic Clouds. *J. Atmos. Sci.*, **62**, 1678–1693, https://doi.org/10.1175/JAS3447.1.

Morrison, H., G. Thompson, and V. Tatarskii, 2009: Impact of Cloud Microphysics on the Development of Trailing Stratiform Precipitation in a Simulated Squall Line: Comparison of One- and Two-Moment Schemes. *Mon. Weather Rev.*, 991–1007, https://doi.org/10.1175/2008MWR2556.1.

——, J. A. Milbrandt, G. H. Bryan, I. Kyoko, S. A. Tessendorf, and G. Thompson, 2015: Parameterization of Cloud Microphysics Based on the Prediction of Bulk Ice Particle Properties . Part II : Case Study Comparisons with Observations and Other Schemes. *J. Atmos. Sci.*, **72**, 312–339, https://doi.org/10.1175/JAS-D-14-0066.1.

Murphy, A. H., 1973: A New Vector Partition of the Probability Score. *J. Appl. Meteorol.*, **12**, 595–600, https://doi.org/doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

Nakanishi, M., and H. Niino, 2009: Development of an improved turbulence closure model for the atmospheric boundary layer. *J. Meteorol. Soc. Japan*, **87**, 895–912, https://doi.org/10.2151/jmsj.87.895.

NCEI, 2017: U.S. billion-dollar weather and climate disasters: Overview. National Centers for Environmental Information. https://www.ncdc.noaa.gov/billions/.

Nelson, S. P., 1983: The Influence of Storm Flow Structure on Hail Growth. *J. Atmos. Sci.*, **40**,

1965–1983.

NOAA, 2019: National Centers for Environmental Information (NCEI) U.S. Billion-Dollar Weather and Climate Disasters. https://www.ncdc.noaa.gov/billions/ (Accessed July 11, 2019).

NOAA National Centers for Environmental Information, 2018: *United States Billion-Dollar Weather and Climate Disasters*. https://www.ncdc.noaa.gov/billions/.

Ortega, K. L., T. M. Smith, K. L. Manross, K. A. Scharfenberg, W. Arthur, A. G. Kolodziej, and J. J. Gourley, 2009: The severe hazards analysis and verification experiment. *Bull. Am. Meteorol. Soc.*, **90**, 1519–1530, https://doi.org/10.1175/2009BAMS2815.1.

——, J. M. Krause, and A. V Ryzhkov, 2016: Polarimetric Radar Characteristics of Melting Hail. Part III: Validation of the Algorithm for Hail Size Discrimination. *J. Appl. Meteorol. Climatol.*, **55**, 829–848, https://doi.org/10.1175/JAMC-D-15-0203.1.

Pan, Y., and M. Wang, 2019: Impact of the Assimilation Frequency of Radar Data with the ARPS 3DVar and Cloud Analysis System on Forecasts of a Squall Line in Southern China. *Adv. Atmos. Sci.*, **36**, 160–172, https://doi.org/10.1007/s00376-018-8087-5.

Park, H. S., a. V. Ryzhkov, D. S. Zrnić, and K.-E. Kim, 2009: The Hydrometeor Classification Algorithm for the Polarimetric WSR-88D: Description and Application to an MCS. *Weather Forecast.*, **24**, 730–748, https://doi.org/10.1175/2008WAF2222205.1.

Putnam, B. J., M. Xue, Y. Jung, N. Snook, and G. Zhang, 2014: The Analysis and Prediction of Microphysical States and Polarimetric Radar Variables in a Mesoscale Convective System Using Double-Moment Microphysics, Multinetwork Radar Data, and the Ensemble Kalman Filter. *Mon. Weather Rev.*, **142**, 141–162, https://doi.org/10.1175/MWR-D-13-00042.1.

——, and Coauthors, 2017a: Ensemble Probabilistic Prediction of a Mesoscale Convective System and Associated Polarimetric Radar Variables using Single-Moment and Double-Moment Microphysics Schemes and EnKF Radar Data Assimilation. *Mon. Weather Rev.*, **145**, 2257–2279, https://doi.org/10.1175/MWR-D-16-0162.1.

——, M. Xue, Y. Jung, G. Zhang, and F. Kong, 2017b: Simulation of Polarimetric Radar Variables from 2013 CAPS Spring Experiment Storm-Scale Ensemble Forecasts and Evaluation of Microphysics Schemes. *Mon. Weather Rev.*, **145**, 49–73, https://doi.org/10.1175/MWR-D-15-0415.1.

——, ——, ——, N. Snook, and G. Zhang, 2019: Ensemble Kalman Filter Assimilation of Polarimetric Radar Observations for the 20 May 2013 Oklahoma Tornadic Supercell Case. *Mon. Weather Rev.*, **147**, 2511–2533, https://doi.org/10.1175/mwr-d-18-0251.1.

Rasmussen, R. M., and A. J. Heymsfield, 1987: Melting and Shedding of Graupel and Hail. Part I: Model Physics. *J. Atmos. Sci.*, **44**, 2754–2763.

——, V. Levizzani, and H. R. Pruppacher, 1984: A Wind Tunnel and Theoretical Study of the Melting Behavior of Atmospheric Ice Particles. III: Experiment and Theory for Spherical Ice Particles of Radius > 500 um. *J. Atmos. Sci.*, **41**, 381–388.

Romine, G. S., C. S. Schwartz, C. Snyder, J. L. Anderson, and M. L. Weisman, 2013: Model bias in a continuously cycled assimilation system and its influence on convection-permitting forecasts. *Mon. Weather Rev.*, **141**, 1263–1284, https://doi.org/10.1175/MWR-D-12-00112.1.

——, ——, J. Berner, K. R. Fossell, C. Snyder, J. L. Anderson, and M. L. Weisman, 2014: Representing forecast error in a convection-permitting ensemble system. *Mon. Weather Rev.*,

142, 4519–4541, https://doi.org/10.1175/MWR-D-14-00100.1.

Rotunno, R., J. B. Klemp, and M. L. Weisman, 1988: A Theory for Strong, Long-Lived Squall Lines. *J. Atmos. Sci.*, **45**, 463–485.

Ryzhkov, A., M. Pinsky, A. Pokrovsky, and A. Khain, 2011: Polarimetric Radar Observation Operator for a Cloud Model with Spectral Microphysics. *J. Appl. Meteorol. Climatol.*, **50**, 873–894, https://doi.org/10.1175/2010JAMC2363.1.

Ryzhkov, A. V., S. E. Giangrande, V. M. Melnikov, and T. J. Schuur, 2005: Calibration issues of dual-polarization radar measurements. *J. Atmos. Ocean. Technol.*, **22**, 1138–1155, https://doi.org/10.1175/JTECH1772.1.

——, M. R. Kumjian, S. M. Ganson, and P. Zhang, 2013: Polarimetric radar characteristics of melting hail. part II: Practical implications. *J. Appl. Meteorol. Climatol.*, **52**, 2871–2886, https://doi.org/10.1175/JAMC-D-13-074.1.

Sammler, W., 1993: An Updated Climatology of Large Hail Based on 1970-1990 Data. *17th Conf. on Severe Local Storms*, St. Louis, MO, Amer. Meteor. Soc., 32–35.

Schwartz, C. S., 2019: Medium-range convection-allowing ensemble forecasts with a variable-resolution global model. *Mon. Weather Rev.*, **147**, 2997–3023, https://doi.org/10.1175/MWR-D-18-0452.1.

——, and R. A. Sobash, 2017: Generating Probabilistic Forecasts from Convection-Allowing Ensembles Using Neighborhood Approaches: A Review and Recommendations. *Mon. Weather Rev.*, **145**, 3397–3418, https://doi.org/10.1175/MWR-D-16-0400.1.

——, and Coauthors, 2010: Toward Improved Convection-Allowing Ensembles: Model Physics Sensitivities and Optimizing Probabilistic Guidance with Small Ensemble Membership.

Weather Forecast., **25**, 263–280, https://doi.org/10.1175/2009WAF2222267.1.

——, G. S. Romine, R. A. Sobash, K. R. Fossell, and M. L. Weisman, 2015: NCAR ' s Experimental Real-Time Convection-Allowing Ensemble Prediction System. *Weather Forecast.*, **30**, 1645–1654, https://doi.org/10.1175/WAF-D-15-0103.1.

Skamarock, W. C., and Coauthors, 2008: A Description of the Advanced Research WRF Version 3. *NCAR Tech. Note NCAR/TN–4751STR,* 113, https://doi.org/10.5065/D6DZ069T.

Skinner, P. S., and Coauthors, 2018: Object-based verification of a prototype Warn-on-Forecast system. *Weather Forecast.*, https://doi.org/10.1175/WAF-D-18-0020.1.

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

Snook, N., and M. Xue, 2008: Effects of microphysical drop size distribution on tornadogenesis in supercell thunderstorms. *Geophys. Res. Lett.*, https://doi.org/10.1029/2008GL035866.

——, ——, and Y. Jung, 2012: Ensemble Probabilistic Forecasts of a Tornadic Mesoscale Convective System from Ensemble Kalman Filter Analyses Using WSR-88D and CASA Radar Data. *Mon. Weather Rev.*, **140**, 2126–2146, https://doi.org/10.1175/MWR-D-11-00117.1.

——, ——, and ——, 2013: Impacts of assumed observation errors in EnKF analyses and ensemble forecasts of a tornadic mesoscale convective system. *17th Conf. on Integrated Observing and Assimilation Systems for the Athmosphere, Oceans, and Land Surface (IOAS-AOLS)*, Austin, TX https://ams.confex.com/ams/93Annual/webprogram/Paper219933.html.

——, ——, and ——, 2015: Multi-Scale EnKF Assimilation of Radar and Conventional

Observations and Ensemble Forecasting for a Tornadic Mesoscale Convective System. *Mon. Weather Rev.*, 1035–1058, https://doi.org/10.1175/MWR-D-13-00262.1.

——, Y. Jung, J. Brotzge, B. J. Putnam, and M. Xue, 2016: Prediction and Ensemble Forecast Verification of Hail in the Supercell Storms of 20 May 2013. *Weather Forecast.*, **4**, 811–825, https://doi.org/10.1175/WAF-D-15-0152.1.

——, F. Kong, K. A. Brewster, M. Xue, K. W. Thomas, T. A. Supinie, S. Perfater, and B. Albright, 2019: Evaluation of Convection-Permitting Precipitation Forecast Products Using WRF, NMMB, and FV3 for the 2016–17 NOAA Hydrometeorology Testbed Flash Flood and Intense Rainfall Experiments. *Weather Forecast.*, **34**, 781–804, https://doi.org/10.1175/waf-d-18-0155.1.

Snook, N. A., M. Xue, and Y. Jung, 2011: Analysis of a Tornadic Mesoscale Convective Vortex Based on Ensemble Kalman Filter Assimilation of CASA X-Band and WSR-88D Radar Data. *Mon. Weather Rev.*, **139**, 3446–3468, https://doi.org/10.1175/MWR-D-10-05053.1.

——, Y. Jung, and M. Xue, 2018: Prediction of the 20 May 2013 Newcastle-Moore EF5 Tornado. *Mon. Weather Rev.*, conditionally accepted.

Snyder, C., and F. Zhang, 2003: Assimilation of Simulated Doppler Radar Observations with an Ensemble Kalman Filter. *Mon. Weather Rev.*, **131**, 1663–1677.

Snyder, J. C., H. B. Bluestein, D. T. Dawson, and Y. Jung, 2017a: Simulations of Polarimetric, X-Band Radar Signatures in Supercells. Part I: Description of Experiment and Simulated ρhv Rings. *J. Appl. Meteorol. Climatol.*, **56**, 1977–1999, https://doi.org/10.1175/JAMC-D-16-0138.1.

——, ——, ——, and ——, 2017b: Simulations of Polarimetric, X-Band Radar Signatures in

Supercells. Part II: ZDR Columns and Rings and KDP Columns. *J. Appl. Meteorol. Climatol.*, **56**, 2001–2026, https://doi.org/10.1175/JAMC-D-16-0139.1.

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

——, and ——, 2015: Assimilating surface mesonet observations with the EnKF to improve ensemble forecasts of convection initiation on 29 May 2012. *Mon. Weather Rev.*, **143**, 3700–3725, https://doi.org/10.1175/MWR-D-14-00126.1.

——, and L. J. Wicker, 2015: On the impact of additive noise in storm-scale EnKF experiments. *Mon. Weather Rev.*, **143**, 3067–3086, https://doi.org/10.1175/MWR-D-14-00323.1.

——, J. S. Kain, D. R. Bright, A. R. Dean, M. C. Coniglio, and S. J. Weiss, 2011: Probabilistic Forecast Guidance for Severe Thunderstorms Based on the Identification of Extreme Phenomena in Convection-Allowing Model Forecasts. *Weather Forecast.*, **26**, 714–728, https://doi.org/10.1175/WAF-D-10-05046.1.

Stensrud, D. J., J.-W. Bao, and T. T. Warner, 2000: Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Weather Rev.*, **128**, 2077–2107, https://doi.org/10.1175/1520-0493(2000)128<2077:UICAMP>2.0.CO;2.

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

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

Straka, J. M., and E. R. Mansell, 2005: A Bulk Microphysics Parameterization with Multiple Ice Precipitation Categories. *J. Appl. Meteorol.*, **44**, 445–466, https://doi.org/10.1175/JAM2211.1.

Stratman, D., N. Yussouf, Y. Jung, T. Supinie, M. Xue, P. S. Skinner, and B. J. Putnam, 2020: Optimal Temporal Frequency of NSSL Phased-Array Radar Observations for an Experimental Warn-on-Forecast System. *Weather Forecast.*, accepted, https://doi.org/10.1175/WAF-D-19-0165.1.

Supinie, T. A., Y. Jung, M. Xue, D. J. Stensrud, M. M. French, and H. B. Bluestein, 2016: Impact of VORTEX2 Observations on Analyses and Forecasts of the 5 June 2009 Goshen County, Wyoming, Supercell. *Mon. Weather Rev.*, **44**, 429–449, https://doi.org/10.1175/MWR-D-15-0171.1.

——, N. Yussouf, Y. Jung, M. Xue, J. Cheng, and S. Wang, 2017: Comparison of the Analyses and Forecasts of a Tornadic Supercell Storm from Assimilating Phased-Array Radar and WSR-88D Observations. *Weather Forecast.*, **32**, 1379–1401, https://doi.org/10.1175/WAF-D-16-0159.1.

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

——, P. R. Field, R. M. Rasmussen, and W. D. Hall, Explicit Forecasts of Winter Precipitation Using an Improved Bulk Microphysics Scheme. Part II: Implementation of a New Snow Parameterization. https://doi.org/10.1175/2008MWR2387.1.

——, ——, ——, and ——, 2008: Explicit Forecasts of Winter Precipitation Using an Improved

Bulk Microphysics Scheme. Part II: Implementation of a New Snow Parameterization. *Mon. Weather Rev.*, **136**, 5095–5115, https://doi.org/10.1175/1520-0493(2004)132<0519:EFOWPU>2.0.CO;2.

Thompson, G., S. Tessendorf, A. Heymsfield, and I. Kyoko, 2018: Improving Graupel and Hail Treatment in the Thompson Microphysics Parameterization. *North American Hail Workshop*, Boulder, CO, NCAR.

Tong, C.-C., Y. Jung, C. Liu, and M. Xue, 2020: Implementation of Radar Data Assimilation Capabilities within Ensemble-Variational Hybrid GSI for the Stand-Alone Regional FV3-Based Convection-Allowing Forecasting System. *30th Conference on Weather Analysis and Forecasting (WAF)/26th Conference on Numerical Weather Prediction (NWP)*, Boston, MA, Amer. Meteor. Soc.

Tong, M., and M. Xue, 2005: Ensemble Kalman Filter Assimilation of Doppler Radar Data with a Comprssible Nonhydrostatic Model: OSS Experiments. *Mon. Weather Rev.*, **133**, 1789–1807, https://doi.org/10.1175/MWR2898.1.

——, and ——, 2008a: Simultaneous Estimation of Microphysical Parameters and Atmospheric State with Simulated Radar Data and Ensemble Square Root Kalman Filter. Part I: Sensitivity Analysis and Parameter Identifiability. *Mon. Weather Rev.*, **136**, 1630–1648, https://doi.org/10.1175/2007MWR2070.1.

——, and ——, 2008b: Simultaneous Estimation of Microphysical Parameters and Atmospheric State with Simulated Radar Data and Ensemble Square Root Kalman Filter. Part II: Parameter Estimation Experiments. *Mon. Weather Rev.*, **136**, 1630–1648, https://doi.org/10.1175/2007MWR2070.1.

Ulbrich, C. W., 1983: Natural Variations in the Analytical Form of the Raindrop Size Distribution. *J. Clim. Appl. Meteorol.*, **22**, 1764–1775.

Vivekanandan, J., W. M. Adams, and V. N. Bringi, 1991: Rigorous Approach to Polarimetric Radar Modeling of Hydrometeor Orientation Distributions. *J. Appl. Meteorol.*, **30**, 1053–1063, https://doi.org/10.1175/1520-0450(1991)030<1053:RATPRM>2.0.CO;2.

Wang, S., M. Xue, and J. Min, 2013: A four-dimensional asynchronous ensemble square-root filter (4DEnSRF) algorithm and tests with simulated radar data. *Q. J. R. Meteorol. Soc.*, **139**, 805–819, https://doi.org/10.1002/qj.1987.

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

Waugh, S. M., C. L. Ziegler, D. R. MacGorman, S. E. Fredrickson, D. W. Kennedy, and W. David Rust, 2015: A balloonborne particle size, imaging, and velocity probe for in situ microphysical measurements. *J. Atmos. Ocean. Technol.*, **32**, 1562–1580, https://doi.org/10.1175/JTECH-D-14-00216.1.

Weber, M. E., J. Y. N. Cho, J. S. Herd, J. M. Flavin, W. E. Benner, and G. S. Torok, 2007: The next-generation multimission U.S. surveillance radar network. *Bull. Am. Meteorol. Soc.*, **88**, 1739–1751, https://doi.org/10.1175/BAMS-88-11-1739.

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.

Wheatley, D. M., N. Yussouf, and D. J. Stensrud, 2014: Ensemble Kalman Filter Analyses and

Forecasts of a Severe Mesoscale Convective System Using Different Choices of Microphysics Schemes. *Mon. Weather Rev.*, **142**, 3243–3263, https://doi.org/10.1175/MWR-D-13-00260.1.

——, 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.

Whitaker, J. S., and T. M. Hamill, 2002: Ensemble Data Assimilation without Perturbed Observations. *Mon. Weather Rev.*, **130**, 1913–1924, https://doi.org/10.1175/1520-0493(2002)130<1913:EDAWPO>2.0.CO;2.

——, and ——, 2012: Evaluating Methods to Account for System Errors in Ensemble Data Assimilation. *Mon. Weather Rev.*, **140**, 3078–3089, https://doi.org/10.1175/MWR-D-11-00276.1.

Wienhoff, Z. B., H. B. Bluestein, L. J. Wicker, J. C. Snyder, A. Shapiro, C. K. Potvin, J. B. Houser, and D. W. Reif, 2018: Applications of a spatially variable advection correction technique for temporal correction of dual-Doppler analyses of tornadic supercells. *Mon. Weather Rev.*, https://doi.org/10.1175/MWR-D-17-0360.1.

Witt, A., M. D. Eilts, G. J. Stumpf, J. T. Johnson, E. D. W. Mitchell, and K. W. Thomas, 1998a: An Enhanced Hail Detection Algorithm for the WSR-88D. *Weather Forecast.*, **13**, 286–303, https://doi.org/10.1175/1520-0434(1998)013<0286:AEHDAF>2.0.CO;2.

——, ——, ——, E. D. W. Mitchell, J. T. Johnson, and K. W. Thomas, 1998b: Evaluating the Performance of WSR-88D Severe Storm Detection Algorithms. *Weather Forecast.*, **13**, 513–

518, https://doi.org/10.1175/1520-0434(1998)013<0513:ETPOWS>2.0.CO;2.

Xue, M., K. K. Droegemeier, and V. Wong, 2000: The Advanced Regional Prediction System (ARPS) - A multi-scale nonhydrostatic atmospheric simulation and prediction model. Part 1: Model dynamics and verification. *Meteorol. Atmos. Phys.*, **75**, 161–193, https://doi.org/10.1007/s007030170027.

——, and Coauthors, 2001: The Advanced Regional Prediction System (ARPS) - A multi-scale nonhydrostatic atmospheric simulation and prediction tool. Part II: Model physics and applications. *Meteorol. Atmos. Phys.*, **76**, 143–165, https://doi.org/10.1007/s007030170027.

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

——, Y. Jung, and G. Zhang, 2010: State estimation of convective storms with a two-moment microphysics scheme and an ensemble Kalman filter: Experiments with simulated radar data. *Q. J. R. Meteorol. Soc.*, **136**, 685–700, https://doi.org/10.1002/qj.593.

Ying, Y., F. Zhang, and J. L. Anderson, 2018: On the selection of localization radius in ensemble filtering for multiscale quasigeostrophic dynamics. *Mon. Weather Rev.*, **146**, 543–560, https://doi.org/10.1175/MWR-D-17-0336.1.

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

——, and K. H. Knopfmeier, 2019: Application of the Warn-on-Forecast system for flash-flood-

producing heavy convective rainfall events. *Q. J. R. Meteorol. Soc.*, **145**, 2385–2403, https://doi.org/10.1002/qj.3568.

——, E. R. Mansell, L. J. Wicker, D. Wheatley, and D. Stensrud, 2013: The Ensemble Kalman Filter Analyses and Forecasts of the 8 May 2003 Oklahoma City Tornadic Supercell Storm Using Single- and Double-Moment Microphysics Schemes. *Mon. Weather Rev.*, **141**, 3388–3412, https://doi.org/10.1175/MWR-D-12-00237.1.

——, D. C. Dowell, L. J. Wicker, K. H. Knopfmeier, and D. M. Wheatley, 2015: Storm-Scale Data Assimilation and Ensemble Forecasts for the 27 April 2011 Severe Weather Outbreak in Alabama. *Mon. Weather Rev.*, **143**, 3044–3066, https://doi.org/10.1175/MWR-D-14-00268.1.

——, 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.

——, 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.

Zhang, Y., F. Zhang, D. J. Stensrud, and Z. Meng, 2015: Practical predictability of the 20 May 2013 tornadic thunderstorm event in Oklahoma: sensitivity to synoptic timing and

topographical influence. *Mon. Weather Rev.*, **143**, 2973–2997, https://doi.org/10.1175/MWR-D-14-00394.1.

Zhu, Y., and J. M. Navon, 1999: Impact of Parameter Estimation on the Performance of the FSU Global Spectral Model Using Its Full-Physics Adjoint. *Mon. Weather Rev.*, **127**, 1497–1517.

Ziegler, C. L., P. S. Ray, and N. C. Knight, 1983: Hail Growth in an Oklahoma Multicell Storm. *J. Atmos. Sci.*, **40**, 1768–1791.

Zrnic, D. S., and Coauthors, 2007: Agile-beam phased array radar for weather observations. *Bull. Am. Meteorol. Soc.*, **88**, 1753–1766, https://doi.org/10.1175/BAMS-88-11-1753.