UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

OPTIMAL DESIGN OF A CONVECTION-ALLOWING ENSEMBLE FROM A MODEL ERROR PERSPECTIVE

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

JEFFREY DUDA
Norman, Oklahoma
2016

OPTIMAL DESIGN OF A CONVECTION-ALLOWING ENSEMBLE FROM A MODEL ERROR PERSPECTIVE

A DISSERTATION APPROVED FOR THE
SCHOOL OF METEOROLOGY

BY

_____
Dr. Xuguang Wang, Chair

_____
Dr. Ming Xue, Co-Chair

_____
Dr. Dave Stensrud

_____
Dr. Alan Shapiro

_____
Dr. Yang Hong

_____
Dr. Yiqi Luo

Acknowledgements

English poet John Donne is credited with the quote, "'No man is an island.'" Without the cognitive, emotional, and financial support of a number of individuals, I would not have been able to conduct the research presented herein and complete this dissertation. I owe a great debt of gratitude to these individuals, named below along with the support they provided.

First, the financial support. I was supported by NSF grants AGS-1046081, AGS-0802888, OCI-0905040, and a grant from the NOAA CSTAR program. The CAPS storm-scale ensemble members analyzed in Chapter 3 were also produced under the support of NOAA CSTAR program (NA17RJ1227). I also thank my advisors, Xuguang Wang and Ming Xue, not just for procuring funds via grant proposals, but also for their guidance throughout my time as a student at the University of Oklahoma. I learned a great deal about how to be a better scientist, researcher, and writer through their advice. I also value the conversations I had with Dave Stensrud throughout my time as a Ph.D. student.

On to the material support. Significant computing resources used for this work were provided by the National Science Foundation XSEDE National Institute of Computational Science at the University of Tennessee, the Texas Advanced Computing Center, and the OU Supercomputing Center for Education & Research (OSCER). OSCER Director Henry Neeman and Petascale Storage Administrator Patrick Calhoun were especially helpful in providing guidance on computing and storage procedures. OSCER systems analysts David Akin and Brett Zimmerman, and HPC application software specialist Joshua Alexander also provided helpful coding and software

maintenance advice. A handful of support staff at XSEDE were also very helpful in addressing smaller issues that occurred occasionally.

A few specific individuals provided especially large amounts of material support. Fanyou Kong offered a tremendous amount of help with coding and accessing resources, especially those associated with the ARPS code and WRF data generated during SFE 2011 and 2012. Fanyou also provided advice on the design of the SKEB ensemble experiments described in Chapter 4. I thank Judith Berner for her consultation and advice throughout the work on the SKEB experiment in Chapter 4. I also thank Dan Dawson for advice on calculating microphysics retrievals presented in Chapter 3 and for offering code to produce the retrievals. I also thank the work done by those in the references upon which this work is built. Scientific research can only advance if we all, as scientists, stand on the shoulders of those who came before us, and therefore everything I accomplished in this dissertation was made possible by the work of those who came before me.

I have been lucky to have strong emotional support from friends and family. My parents, Dennis and Cynthia, have supported me through all of the ventures of my life and I am grateful for that. While I do not frequently get to enjoy the company of many of my closest friends due to distance, I still value their presence in my life. To Logan, Darren, Zac, Mason, and Petar, thank you for all the good times we have had and for providing support when I needed it. Finally, while our time together has come to an end, I would not have made it this far nor enjoyed life as much without my best friend over nine years, Megan. Thank you for your support and for bringing the gift of love into my life. You made my life better than it otherwise would have been.

Table of Contents

List of Tables

List of Figures

Abstract

As computer technology continues to improve, resources are becoming increasingly available for running an ensemble of NWP simulations with sufficient resolution that convection parameterization is no longer needed and individual storms can be captured on the grid scale. Due to past limitations on technology, research into the optimal design of convection-allowing ensembles has remained limited.

The purpose of ensembles is to account for forecast uncertainty that arises due to errors in the forecast process. There are errors in the initial conditions resulting from incomplete spatiotemporal sampling of the atmosphere and measurement error. There are errors in the lateral boundary conditions that drive limited-area models (as is the case for current experimental convection-allowing ensembles). There are also errors associated with the model formulation caused by numerical error from discretization of the model grid, truncation of numerical schemes, and inadequate subgrid-scale physics parameterizations that provide source and sink terms for the resolved-scale variables. This dissertation focuses on methods of accounting for uncertainty in assorted model physics components.

Three research projects were performed using the WRF model with 4 km grid spacing. One investigated methods of accounting for microphysics uncertainty. Another investigated the impact of adding a stochastic model error representation scheme. The third project investigated methods of accounting for uncertainty in the land-surface model component.

Two methods of perturbing the microphysics were compared: perturbing fixed parameters within a single scheme and using multiple schemes. The latter, termed

mixed microphysics, was found to generate somewhat more skillful and more reliable probabilistic forecasts of precipitation, although the former, termed perturbed parameter microphysics, also performed well.

A stochastic model error scheme that was originally created to improve large-scale forecasts in the ECMWF model was converted for use in the WRF model. The stochastic kinetic energy backscatter, or SKEB, scheme injects kinetic energy at all scales using a forced power spectrum. Its utility in a convection-allowing ensemble was analyzed and found to add significant ensemble spread while also reducing ensemble mean error. The SKEB scheme does not perturb any moisture variables, but was successful in causing slight improvements to precipitation forecasts.

The final portion of this dissertation features an exploratory study to determine a perturbation strategy for the land-surface model component. Literature review revealed many uncertainties in current land-surface models that have not been accounted for in prior experimental convection-allowing ensembles. A set of perturbations that reflect uncertainty in the calculation of sensible and latent heat flux was determined and applied to a small set of cases. The sensitivity of convection forecasts to these perturbations was assessed and compared to that from other physics perturbations (microphysics and PBL physics). It was also determined that adding land-surface model perturbations to other existing physics perturbations further increases ensemble diversity and improves probabilistic forecasts of precipitation.

Overall, it is encouraging that the WRF provides a wealth of resources for accounting for model physics uncertainty in convection-allowing NWP forecasts.

Chapter 1. Background and Motivation

1.1 Introduction

Ensemble forecasting is a method of using multiple numerical weather prediction (NWP) model simulations to account for the inherent uncertainty a weather forecast. A model ensemble consists of multiple independent (aka, deterministic) NWP simulations that cover the same forecast time. Individual realizations of the ensemble, called members, do not need to be initialized at the same time or use the same NWP model core, boundary conditions, or model physics options. Variations in any of these are sufficient to render an ensemble forecast. The number of members, referred to as ensemble size, can vary from a few to as many as one desires or as much as computational expense allows. For large-scale atmospheric ensemble forecasting, ensemble sizes typically range from 10 to as many as 50. Such ensembles are typically referred to as ensemble prediction or ensemble forecast systems (EPS or EFS). Either term may be used interchangeably throughout this dissertation.

Ensemble forecasts aim to account for the sources of error in a deterministic NWP forecast. One major source of error is the initial condition. Initial condition error results from an incomplete and inaccurate specification of the atmospheric state: the atmosphere is not measured at every location; instruments that sample the atmosphere contain error; and interpolation is required to map observations to the model grid, which can also generate error. Clearly, any NWP forecast initialized using data with nonzero error will contain nonzero error. The boundary conditions supplied to drive the model integration are a second source of error for the same reasons as for initial condition error. Boundary condition error is especially important for model domains with limited

areal extent, but it is present in all model domains due to incomplete and errant specification of the lower (i.e., surface) and upper (i.e., upper atmosphere or edge of space) model domain boundaries. A third source is from an imperfect construction of the NWP model. Error will result merely from discretizing the spatial and temporal dimensions in an NWP model since the real atmosphere is continuous. However, this source of error is controllable and at least somewhat predictable since the numerical methods used in the models have known error characteristics. Perhaps more importantly, there are many atmospheric processes that occur on a temporal and spatial scale too small to be resolved by the model domain, called subgrid-scale processes. A parameterization scheme is required to include the effects of these processes on the scales the model can resolve. Obviously, not parameterizing a real process at all will result in forecast error. So merely including the parameterization should reduce error, but the parameterization itself represents an additional source of model error. This dissertation focuses on methods of accounting for model error from uncertainties in physical parameterization schemes. The focus is also on convective-scale modeling (also referred to as storm-scale), which refers to using a rather highly refined model grid capable of approximately containing features as small as thunderstorms. Such model domains are limited in horizontal extent, and thus boundary conditions are a significant issue, but they will not be covered here. Initial condition error will also not be investigated.

What follows is background discussion of ensemble forecasting followed by a history of the uses of ensembles for weather prediction, and finally discussion

motivating the exploration of methods for accounting for sources of model error in a convection-allowing ensemble presented in this dissertation.

1.2 Background on ensemble forecasting

A general goal of ensemble forecasting is to assess the uncertainty in a forecast. Another way to describe this goal is to inform how a forecast may err if some component in the forecast process is wrong. For example, suppose the output from a single NWP forecast for 2-m temperature is 280 K. If that forecast is wrong, a user has no information on what the actual verifying 2-m temperature may be. The forecast value of 280 K may represent the best guess from the NWP forecast system, but it only contains that single piece of information. An ensemble provides an extra dimension of information (additional moments of the forecast distribution, whereas the singular NWP forecast only provides one moment) that can be very useful to a user. An ensemble forecast for the same variable verifying at the same location and same time with an ensemble mean of 280 K could forecast a range from 276 K to 284 K with a fairly symmetric distribution, or it could forecast a range from 279 K to 287 K with a highly right-skewed distribution. This extra information could cause a user to make a different decision on a meteorologically sensitive issue than the one they would make if only a single-valued forecast was available. This simple example illustrates the value that ensembles can provide. Some basic theory and additional examples can be found in Richardson (2001) and Zhu et al. (2002). It should be noted that every forecast contains some uncertainty and therefore can be represented as a distribution. Therefore, it is also consistent to say that the goal of ensemble forecasting is to mimic the underlying forecast distribution.

A perfect ensemble requires an infinite ensemble size to fully populate the forecast distribution and eliminate sampling uncertainty, and therefore cannot be realized in practice. However, an ideal ensemble, which can be realized, produces probabilistic forecasts with certain desirable characteristics. One characteristic is statistical reliability. Reliability is the amount of agreement between a probabilistic forecast and the corresponding frequency of occurrence of the event for a given probability. Reliability is one component of the Brier score (Brier 1950), and hence can be analyzed numerically (see appendix for a breakdown of the Brier score). It can also be analyzed using a graphical technique called a reliability diagram (see Chapter 2). For instance, suppose an ensemble forecasts a probability of 45% that 24-hour precipitation will accumulate to greater than 10 mm. The ensemble would be perfectly reliable if the verifying 24-hour precipitation accumulation exceeded 10 mm 45% of the time when such a probability was forecast. Clearly, such a forecast must be issued many times so that such a probability can be numerically realized, as any single instance of an event carries a probability of occurrence of either 0% or 100%. Hence, like the other characteristics of an ideal ensemble, reliability is best assessed over a large enough period of time for a sufficiently large sample to be obtained.

Another desirable trait of an ensemble is high resolution. Resolution is another component of the Brier score and can also be evaluated both numerically and graphically using a reliability diagram (see Chapter 2). Resolution measures the difference between forecast probability and climatological frequency, and is consistent with forecast sharpness (probabilities close to 0% and 100%; Murphy 1993). An ensemble that consistently forecasts a probability of occurrence of an event close to the

climatological frequency is of little value since similar forecast accuracy could be obtained by using the climatological value for every forecast. Therefore, an ensemble with higher resolution forecasts probabilities greatly different from the climatological probability. However, high resolution is meaningless if the forecasts are unreliable (i.e., highly error prone). An ideal ensemble has both high resolution and perfect reliability (Toth et al. 2001).

Another characteristic of an ideal ensemble is that the verifying value is statistically indistinguishable from any given ensemble member. This characteristic can be assessed using rank histograms (Hamill 2001; see Chapter 2). A rank histogram shows where within the forecast distribution the verifying value tends to be located over a large number of samples. The rank histogram for an ideal ensemble is flat, which means that the verifying value could be any one of the ensemble members (Hou et al. 2001). This characteristic can also be expressed by stating that the verification and ensemble members appear to be drawn from the same distribution, the "truth" distribution.

Yet another trait is that the ensemble spread increases with time at a similar rate to the ensemble error. This trait is desirable is because it is assumed that the initial condition and model perturbations are drawn from the truth distribution and thus represent error, so the divergence of ensemble member forecasts with time should be similar to the divergence of any given member (or the ensemble mean) from the true atmospheric state. This characteristic can be analyzed by examining time series of ensemble diversity and ensemble error. A common measure of each quantity is the

ensemble standard deviation and the root mean square error of the ensemble mean (RMSE), defined below:

$$standard\ deviation = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} \quad (1.1),$$

$$root\ mean\ square\ error = \sqrt{\frac{1}{M}\sum_{M}\left(\overline{x_f} - x_t\right)^2} \quad (1.2).$$

In (1.1) and (1.2) N represents ensemble size, an overbar represents an ensemble mean value, x_i represents the value of a field from one ensemble member, the subscript f denotes a forecast value, the subscript t denotes the true value (while the true value is never known and observed values contain error, observed values can be used when averaging over a large number of cases so long as there is no long term bias), and M represents the number of independent cases. If the ensemble members and true state are being drawn from the same distribution then, after averaging over a large number of cases, the mean squared quantities in the formulas above should be approximately the same (e.g., Wang and Bishop 2003). Due to sampling uncertainty, the ratio of spread to RMSE should not be exactly 1.0, but rather $\sqrt{N-1}/\sqrt{N+1}$ (Leutbecher and Palmer 2008). It should be noted that there are other measures of ensemble diversity and quality (e.g., Grimit and Mass 2007). However, standard deviation and RMSE of the ensemble mean are arguably the most commonly used measures, and will also be used throughout this dissertation.

1.3 History of ensemble forecasting

Operational ensemble forecasting in a large scale capacity (i.e., for entire regions, nations, or collections of nations) officially began at the National Meteorological Center (now the National Center for Environmental Prediction; NCEP) and the European Center for Medium-Range Weather Forecasts (ECMWF) in December 1992 (Tracton and Kalnay 1993; Molteni et al. 1996). Other major weather agencies such as the Canadian Meteorological Center followed within a few years (e.g., Houtekamer et al. 1996). The theory and experimentation that provided the basis for ensemble forecasting methods was developed and conducted many years prior, however. During the 1950s through the 1970s many approaches to accounting for forecast error from newly developed NWP models were investigated, headed by pioneers such as Edward Lorenz, Cecil Leith, Philip Thompson, and Edward Epstein (Lewis 2005). Philip Thompson was among the first to discuss how forecast errors were related to initial condition errors and how such errors would impact predictability (Thompson 1957). Edward Lorenz is frequently credited for discovering that the atmosphere is a chaotic dynamical system, and provides a detailed lecture and demonstrations of chaos in his book *The Essence of Chaos* (Lorenz 1993). Lorenz showed, using the results from an accidental discovery involving the use of truncated numerical values in intermediate calculations of an NWP simulation, that the atmosphere is nonperiodic and NWP simulations exhibit sensitive dependence to small changes in initial conditions, which today atmospheric scientists and forecasters recognize as a fundamental limit on the predictability of atmospheric phenomena. Edward Epstein investigated stochastic-dynamic prediction – the notion of integrating

individual moments of a probability distribution forward in time – and Cecil Leith investigated the use of Monte Carlo methods – taking random draws from assumed probability distributions that represent initial condition uncertainty – to account for initial condition error (Epstein 1969; Leith 1974).

While currently existing methods to account for forecast uncertainty (from initial condition uncertainty) in operational model ensembles still resemble some of the earlier methods suggested by these pioneers, recent developments, especially since the late 1990s, have diverged somewhat from the earlier ideas. In particular, the assumption that the forecast model itself was perfect, and thus only initial condition uncertainty was worth investigating, has been dropped. Stensrud et al. (2000) was one of the earliest papers to document the use of mixed physics parameterizations in ensemble forecasts and demonstrated that a model error approach can result in ensemble diversity that cannot be represented by only initial condition uncertainty. Probabilistic forecasts of many atmospheric quantities were improved using this technique. The current operational Short-Range Ensemble Forecast (SREF) system run by NCEP, a multi-model, multi-physics, and initial condition ensemble, was born out of this paper combined with the investigations of Du et al. (1997), Hamill and Colucci (1997), and Hou et al. (2001). Since then, many other papers covering the development of a mixed physics approach in global scale and mesoscale ensembles have been published (e.g., Eckel and Mass 2005; Clark et al. 2008; Hacker et al. 2011; Berner et al. 2011; Stensrud and Yussouf 2003, 2007; Wandishin et al. 2001). The diversity in many of these studies was generated using different planetary boundary layer, cloud microphysics, deep convection, and radiation parameterization schemes. However, the research also was

limited to coarser grid domains and simpler physics options. Research using these methods at higher resolutions and using more sophisticated physics is less common.

Nearly contemporaneously with the addition of model error methods, model ensembles run at grid spacings sufficient to represent individual thunderstorms emerged, with developments continuing through the current day. Until very recently (within the last five years or so) the use of these storm-scale ensembles has been limited to experimental and research purposes over very small horizontal domains and very short time ranges. But with continued increases in computational resources over the next few years, model domains, forecast lengths, and ensemble configuration have gradually expanded, and operational use of storm-scale ensembles over large regions such as the continental United States (CONUS) is likely to begin soon.

The earliest application of storm-scale ensembles was for severe storm forecasting over the Great Plains of the U.S. The Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma was among the first, if not the very first, entity to carry out a storm-scale ensemble forecast in real time for the first Verifications Of Rotation in Tornadoes Experiment in 1995 (Sathye et al. 1996). Other experimental uses of storm-scale ensembles followed in later years using similar configurations (e.g., horizontal domain extent of 500 km or less, forecast length of 12 hours or less; Elmore et al. 2003). By the mid-2000s, CAPS had begun much more expansive real-time experimental storm-scale forecasting of severe weather in support of the National Oceanic and Atmospheric Administration Hazardous Weather Testbed (NOAA/HWT) spring forecasting experiment (SFE), conducted annually since at least 2000 (see, e.g., Kong et al. 2007). In 2015, what is now called the Storm-Scale Ensemble Forecast

system (SSEF) was run using 20 members, each with 3 km grid spacing over a domain that encompassed the entire CONUS with a forecast length of 60 hours. The SSEF design was used as a model for the design of storm-scale ensembles used to conduct the experiments in this dissertation.

1.4 Motivation for this work

Technology advances have allowed for continual improvement in available computational resources for running NWP models. As a result, grid spacing has gradually decreased, allowing for smaller scale features to be resolved by the models. A particularly major achievement made possible by reduced grid spacing is the ability to explicitly resolve thunderstorms. Traditional scaling arguments suggest that since individual thunderstorms have length scales of O(10 km), then grid spacings of O(1 km) should be sufficient to resolve them. Weisman et al. (1997) suggested a threshold grid spacing of 4 km for deep moist convection. Bryan et al. (2003) later reduced that threshold, suggesting that grid spacings of O(100 m) are required to sufficiently resolve all convective-scale circulations. This has left the range of model grid spacings between as a sort of gray zone wherein domains are referred to as "convection allowing" as opposed to "convection resolving" because traditional deep convection parameterization was not meant to be used at grid spacings of O(1 km) (Molinari and Dudek 1992) where individual deep convective storms are not fully resolved. Without the use of hybrid convection parameterizations, greater emphasis falls on the cloud microphysics component, in particular, to handle the effects of thunderstorms. Accordingly, the opportunity to account for model error by varying the microphysics has been opened. Other physics components, such as planetary boundary layer, radiation, and land-

atmosphere exchange (also referred to as the land-surface model) are also important subgrid scale physical processes. Research into accounting for model error by varying these physics components in convective-scale ensembles is relatively new, although there have been many studies over the past several years (e.g., Bouttier et al. 2012; Johnson et al. 2011ab; Johnson and Wang 2012, 2013; Kong et al. 2007, 2011; Leoncini et al. 2010; Romine et al. 2014; Ropnack et al. 2013; Xue et al. 2008, 2009). However, the sources of model error are commonly lumped together in prior studies. The research in this dissertation isolates the impact of accounting for sources of model error from separate physics components, specifically, microphysics and land-surface model.

There remains much research to be conducted on specific methods for accounting for error from the microphysics. Many previous studies have indicated the sensitivity of convective-scale NWP forecasts to microphysics scheme. In particular, simulations of supercell thunderstorms and squall lines have been found to be sensitive to the complexity of the microphysics scheme (Morrison et al. 2009; Dawson et al. 2010; Bryan and Morrison 2012; Putnam et al. 2014), the value of particle size distribution parameters (Gilmore et al. 2004; Snook and Xue 2008; Tong and Xue 2008), and treatment of specific water phase change processes (Morrison and Milbrandt 2011). However, research is absent on which of these particular methods of accounting for microphysics uncertainty in a convective-scale ensemble provides better probabilistic forecasts of deep moist convection.

Research on accounting for uncertainty in the land-surface model (LSM) component for convective-scale ensemble forecasts is generally nonexistent. Many prior studies on LSMs have focused on implementation of changes and improvements to

individual schemes, with research aimed at global-scale and long-term forecasting (e.g., the Project for Intercomparison of Land-surface Parameterization Schemes; Henderson-Sellers et al. 1996) rather than high-resolution, short-term forecasting. Other research looking at LSM uncertainty in storm-scale forecasts have concentrated on the soil state uncertainty rather than the physical processes within the schemes (e.g., Sutton et al. 2006 and Aligo et al. 2007). LSM uncertainty has been accounted for to some degree in past iterations of the SSEF for the NOAA/HWT SFE (e.g., Xue et al. 2011). However, the degree to which the uncertainty has been accounted for is rather limited, and probably does not adequately reflect the full degree of uncertainty present within the LSM component. Therefore, more thorough research into methods for perturbing the LSM component in a storm-scale ensemble is warranted.

Finally, an investigation into accounting for an additional type of model error called stochastic model error is presented. Stochastic model error is meant to account for subgrid-scale uncertainty that is not, or cannot, be represented by existing physical parameterization schemes. Schemes using random numbers have thus been developed to attempt to account for this type of error. Much of the preliminary work on accounting for stochastic model error was conducted using the ECMWF ensemble system (e.g., Buizza et al. 1999; Shutts 2005; Berner et al. 2009) as an example. It may seem silly to use random numbers in a weather forecast, but even Edward Lorenz recognized the utility in random number processes in the 1970s (Lorenz 1975). Research conducted so far on stochastic model error has used global-scale models which do not resolve convective-scale processes. This fact has motivated the testing of one particular stochastic model error scheme in a storm-scale ensemble.

These three major topics regarding model error sources in a storm-scale ensemble are covered in this dissertation, with one chapter devoted to each. Methods of verification, a very important aspect of determining optimal ensemble design, are presented in Chapter 2. In Chapter 3, separate methods for accounting for microphysics uncertainty are compared. In Chapter 4, a stochastic kinetic energy backscatter scheme to account for random errors is discussed. In Chapter 5, methods for accounting for land-surface model uncertainty are discussed. Chapter 6 provides a closing discussion on what has been learned from these investigations.

Chapter 2. Forecast Verification

2.1 Introduction

Verification is a critical component to the forecast process despite being performed after the associated meteorological event has passed. Forecasts must be vetted for their accuracy so their value to users can be determined. Verification is also the best means by which forecast systems (whether human, machine, or hybrid) can be compared. Verification is not a simple process. Verification is also not without its uncertainties. There are many different methods of verifying a given forecast, especially when considering high-resolution forecasts. Many of the common methods, as well as others used in this dissertation, are described below.

2.2 Essential probabilistic verification metrics

Perhaps the simplest verification method is based on the difference between the forecast value and the verifying value, called simply, error. Ideally, the comparison should be made between the forecast and the true atmospheric state. Since the true atmospheric state is never known, an observation or analysis (from a very short range forecast or interpolation operator) is used as a substitute. Clearly, error at a single point in space or time does not provide much information about the overall quality of a forecast. Therefore, mean error – typically called bias – is calculated over a reasonably sized spatiotemporal domain:

$$bias = \frac{1}{N}\sum_{i=1}^{N} f_i - o_i \tag{2.1}$$

where N is the number of spatiotemporal points over which the bias is computed, f represents a forecast value, and o represents a verifying value. The sign of the bias is very useful in determining deficiencies in the forecast. The absolute value of the summand can also be used to calculate the mean absolute error, in which sign does not matter. However, a simple and strictly proper scoring rule that is commonly used is the mean squared error, in which the summand in (2.1) is squared. The root mean square error (RMSE) results when taking the square root of the result:

$$RMSE = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(f_i - o_i)^2} \tag{2.2}$$

The mean squared error approach can be applied to probabilistic forecasts. In this case, f and o are restricted to the probability range [0,1] and the radical is removed, resulting in the Brier score (BS; Brier 1950; eqn. 2.3), perhaps the most common method of verifying probabilistic forecasts. The Brier score can be arithmetically decomposed into three terms as follows:

$$BS \equiv \frac{1}{M}\sum_{i=1}^{M}(f_i - o_i)^2 = \frac{1}{M}\sum_{k=1}^{K} N_k(f_k - \overline{o_k})^2 - \frac{1}{M}\sum_{k=1}^{K} N_k(\overline{o_k} - \bar{o})^2 + \bar{o}(1-\bar{o})$$

$$= reliability - resolution + uncertainty \tag{2.3}$$

In the decomposition, M refers to the number of independent forecasts, K the number of probability categories with values defined as $f_k = \frac{i}{K}\ \forall\ i \in [0,1,\dots,K-1,K]$ (equivalent to the number of ensemble members plus one), N_k is the number of occurrences of forecast probability f_k, $\overline{o_k}$ is the conditional observed frequency for forecasts of probability f_k, and $\bar{o}$ is the climatological frequency of occurrence of the event. The first term on the RHS of (2.3) is the mathematical expression for reliability.

The second term represents resolution, and the third term is called uncertainty and is a function only of the climatology of the event, and is independent of the forecast system. It can be shown that the Brier skill score (BSS), which is a normalized BS compared to that from a reference forecast, can be expressed as

$$BSS = \frac{resolution - reliability}{uncertainty} \quad (2.4).$$

Unlike the BS which is 0.0 for a perfect forecast and has no upper limit, the BSS is 1.0 for a perfect forecast, 0.0 for a forecast that is identical in quality to the reference, and negative for a forecast that is poorer than the reference. When the reference forecast is climatology, the BSS gives an indication of whether the forecast system has any skill relative to climatology. Since BSS = 0 implies resolution and reliability are the same, a "no skill" line can be drawn on an attributes diagram that represents half the distance between perfect reliability and zero resolution (Fig. 2.1). All three of the BS components can be visualized on an attributes diagram. An inset showing the distribution of f_k can be added, which indicates the amount of weight the values at any point contribute to either component.

A rank histogram is constructed by ordering ensemble members and considering the space between each consecutive pair of members as bins. Therefore, a rank histogram has N+1 bins for N ensemble members. A count is added to a bin when the verifying value falls in the space between the corresponding pair of ensemble members (Fig. 2.2). A large number of samples/forecasts are needed to obtain a robust estimate of the ensemble quality using a rank histogram. The shape of the rank histogram provides an indication of ensemble performance (Hamill 2001). The ideal rank histogram is perfectly flat, indicating a proper distribution of ensemble members where the verifying

value is equally likely to be any one of the ensemble members, meaning any of the ensemble members are a possible representation of the truth. The rank histogram for a biased ensemble contains an uneven distribution of frequency mass between the left and right halves with the specific shape (e.g., linear slope, local maximum, or the largest counts in the outermost bin) reflecting the severity and consistency of the bias. A "U-shaped" rank histogram, where the outermost bins have the largest counts, indicates an ensemble with too little variability (aka, underdispersive). Conversely, an upside-down-U-shaped rank histogram, with most counts towards the middle of the rank histogram, indicates an ensemble with too much spread (overdispersive).

2.3 Contingency table verification

A large number of commonly used verification metrics require a contingency table which also requires dichotomous forecasts. Dichotomous forecasts are those for which the forecast can be reduced to a "yes" or "no" that a meteorological event will occur. For continuous fields, this can be done by applying a threshold value to be exceeded or not exceeded. The corresponding verification value is also "yes" or "no" regarding whether the event actually occurs or the threshold value is exceeded. A contingency table contains the frequency of the four possible scenarios involving dichotomous forecast verification (Table 2.1).

Table 2.1. A 2 x 2 contingency table for dichotomous forecasts.

	forecast: yes	forecast: no
observation: yes	hit (A)	miss (B)
observation: no	false alarm (C)	correct null (D)

Counts are derived from each grid point where the forecast and observation fields are both valid. Contingency table metrics used for verification in this dissertation are defined below.

2.3.1 Probability of detection (POD)

$$POD = \frac{A}{A + B} \tag{2.5}$$

The probability of detection is defined as the fraction of observed events that were forecast to occur. POD ranges from 0.0 (no observed events were forecast) to 1.0 (each event was forecast to occur). A higher POD corresponds to a more accurate forecast.

2.3.2 Probability of false detection (POFD)

$$POFD = \frac{C}{C + D} \tag{2.6}$$

The probability of false detection is defined as the ratio of the number of events that were incorrectly forecast to occur to the number of times the event did not occur. Like POD, POFD also ranges from 0.0 to 1.0. However, unlike POD, a perfect forecast has a POFD of 0.0. POFD is sometimes confused with the false alarm rate, defined as the ratio of the number of events that were incorrectly forecast to occur to the number of times the event did occur. The false alarm rate was not used in this dissertation.

The POFD can be combined with the POD and plotted graphically to create a Receiver Operating Characteristic (ROC) curve (Mason 1982; Fig. 2.3), which is a method of verifying probabilistic forecasts. Typically the area under the ROC curve is calculated to integrate the contribution from each forecast threshold. Since probabilistic thresholds of ≤0% (every forecast is a "yes") and >100% (every forecast is a "no") are

included to fill the real number space and complete the ROC curve, a forecast with no skill has an area of 0.5. ROC areas range from 0.0 to 1.0, with a perfect forecast having a ROC area of 1.0.

2.3.3 Frequency bias

$$frequency\ bias = \frac{A + C}{A + B} \tag{2.7}$$

Frequency bias is defined as the ratio of the number of events forecast to occur to the number of events that occurred. Frequency bias can also be thought of as an areal bias in the context of the area of a forecast object or areal coverage of an event above some threshold. Frequency bias ranges from 0.0 to infinity, with a perfect forecast having a frequency bias of 1.0.

2.3.4 Equitable threat score (ETS)

$$ETS = \frac{A - r}{A + B + C - r} \tag{2.8a}$$

$$r = \frac{(A + C)(A + B)}{A + B + C + D} \tag{2.8b}$$

The equitable threat score is a threat score with adjustments made for getting "yes" forecasts correct by random chance, represented by r. ETS ranges from -⅓ to 1.0, with any value less than 0.0 indicating "no skill" relative to a random forecast. There is a similar score that accounts for achieving correct "no" forecasts by chance also, called the Heidke skill score, but it is not commonly used in verifying precipitation forecasts from mesoscale and storm-scale models, and was not used in this dissertation.

2.4 Neighborhood verification

It has been shown that traditional verification metrics such as mean error, RMSE, and ETS may not adequately reflect the true accuracy of high-resolution forecasts compared to coarser-resolution forecasts (Mass et al. 2002). This is because these scores can be very sensitive to small displacement errors, both in space and time, of features that may otherwise be subjectively judged to be skillfully forecast. Ahijevych et al. (2009) discusses this issue further and reviews other methods of verification that may be more appropriate for high-resolution QPF verification that is performed throughout this dissertation. The neighborhood approach is used heavily as the chosen method of alternative verification throughout this dissertation (Ebert 2009; Schwartz et al. 2010). Some of the verification metrics discussed in section 2.3 can be extended to a neighborhood.

The ETS can be extended to neighborhood by allowing hits, misses, and false alarms to be determined using a neighborhood around each grid point. Clark et al. (2010b) provide more details and examples. NETS is best used to verify deterministic forecasts rather than ensemble mean forecasts, especially for feature-based fields like precipitation and reflectivity (A. Clark 2015, personal communication).

A ROC curve can be constructed using probabilistic forecasts based on spatial and/or temporal neighborhoods either instead of, or in addition to, ensemble members.

Roberts and Lean (2008) and Schwartz et al. (2010) introduce the fractions skill score (FSS), the neighborhood based analog to the BSS. The FSS is defined as

$$FSS = \frac{FBS_{worst} - FBS}{FBS_{worst}} \quad \text{(2.9a)},$$

where

$$FBS = \frac{1}{N}\sum_{i=1}^{N}(NF_i - NO_i)^2 \quad (2.9b),$$

where NF_i and NO_i are neighborhood probabilities of an event, the former including spatiotemporal coverage in the forecast as well as ensemble coverage, and the latter using only observed coverage, at each grid point, and N is the total number of grid points. The FBS is essentially identical to the BS except for the fact that the observed value is allowed to take on values other than 0.0 and 1.0. In theory, the "worst" (highest) FBS occurs when there is absolutely no overlap between forecast and observed neighborhoods with a nonzero probability for a given event. In that case, one of the terms in the square of the summand can be eliminated, giving

$$FBS_{worst} = \frac{1}{N}\left[\sum_{i=1}^{N} NF_i^2 + \sum_{i=1}^{N} NO_i^2\right] \quad (2.9c).$$

Therefore, like the BS and BSS, the FBS and FSS of a perfect forecast is 0.0 and 1.0, respectively, whereas negative FSS values indicate poorer performance compared to a reference forecast, in this case, the poorest forecast field possessing the same spatial forecast coverage, but rearranged such that the forecast probability at any point is as far from the truth as possible. In essence, the FBS measures the difference in spatiotemporal coverage of neighborhood probability forecasts and observations and, like other neighborhood based verification metrics, partially rewards forecasts for being "close" to observations without requiring the forecast location to be perfect. Roberts and Lean (2008) also discuss how the FSS can be used alternately to determine at which spatial scales a forecast becomes of sufficient quality, with comparisons made using the smallest neighborhood size at which FSS exceeds a certain value. In this dissertation, a

simple comparison between neighborhood coverage between differently configured ensembles is preferred, so FBS is used preferentially to FSS.

2.5 Object based verification

Another alternative method of verification is object based verification. Object based verification is best applied to fields such as precipitation or radar reflectivity that are more feature based (i.e., contain small objects against a background of zero values) rather than plainly continuous fields such as temperature or geopotential height. The reason for using object based verification measures is the same as for using neighborhood verification measures. The choice of which to use is more arbitrary since well-established methods of either neighborhood or object based verification have yet to emerge. The general formulation of object based verification is to identify features or objects in a given field. Objects are identified generally based on a fuzzy logic engine, and objects in the forecast are compared to objects in observations. A wide variety of metrics can be computed based on the comparison between forecast and observed objects. The Method of Object based Diagnostic Evaluation is a popular object based verification tool (Davis et al. 2006). It has been used for storm-scale verification, both in a deterministic sense (e.g., Duda et al. 2013) and in an ensemble sense (e.g., Johnson and Wang 2012).

I chose arbitrarily not to use object based verification in this dissertation.

Figures

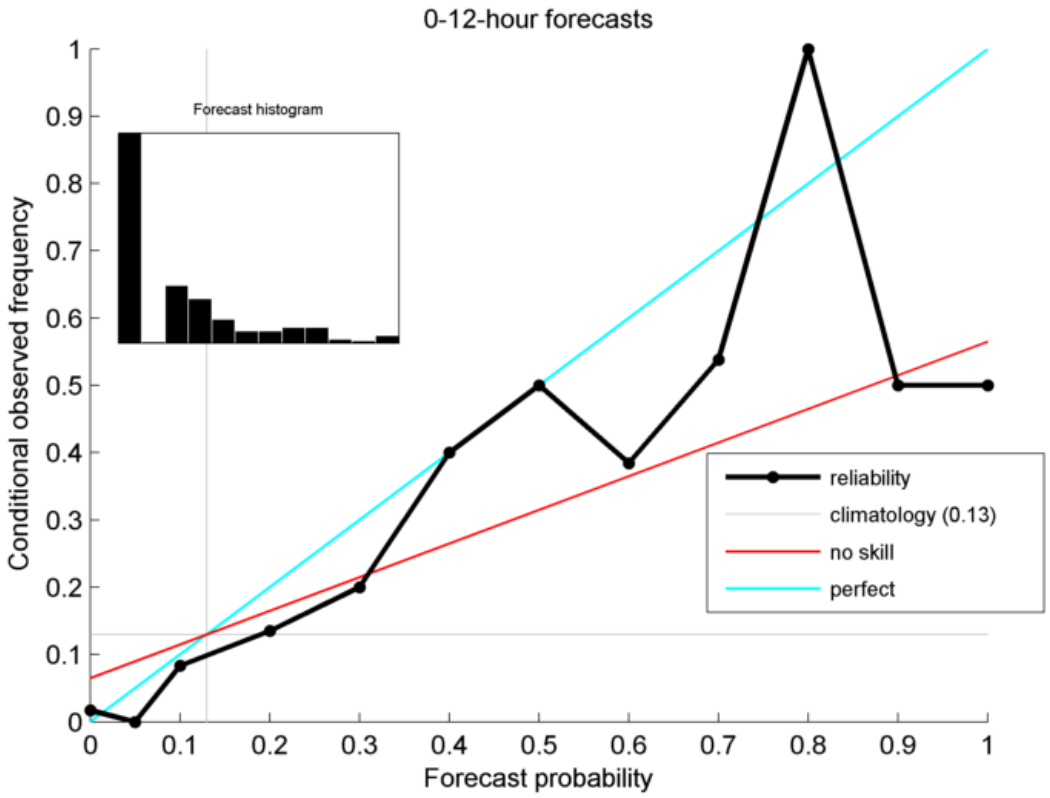

Fig. 2.1. Example reliability diagram for 0-12 hour probabilistic forecasts of measureable precipitation. Reliability is visualized as the sum of the vertical distances between the forecast values (black line) and perfect reliability (cyan line), while resolution is the sum of the vertical distances between the forecast values and climatology (horizontal light gray line). A forecast histogram is included in the inset.

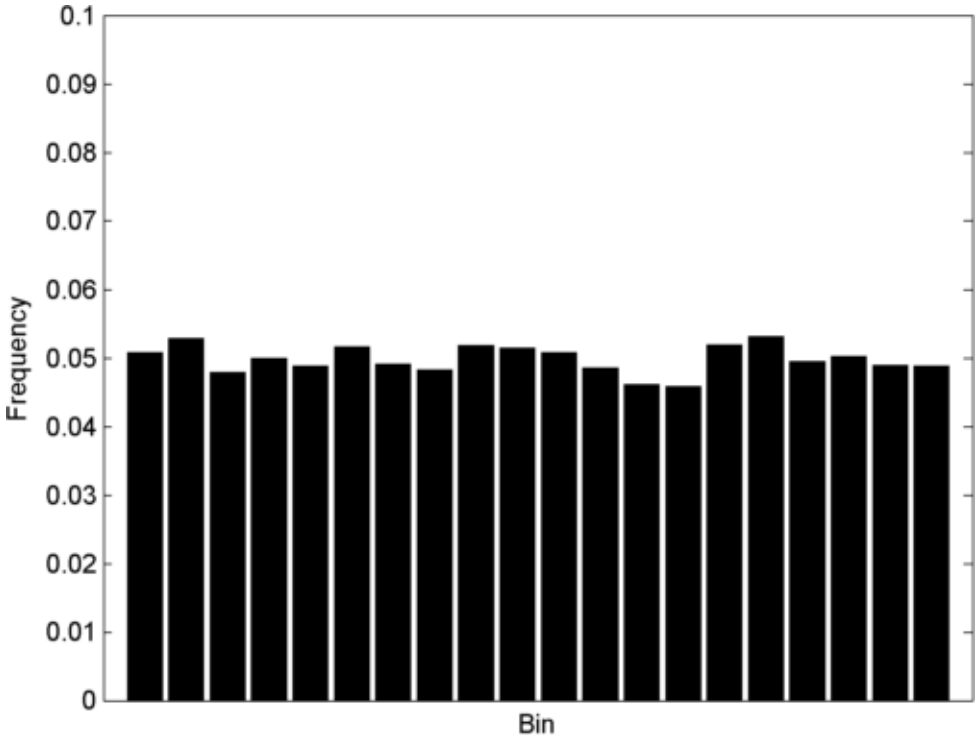

Fig. 2.2. Example rank histogram for a 20-member ensemble over 10,000 random forecasts.

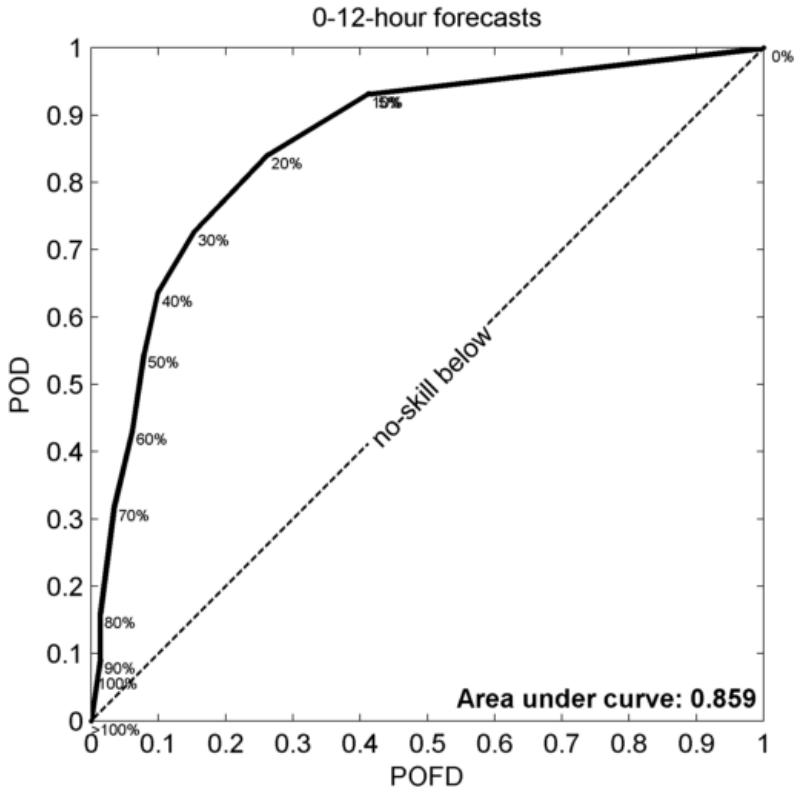

Fig. 2.3. Example ROC curve using the same forecast data used for Fig. 2.1. Percentages at each point refer to the probability threshold used to calculate each POD/POFD pair.

[1]Chapter 3. Comparing Methods of Accounting for Microphysics Uncertainty

3.1 Introduction

Research on methods of perturbing model physics in convective-scale ensemble forecasts is young. Although research related to convective-scale ensemble prediction has increased in the last few years (e.g., Kong et al. 2006, 2007ab; Xue et al. 2007; Mittermaier 2007; Clark et al. 2008, 2010; Schwartz et al. 2010; Vie et al. 2011; Johnson et al. 2011ab; Johnson and Wang 2012, 2013; Johnson et al. 2013ab; Ropnack et al. 2013; Caron 2013), many questions addressing the optimal design of convection-allowing model ensembles remain. Perhaps the most actively researched convective-scale features are those that produce precipitation. Specifically, convective storms and mesoscale convective systems, which are dominantly warm-season phenomena, are explicitly simulated without cumulus parameterization.

Warm-season QPF skill has historically been poor compared to that from the cold season (Ralph et al. 2005). Additionally, growth rates of forecast errors for convective scales can be highly nonlinear (Hohenegger and Schar 2007). Therefore, extensive research that optimizes convective-scale ensemble model forecasting to improve warm-season probabilistic quantitative precipitation forecasting is needed.

There are many ways to address model error in an ensemble. One particular way is to represent error stochastically (Buizza et al. 1999; Berner et al. 2009). Another way is by using multiple models and varied physics options within a given model (e.g., Doblas-Reyes et al. 2000; Hou et al. 2001; Ebert 2001; Hagedorn et al. 2005; Xue et al.

[1] This chapter is published as 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. Wea. Rev.*, **142**, 2198–2219.

2007, 2008, 2009, 2011; Kong et al. 2007b, 2010, 2011, Candille 2009; Berner et al. 2011; Hacker et al. 2011; Charron et al. 2011; Johnson et al. 2011ab; Johnson and Wang 2012). Since convection is explicitly represented by microphysics parameterizations in high resolution NWP simulations, warm-season QPF can be very sensitive to uncertainties in microphysics (MP) parameterizations. Earlier studies have investigated the sensitivity of convective-scale features like supercell thunderstorms and squall lines to complexity of, and the values of parameters used in, particle size distributions in microphysics schemes (Gilmore et al. 2004; Snook and Xue 2008; Tong and Xue 2008; Morrison et al. 2009; Dawson et al. 2010; Putnam et al. 2013). Great sensitivity of accumulated precipitation and cold pool size and intensity to parameters that describe rain, graupel, and hail size distributions has been found. It has also been found that simulations using multi-moment MP schemes produce more realistic reflectivity structures and more realistic stratiform precipitation regions in squall lines. However, most of these studies were conducted in an idealized and/or deterministic framework.

In this chapter the use of varied microphysics in a convection-allowing forecast in an ensemble and real data framework is examined. Two approaches to addressing model error from uncertainties in the microphysics are tested. In one approach the values of some parameters within a single MP scheme are varied. This way one can address uncertainty by sampling the distribution of possible values of parameters significant to cloud and precipitation physics. The perturbed parameters in this study include the intercept parameters for precipitation particle size distributions (PSD) and graupel density. The resultant ensemble is hereafter denoted as perturbed parameter

microphysics (PPMP). A single-moment MP scheme is used for testing this approach. These parameters were chosen based on earlier studies that examined the uncertainty ranges of PSD-related parameters and showed great sensitivities of modeled storm dynamics and precipitation forecasts to these parameters (Gilmore et al. 2004; Snook and Xue 2007; Tong and Xue 2008; Jung et al. 2010; Yussouf and Stensrud 2012). In the second approach separate MP parameterizations are used. Not only can this approach address uncertainty in parameter values, but it can also address uncertainty in the microphysical processes within the parameterization schemes. The resultant ensemble is hereafter denoted as mixed microphysics (MMP). Most of the MP parameterizations used in the MMP ensemble predict two moments of at least one of the precipitating species. Therefore this experiment also enables an investigation into whether an ensemble using various sophisticated MP parameterizations is superior to one using simpler and more computationally efficient MP parameterizations with perturbed parameters. This research presents a step in the investigation of how to best sample the microphysics errors in a convection-allowing ensemble for warm-season QPF.

The purpose of this research is two-fold. One purpose is to examine the effectiveness of different approaches to accounting for model microphysics error in convective-scale PQPF via conducting various verifications. The mixed results from a comparison of the MMP and PPMP ensembles as shown later motivated investigation of the combination of the approaches; the combination of the MMP and PPMP ensembles formed a third ensemble, called the pooled ensemble. The second purpose is to examine and compare the systematic behaviors of the various microphysical

variables over a broad range of cases for the MP schemes that comprise the MMP and PPMP ensembles in convection-allowing forecasts to facilitate ensemble design in the future.

3.2 Experimental setup

3.2.1 Model description

Version 3.2.1 of the Weather Research and Forecasting (WRF) model, Advanced Research WRF (ARW) dynamic core (Skamarock et al. 2008), was used as the NWP model. The WRF-ARW was the primary NWP model used in the Storm-Scale Ensemble Forecast system (SSEF) conducted by the Center for Analysis and Prediction of Storms and used in the NOAA Hazardous Weather Testbed (HWT) 2011 Spring Experiment (Kong et al. 2011). The 2011 Spring Experiment extended from late April to early June, which included 35 cases (generally one case each weekday from 27 April to 10 June; Kong et al. 2011). Forecasts were initialized at 0000 UTC and ran for 36 hours over the contiguous United States model domain (see Fig. 3.1).

Ten members were designed to investigate which approach to addressing model error due to microphysics uncertainty results in a more skillful ensemble. All other aspects of the model configuration, including initial and lateral boundary conditions, as well as other physics parameterizations, were identical among the ten members. The members used six different MP schemes including Thompson (Reisner et al. 1998; Thompson et al. 2008), Ferrier (Ferrier et al. 2002, 2003; Ferrier 2005), Morrison (Morrison et al. 2009), Milbrandt-Yau (M-Y; Morrison and Milbrandt 2011), WRF single-moment – 6 class (WSM6; Hong and Lim 2006a), and WRF double-moment – 6 class (WDM6; Lim and Hong 2010). Relevant characteristics and parameters of the

members and MP schemes are given in Table 3.1 and a description of the differences between the members is given in section 3.6. The MMP ensemble contained six members which used the Thompson, Ferrier, M-Y, Morrison, WDM6, and WSM6 schemes. The PPMP ensemble contained five members, each of which used the WSM6 scheme but with different values of the rain and graupel intercept parameter of the respective PSDs and different graupel densities. The graupel density ranges between values typical of graupel and hail, while the intercept parameters take values within observed uncertain ranges (see discussions in Tong and Xue 2008 on the uncertain ranges of values). The WSM6 scheme was a member of both ensembles. To investigate whether a combination of the above two approaches to addressing microphysics error is superior to either approach alone, a third ensemble – the "pooled" ensemble – was comprised of members from the PPMP and MMP ensembles.

Table 3.1. Description of the ten SSEF members used in this study. The first six members (in **bold**) comprised the MMP ensemble, whereas the bottom five members (in *italics*) comprised the PPMP ensemble.

Member name and scheme (perturbation)	# of classes[!]	Double moment species	N_{0r} (m^{-4})	N_{0g} (m^{-4})	ρ_g (kg m^{-3})	α_r	α_g
Thompson	6	rain, ice	-	varies	400	0	0
Ferrier+[o]	3	-	$8x10^6$	-	-	-	-
Morrison	6	rain, graupel, snow, ice	-	-	400	0	0
Milbrandt-Yau*	7	all but vapor	-	-	400	2	0
WDM6	6	rain, cloud water	-	$4x10^6$	500	1	0
WSM6 (control)	6	-	$8x10^6$	$4x10^6$	500	0	0
WSM6 (M1)	6	-	$8x10^6$	$4x10^4$	913	0	0
WSM6 (M2)	6	-	$8x10^7$	$4x10^6$	500	0	0
WSM6 (M3)	6	-	$8x10^5$	$4x10^2$	913	0	0
WSM6 (M4)	6	-	$8x10^5$	$4x10^3$	913	0	0

*Milbrandt-Yau scheme adds hail as a separate class with ρ_h = 900 kg m^{-3} and α_h = 3
[o]Ferrier+ scheme treats all condensate species as one
[!]Water classes (species) include vapor, liquid cloud drops, frozen cloud ice particles, rain, snow, graupel, and hail. Hail was the least common species included.

The initial conditions were generated by the Advanced Regional Prediction System three-dimensional variational data assimilation and cloud analysis system (Gao et al. 2004; Xue et al. 2003; Hu et al. 2006) using 0000 UTC operational 12 km North American Mesoscale (NAM) model analyses as the background field. The data assimilated include Weather Surveillance Radar – 1988 Doppler radial velocity and reflectivity, surface pressure, horizontal wind, temperature, and specific humidity from the Oklahoma Mesonet, surface aviation observations, and wind profiler data. The boundary conditions were provided by 12-km NAM model forecasts initialized at 0000 UTC. The Noah land-surface model (Ek et al. 2003), and the MYJ boundary-layer

parameterization (Mellor and Yamada 1982; Janjić 2002) were used for all ten members.

3.2.2 Verification methodology

Verification was performed over a portion of the model domain containing the central and southern United States (Fig. 3.1). This domain was chosen because the environment in this region during the late spring and early summer supports organized and intense convection with tight spatial gradients. Therefore the verifications in this region are a representative evaluation of the skill of convective-scale warm-season QPF. A variety of verification metrics were used, including grid-point-based, contingency-table-based, and neighborhood-based metrics. These metrics are introduced as they are encountered. Probabilistic verification was computed using uncalibrated model output. Therefore, probability of precipitation (PoP) was computed as the proportion of members in which a specified 1-hour precipitation amount threshold[2] was exceeded. PoP values across the verification domain constituted the probabilistic QPFs (PQPF) verified in this study.

Verifying precipitation data were provided by the National Mosaic and Multi-Sensor Quantitative Precipitation Estimation (QPE) project (NMQ) of the National Severe Storms Laboratory (Zhang et al. 2011). The NMQ QPEs were regridded to the verification domain using bilinear interpolation. Verifications were performed on one-hour accumulated precipitation fields. The NMQ QPEs have been used in other studies to verify storm scale precipitation forecasts (e.g., Johnson and Wang 2012, 2013; Johnson et al. 2013ab).

[2] The thresholds tested included, in mm, 0.254, 2.54, 6.35, 12.7, and 25.4, which convert to the following in inches: 0.01, 0.10, 0.25, 0.50, and 1.00.

The MMP, PPMP, and pooled ensembles contained six, five, and ten members, respectively. Earlier studies have shown that verification scores can be sensitive to ensemble size (Richardson 2001). The purpose of the study is to investigate different methods of accounting for microphysics scheme errors in a storm-scale ensemble. To minimize the impact of different ensemble sizes on this purpose, a single ensemble size was used for verification. Since the PPMP ensemble had the smallest size at five members, five members were selected randomly from each of the MMP and pooled ensembles to constitute a resample. One-hundred such resamples were obtained for each of the MMP and pooled ensembles. The verification scores were averaged over these resamples for the MMP and pooled ensembles before being compared with the scores from the PPMP ensemble. The choice of 100 resamples follows earlier studies using bootstrap resampling techniques in statistical significance tests (e.g., Wang and Bishop 2005). In addition, each five-member resample from the pooled ensemble was selected such that each resample was randomly composed of either two or three members from the MMP and PPMP ensembles to ensure each of the MMP and PPMP ensembles were equally represented in the resample of the pooled ensemble.

3.3 Verification of quantitative precipitation forecasts

3.3.1 Verification of individual members

The existence of particularly skillful or unskillful members could impact the overall ensemble performance (Ebert 2001). Therefore, individual members were first verified. Fig. 3.2 shows the average hourly precipitation forecast errors. A periodic trend, featuring local maxima around 0000 UTC (0- and 24-hour forecasts) was observed. This trend has been commonly observed in the SSEF in the past (e.g., Clark

et al. 2009; Johnson and Wang 2012) and in other verification scores in this chapter. The relative maxima around forecast hours 0 and 24 identify "convection maxima", or diurnal peaks in convective activity around 0000 UTC, which corresponds to 7 or 8 PM local time in the central and eastern U.S. The convection maximum around forecast hour 0 will hereafter be referred to as the first convection maximum, while that around forecast hour 24 will be called the second convection maximum. The convection minima, on the other hand, refer to the reduced convective activity centered near forecast hours 12 and 36.

Significant variation among members accompanied the mean precipitation errors in the first few forecast hours (i.e., f01[3]-f04) as each member adjusted from the initial conditions that contained cloud and hydrometeors created by the cloud analysis with radar reflectivity data assimilation (Fig. 3.2). At the first convection maximum, the Thompson and M-Y schemes produced too little precipitation whereas all other schemes produced too much. The overproduction was very large in the WDM6, WSM6 (control), and WSM6-M1 schemes. All members produced too little precipitation during the convection minima. During the second convection maximum large variation was found in the mean errors as some members produced too much precipitation (all from the PPMP ensemble) whereas others produced too little. In particular the WSM6-M3 and WSM6-M4 members had a larger positive bias in the f18-f24 period while the M-Y scheme had a much larger negative bias in the f18-f30 period.

The frequency biases (the ratio of the number of grid points at which forecast precipitation exceeded a threshold to the number of grid points at which precipitation was observed to have exceeded that threshold) of the WSM6-M3 and WSM6-M4

[3] f01 means forecast hour 1. This symbolism will be used throughout the remainder of the chapter.

schemes were not appreciably higher than those of other members at f18-f24 for lower thresholds (Fig. 3.3), but they increased significantly for the higher thresholds, indicating those members over-forecast heavy precipitation episodes. The overall negative error for the M-Y scheme is consistent with its frequency biases, which were almost exclusively below 1.0, indicating a consistent trend to under-forecast precipitation at all thresholds. The increased variability in frequency biases among the different schemes at high thresholds is related to the counts in each cell of the 2x2 contingency table. It is rare for 12.7 mm or 25.4 mm of precipitation to fall in one hour, occurring only within the cores of stronger thunderstorms. Therefore, the contingency table has small numbers of hits, false alarms, and misses at these thresholds – a few thousandths of a percent of the total number of grid points at which verification was performed. Since frequency bias depends on the number of hits, false alarms, and misses, seemingly large differences in frequency bias can appear between different schemes.

Regarding equitable threat scores (ETS), a large decrease in skill from very high values within the first one or two forecast hours was observed (Fig. 3.4). This higher skill at early forecast hours can be attributed to the initialization of pre-existing convection within the initial condition resulting from the assimilation of radar data (e.g., Xue et al. 2008; 2013), the effects of which wore off quickly. The benefits to assimilating radar data for very short range convective scale QPF on the order of a few hours have been observed in previous SSEF forecasts (e.g., Johnson and Wang 2012; Kong et al. 2011; Xue et al. 2008, 2009, 2011, 2013). ETSs were very similar among the members with a few exceptions; the Thompson and Morrison schemes had higher

ETSs than the other members for most forecast hours at the lightest threshold. The improved performance of those schemes dwindled with increasing threshold. No member appeared to perform worse except for the WDM6 scheme around f30. ETSs for the PPMP members were tightly clustered, and no single member stood out except for around f06, when the WSM6-M3 and WSM6-M4 schemes were slightly inferior at moderate thresholds. ETSs were highly variable at the highest thresholds, owing to the decreased numbers of hits, false alarms, and misses in the contingency table due to limited samples.

Verification metrics based on grid-point values such as the ETS may give a misleading interpretation of the skill of high-resolution precipitation forecasts due to the small horizontal scale of features compared to the horizontal scale of acceptable spatial errors (Baldwin et al. 2001). For that reason it is more appropriate to consider neighborhood-based verification metrics. One such measure of neighborhood-based verification is applied to the Brier score and is referred to as the fractions Brier score (FBS; Roberts and Lean 2008; Schwartz et al. 2010). Despite being a probabilistic verification metric, the FBS can be computed for deterministic forecasts by constructing probabilistic forecasts based on spatial neighborhoods (Theis et al. 2005). A square neighborhood of radius 48 km (12 grid points) was used for all verifications in this chapter applied to a neighborhood (Johnson and Wang 2012), including those for both deterministic and ensemble forecasts. Thus, FBS is a mean square difference between the forecast PoP in a spatial neighborhood and the proportion of observed precipitation events in a neighborhood around a given point. A diurnal cycle was present in the FBSs, and the highest FBSs (poorest performance) occurred at the second convection

maximum for low thresholds and at the first convection maximum for the highest thresholds (Fig. 3.5). The Thompson and M-Y schemes were commonly the best for a given threshold and forecast lead time. However, the WSM6-M3 and WSM6-M4 schemes had better FBSs around the second convection maximum at the 2.54-mm threshold. The WDM6 scheme commonly had the worst FBSs around the second convection maximum, except at the highest threshold, indicating consistently poor performance of that member around that time of forecast. Among the PPMP members, the WSM6-M2 scheme typically had the worst FBS around the convection maxima, while the WSM6-M3 scheme frequently had the worst FBS away from the convection maxima.

The area under the relative operating characteristic curve (ROC; Mason 1982) for each member was computed using neighborhood forecasts verified against single-point observations. The ROC areas indicate that the M-Y scheme was commonly the worst performing scheme, especially around the second convection maximum (Fig. 3.6). The Thompson and Morrison schemes had larger (better) ROC areas for the lower thresholds, but the WSM6-M3 and WSM6-M4 schemes had larger areas at the highest thresholds and for later forecast hours. For the lower thresholds the envelope of scores from the MMP ensemble members tended to contain those of the PPMP ensemble members, implying more variability in skill within the MMP ensemble than within the PPMP ensemble. At higher thresholds, however, the ROC areas of the PPMP ensemble members tended to be larger than most of those of the MMP ensemble members.

The metrics discussed here indicate that members of the MMP ensemble commonly under-predicted both the amount and areal coverage of precipitation,

especially at high thresholds. In contrast, members of the PPMP ensemble were generally less biased. However, these PPMP members were also generally less skillful, with a few exceptions. Superior skill was frequently noted with the Thompson and Morrison schemes in agreement with Clark et al. (2012), and inferior performance was commonly seen with the WDM6 scheme, although the ordering of skill was somewhat a function of metric, forecast lead time, and threshold. The WSM6 scheme frequently ranked near the middle of all schemes and near the middle of the WSM6 perturbation schemes.

3.3.2 Verification of the MMP, PPMP, and pooled ensembles

All of the following discussion refers to the mean values of verification scores from the 5-member resamples except where otherwise noted. The significance of the difference was determined using the standard deviation of the scores from the resamples, similar to the method used in Wang and Bishop (2005). Here, means differing by three standard deviations or more are declared to represent ensembles having statistically significantly different skill. The larger value of the standard deviations between the ensembles compared was used. After about f04, the magnitude of the mean error of the ensembles was around one order of magnitude less than the season-average rain amounts per grid point, which ranged from 0.10 to 0.30 mm (Fig. 3.7b). This was especially true during f12–f27. For early and late forecast hours the errors were about of the same order of magnitude as the average precipitation amount at each grid point, which indicates that bias at these forecast hours could be significant. The ensembles were biased low during f04–f18 and f24–f36 (Fig. 3.7a). The bias was variable among different ensembles during f18–f25 and in the first 3 hours of the

forecast. For the first 3 hours the ensembles produced too much precipitation, with the PPMP ensemble producing the most precipitation and the MMP ensemble being the least positively biased. During f21-f23, the PPMP and MMP ensembles had similar absolute biases, with the MMP ensemble biased low and the PPMP ensemble biased high. During f18-20 and f24-25, the PPMP ensemble was less negatively biased than the MMP ensemble. The exceptionally low bias of the M-Y scheme is the leading cause of this low bias in the MMP ensemble, as the remaining members of the MMP ensemble were less biased during that period. The bias for the pooled ensemble was, as expected, between that of the MMP and PPMP ensembles. Around the second convection maximum the pooled ensemble had a nearly zero mean error. The bias for the full 10-member pooled ensemble was nearly identical to that of the resampled pooled ensemble. The same was true for the full 6-member MMP ensemble compared to the resampled MMP ensemble. The frequency biases were consistent with the mean errors (not shown), regardless of precipitation accumulation thresholds.

The reliability of the ensembles was evaluated through examination of rank histograms (Hamill 2001). Easily recognizable signatures that represent the dispersion and bias characteristics of ensembles can be illustrated by rank histograms. Note that to reduce the impact of ensemble size on evaluating the flatness of the rank histogram, resampled ensembles with an ensemble size of five were used. The under-dispersive nature of all ensembles is apparent (Fig. 3.8). The under-dispersion was generally worse during the first 10 forecast hours and better around the second convection maximum. The left skew of the histograms is consistent across all forecast hours with varying degrees of magnitude. Also, while it has been shown that a composite of signals may

result in rank histograms that give a misleading impression of the reliability of an ensemble (Hamill 2001), rank histograms computed from a few localized subsets and a scattered subset of the domain (counts were obtained from grid points widely separated in space) revealed the same signatures as those using all data points. Therefore the aggregated signatures are unlikely to be the result of a combination of high bias in some regions of the domain and low bias in other regions. Although all ensembles are under-dispersive, a way of quantifying the extent of under-dispersion is by the proportion of values that fall within the outer bins of the histogram normalized by the proportion of values expected to be located in each bin of a flat histogram (*ext*; upper-right corner of each panel in Fig. 3.8). The values of *ext* indicate that the MMP and PPMP ensembles were under-dispersive to about the same extent, although the value of *ext* indicates that the MMP ensemble was slightly better than the PPMP ensemble. The value of *ext* for the pooled ensemble is in between that of MMP and PPMP. Note that in this study, the ensembles only accounted for the microphysics scheme errors. Early work analyzing the SSEF that included perturbations to represent other sources of forecast errors has shown flatter histograms (cf. Fig. 8a of Clark et al. 2009; Fig. 15 of Xue et al. 2011).

We now consider the neighborhood-based FBS. A diurnal cycle was observed with the FBSs. For the highest two thresholds the highest (worst) scores occurred during the first convection maximum, whereas for the remaining thresholds the worst scores occurred around the second convection maximum (Fig. 3.9). This behavior is identical to that seen in the FBSs for the individual members (section 3.3.1). The best FBSs occurring at f01-f02 were likely due to the improved initial conditions due to the assimilation of radar data. At most thresholds and forecast lead times the MMP

ensemble performed better than the PPMP ensemble. The biggest exception was around the second convection maximum for the 2.54-mm and 6.35-mm thresholds, when the skill of the MMP ensemble was equal to that of the PPMP ensemble. The FBSs of the pooled ensemble were generally smaller than those of the PPMP and nearly identical to those of the MMP ensemble. The largest difference between the MMP and pooled ensembles occurred around the convection maxima at the highest two thresholds where the MMP ensemble had statistically significantly better FBSs than the PPMP ensemble. The FBSs of the full 10-member pooled ensemble were generally similar to those of the resampled pooled ensemble. However, the full 10-member pooled ensemble was generally better around the second convection maximum for the lower thresholds, thus showing the positive impact from using a larger ensemble during periods of intense convective activity at longer lead times.

Probabilistic forecasts derived from spatial neighborhoods are further evaluated by computing the area under ROC curves. Scores decreased only slightly with increasing threshold and forecast hour (Fig. 3.10), and all scores were greater than the no-skill value of 0.5. The spread among resampled scores was small. This is probably due to the larger number of samples used to compute scores (several thousand for neighborhood verification). At the lowest two thresholds the MMP ensemble was generally more skillful than the PPMP ensemble. At the highest two thresholds the PPMP ensemble was more skillful than the MMP ensemble. At the low thresholds the ROC area of the pooled ensemble was nearly identical to that of the MMP ensemble. At the high thresholds the ROC area of the pooled ensemble was between that of the MMP and PPMP ensemble with better scores than the MMP for most lead times. These results

are consistent with the ROC area for the individual members (section 3.3.1). The ROC area for the full 10-member pooled ensemble was larger than that for the other ensembles at all forecast hours and precipitation thresholds, suggesting the positive impact of using a larger ensemble. This result is consistent with Clark et al. (2011).

The ensemble verification presented here suggests that the MMP ensemble was more skillful than the PPMP ensemble according to FBSs for most lead times and thresholds and less skillful than the PPMP at high thresholds according to ROC areas. While this result may seem contradictory, each metric assesses different aspects of the ensemble PQPF. The FBS measures the mean squared error of the neighborhood PQPF relative to the neighborhood coverage of observed precipitation, whereas the ROC measures the ability of the ensemble to distinguish between "yes" and "no" forecasts for precipitation exceeding a threshold. The pooled ensemble did not outperform the MMP or PPMP ensemble except for at a small number of thresholds, forecast hours, and metrics.

The resampling allowed for an examination of the impact of ensemble size in the quality of the forecasts. Differences between the full 6-member and the 5-member resampled MMP ensemble were generally imperceptible. However, when differences were found it was generally the full 6-member MMP ensemble that was superior. There were larger, more noticeable differences between the full 10-member and 5-member resampled pooled ensembles than between the full and resampled MMP ensembles. The full 10-member pooled ensemble was always better than the resampled pooled ensemble.

The preceding discussion implies that the relative skill of using a combination of mixed MP schemes and perturbing parameters within a single MP scheme to sample the microphysics errors can be dependent on methods of verification and the precipitation thresholds. Using a combination of the two methods generally is no more skillful than using either approach separately. Ensemble size seems to have as large an impact on the quality of the forecasts as the method used to construct the ensembles.

3.4 A comparison of the microphysical parameters

Microphysics parameters such as the intercept parameters for PSDs and particle density have been shown to impact the simulation of convective systems. Specifically, Gilmore et al. (2004), Snook and Xue (2008), Dawson et al. (2010), and Morrison and Milbrandt (2011) have shown that the dynamics and precipitation patterns in supercells, in particular, are sensitive to the values of the rain and hail intercept parameters, graupel and hail density, and the number of predicted moments of the PSD for precipitating species in bulk microphysics schemes. Motivated by these earlier studies, two methods to account for microphysics scheme errors in convection allowing ensembles were tested within the CAPS SSEF system in 2011. In the PPMP ensemble, the rain and graupel intercept parameters and the graupel density were perturbed within a single microphysics scheme. Alternatively, the MMP ensemble contained a variety of microphysics schemes. To facilitate developing a strategy to optimally account for microphysics scheme errors in the ensemble, in addition to evaluating the ensembles using objective verification scores, we examined the systematic behaviors of various microphysical variables from the PPMP and MMP ensemble members. The following analysis is meant to answer the following questions: How did the prescribed intercept

parameters for rain and graupel in the WSM6 schemes compare to those retrieved from the double-moment microphysics schemes? How different are the microphysics parameters between single and double-moment schemes? How did the WDM6 and WSM6 schemes in particular compare?

Using the method discussed in section 3.6, various parameters in these MP schemes were retrieved from the prognosed moments' fields. Parameters were not retrieved from the Ferrier+ scheme due to significant differences between that scheme and the others regarding the treatment of the water species. Numerous additional assumptions regarding the rain and graupel PSDs would be required to retrieve parameters from the Ferrier+ scheme. Since the graupel intercept parameter is diagnosed from the mixing ratio internally in the Thompson scheme, the graupel intercept parameter was not retrieved from it. We chose specifically to examine mean mass particle diameter, particle surface area, number concentration, mixing ratio, and intercept parameter for the rain and graupel PSDs. The former two parameters serve as proxies for the tendencies for rain evaporation and cold pool formation (Dawson et al. 2010) which can have a large impact on the organization and lifetime of convective systems (Rotunno et al. 1988; Snook and Xue 2008).

Vertical profiles of horizontally averaged microphysical parameters for rain are shown in Fig. 3.11, while those for graupel are shown in Fig. 3.12. We only consider rain and graupel since the PPMP ensemble only perturbed the parameters associated with these two species. The average was computed over only those grid boxes at which the mixing ratio exceeded 10^{-6} kg kg^{-1}, a general threshold to distinguish precipitating from non-precipitating grid boxes (the counts can be seen in Figs. 3.11f and 3.12f). The

profiles were also averaged over all forecast hours. For rain the discussion will focus generally below 3 km, a common height of the freezing level during the experiment, where rainfall is the dominant precipitation type. On the other hand, the discussion for graupel will focus on the lowest 10 km.

As pointed out earlier, the prescribed values of intercept parameters in the PPMP members were chosen based on values reported in the literature (Snook and Xue 2008; Tong and Xue 2008; Yussouf and Stensrud 2012). Fig. 3.11d shows that the effective rain intercept parameters sampled by the MMP members generally laid in a similar range as those prescribed to the members of the PPMP scheme, although that for the WDM6 scheme is anomalously high, and the $8x10^5$ m^{-4} value used for the PPMP members seems somewhat low. Other microphysical variables such as raindrop size, total raindrop surface area, rain number concentration and mixing ratio sampled by the MMP members also in general laid in the same range as those in the PPMP ensemble (Fig. 3.11). The Morrison scheme was the only scheme to give meaningful results for the graupel intercept parameter (Fig. 3.12d), and the values from that scheme line up nicely with the values from the unperturbed WSM6 scheme ($4x10^6$ m^{-4}) except for the vertical variation of nearly three orders of magnitude. Except for the Morrison scheme, which exhibited different behavior than the other schemes, the WSM6 schemes in the PPMP ensemble tended to produce less of the various graupel quantities examined, including surface area, number concentration, and mixing ratio (Fig. 3.12). The Thompson scheme produced the least graupel mixing ratio, however (Fig. 3.12e). The effective graupel intercept parameters retrieved from the Morrison scheme most closely

resemble those of the control value of 4 x 10^6 m^{-4} (Fig. 3.12d), suggesting that the perturbed values of the graupel intercept parameters may be too low.

Among the single-moment members that comprised the PPMP ensemble, the behavior was found to be consistent with the prescribed intercept parameters. WSM6-M2 was designed to represent small raindrops. For a given mixing ratio, mean drop size decreases with increasing intercept parameter. WSM6-M2 was the only scheme for which the rain intercept parameter was perturbed high compared to the control value (Table 3.1). While the rain mixing ratios are not identical between the single-moment members, differences of approximately 1 to 2 orders of magnitude are required for this rule to be violated. That was not the case at any level or time in this experiment (Fig. 3.11e). Similarly, the larger rain drop sizes for WSM6-M3 and WSM6-M4 compared to those of the WSM6-M2 and WSM6 (control) schemes are also consistent given that the rain intercept parameter for WSM6-M3 and WSM6-M4 is lower than it is for WSM6 (control) and WSM6-M2. Given the similarity in mixing ratio, it is also sensible that WSM6-M2 also had the largest number concentration of rain and the largest total raindrop surface area, while WSM6-M3 and WSM6-M4 had smaller number concentrations and surface areas. The vertical profiles of rain number concentration and raindrop surface area were very similar between pairs of members (e.g., M3 compared to M4 and control compared to M1) that used the same set of intercept values because these parameters are derived from the only free moment, mixing ratio. Slight differences in the mixing ratios correspond to slight differences in these other retrieved parameters. Similar arguments apply to the retrieved graupel variables. WSM6-M3 and WSM6-M4 were meant to contain large hail-like particles. Therefore it makes sense

that these schemes have the largest graupel (Fig. 3.12a), smallest number concentration (Fig. 3.12c), and smallest surface area (Fig. 3.12b). Similarly, WSM6 and WSM6-M2 contained smaller and more numerous graupel particles. Interestingly, the graupel mixing ratios above about 3 km were ranked according to prescribed graupel intercept parameters for the WSM6 schemes.

The profiles from the double-moment schemes contained more variation in the vertical than those from the single-moment scheme members (Figs. 3.11 and 3.12). This behavior is likely the result of the ability of the double-moment schemes to account for size sorting of rain and graupel and other processes causing size distribution to change, since they contain separate predictive equations for the additional moment of number concentration (see, for example, Riesner et al. 1998, Milbrandt and Yau 2005, Morrison et al. 2005, and Lim and Hong 2010). The overall increase in mean mass diameter of raindrops towards the surface in the double-moment schemes is an indication that smaller raindrops are evaporating during their descent from the mid-levels; some completely evaporate before reaching the ground. This is consistent with the decrease of number concentration and surface area towards the surface, although there is some anomalous behavior seen in the Thompson scheme, where the number concentration and intercept parameter increase in the lowest two to three model levels (Fig. 3.11c,d). The rain mixing ratio generally decreases from a height of about 3 km towards the ground. The general decrease supports the notion that rain is evaporating in unsaturated downdrafts or due to entrainment of unsaturated environmental air. Graupel mixing ratio and number concentration also decrease below about 3 km (except for Thompson; Fig. 3.12e) as graupel melts below the freezing level. The Morrison scheme is the only

one that predicts number concentration for graupel that gave meaningful results. In the Morrison scheme, only larger particles are left below the freezing level as the smaller particles completely melt (the lavender trace in Fig. 3.12a). Among the double-moment schemes considered, WDM6 produced the smallest rain drops and very large rain number concentrations compared to the other schemes. This behavior of WDM6 has been noted and modifications to the scheme to reduce the number concentration are underway by the scheme developers (K.-S. Lim 2012, personal communication). The M-Y scheme tended to produce larger raindrops and larger rain surface area than the Morrison scheme. This is opposite of what was found in Morrison and Milbrandt (2011) regarding comparisons between those schemes. However, the versions of the schemes used in that study are slightly different from those used in this study. Namely, the shape parameters of the drop size distribution differ ($\alpha_r = 0$ was used in the M-Y scheme of Morrison and Milbrandt (2011), whereas $\alpha_r = 2$ in this study). However, the graupel parameters examined in this study are identical to those in Morrison and Milbrandt (2011), and the similar mixing ratios between the MY and Morrison schemes in this study disagree with the results of Morrison and Milbrandt (2011), which found that the MY scheme tended to produce more graupel. Also, above 3 km, the MY and Morrison schemes produced more graupel mixing ratio (Fig. 3.12e) than the single-moment schemes for graupel.

The design of the WDM6 and WSM6 schemes makes them more similar to each other than any other pair of schemes in this study (not including the WSM6 schemes with perturbed parameters). The major differences between the WDM6 and WSM6 schemes include the number of moments predicted for rain, the shape parameter of the

rain drop size distribution, and the addition of cloud condensation nuclei as a prognostic variable in the WDM6 scheme. There are also minor differences in how some microphysics processes like accretion rates are parameterized. The retrieved rain variables for WDM6 and WSM6 were quite different, however. The WDM6 scheme produced the smallest rain drops overall, corresponding to the largest drop concentration and surface areas as well. It also produced more rain mixing ratio, slightly greater radar reflectivity (not shown), but similar domain-accumulated precipitation (also not shown). Since the graupel distributions had the same shape parameter, differences in the graupel variables were reduced compared to those in the rain variables. The WSM6 scheme tended to produce less graupel mixing ratio near the surface (Fig. 3.12e). This is likely related to how some of the graupel and rain processes were parameterized. It is beyond the scope of the study to isolate the specific impact from incorporating the major differences one at a time. However, the results herein are generally in agreement with those obtained in Lim and Hong (2010) in which the same versions of both schemes were compared.

3.5 Summary and conclusions

Two approaches to accounting for uncertainty of the MP parameterization in warm-season QPF in a convection-allowing ensemble were examined. The two approaches include a set of completely different MP parameterizations (MMP) and a set of different numerical values of a set of parameters within a single MP scheme (PPMP). The combination of the two approaches was also tested. These approaches were implemented within the storm-scale ensemble framework of the Center for Analysis and Prediction of storms at the University of Oklahoma, and tested during the NOAA HWT

2011 Spring Experiment (27 April – 10 June 2011). An ensemble of convection-allowing WRF-ARW simulations were run in real-time at 4 km resolution over the contiguous United States for 35 cases during the period. The two ensembles tested – MMP and PPMP – along with an ensemble comprised of members from both ensembles, had six, five, and ten members, respectively (one member belonged to both ensembles). To minimize the impact of ensemble size on the comparison between approaches to accounting for microphysics errors, scores for the MMP and pooled ensembles were obtained from an average of 100 five-member resamples.

It was found that, in general, the MMP ensemble was more skillful than the PPMP ensemble (based on WSM6) with variations dependent on the metric chosen, the forecast lead time, and precipitation threshold. The MMP ensemble was more skillful for FBSs for most lead times and thresholds, but the PPMP ensemble was more skillful for high thresholds when ROC areas were examined. The rank histograms of the MMP ensemble were slightly flatter than those of the PPMP and pooled ensembles. The MMP ensemble was also somewhat more biased, but mostly due to one member. Also, the combined approach where both sophisticated MP schemes and perturbed parameters within a simpler single-moment MP scheme were used was no more skillful than the better of the MMP and PPMP ensembles. Since all PPMP members were based on WSM6, the skill of those members was limited by the skill of WSM6. Therefore, the results obtained here may not apply to perturbed parameter ensembles generated by perturbing other microphysics schemes. The similarity in the performances of the MMP and PPMP ensembles was sensible, however, since the WSM6 scheme frequently

ranked in the middle of the microphysics schemes considered and near the middle of the WSM6 perturbation schemes as shown in section 3.3.1.

The QPF from individual members was also examined. Verification scores varied more among members of the MMP ensemble than among members of the PPMP ensemble for low and moderate thresholds. Large variation was found in the bias during the first and second convection maxima. The WSM6-M3 and WSM6-M4 schemes showed the largest positive bias in the mean errors and in the frequency biases at highest two thresholds during the second convection maximum. The M-Y scheme showed the largest negative bias. PPMP members were generally less biased, but were also less skillful at all but the highest thresholds. The skill of individual members depended on the thresholds and methods of verification. For example, for ETSs, the Thompson and Morrison schemes were the most skillful at the lightest two thresholds whereas the WDM6 scheme was the least skillful at low and moderate thresholds at later lead times. For FBSs, the Thompson scheme showed the highest skill at moderate thresholds, and both Thompson and Morrison showed the highest skill at the highest threshold. The Ferrier+ and WDM6 schemes were the least skillful at moderate thresholds at the second convection maximum. For ROC areas, the WSM6-M3 and WSM6-M4 schemes were the best at the highest thresholds. The M-Y scheme was the least skillful at the second convection maximum.

In addition to evaluating the ensembles using objective verification scores, the systematic behaviors of the members using retrieved microphysics parameters relevant to this study were examined. The range of the retrieved rain intercept parameters from the double-moment members is, in general, consistent with the range of perturbed rain

intercept parameters for the PPMP members. The retrieved graupel intercept parameter values suggest the perturbed graupel intercept values used in the PPMP members were too low. The behavior of the PPMP members was found to be consistent with the prescribed values of the rain and graupel intercept parameters and graupel density. Behavior was more variable for the double-moment members, and causes of differences in retrieved parameters among these MP schemes are more complex than differences among the single-moment members owing to the differences in parameterizations of individual processes within each scheme. The WDM6 scheme tended to produce a very large number of small raindrops, whereas the M-Y scheme tended to produce very large raindrops.

Although resampling of the pooled ensemble with an equal size as the PPMP and MMP ensembles was no more skillful than the better of the PPMP and MMP ensembles, the full 10-member pooled ensemble showed better skill than other ensembles tested for nearly all lead times, thresholds, and metrics. This result brings up an interesting point in that increasing ensemble size had a significant impact in the skill of the ensembles tested (see also Clark et al. 2011; Ebert 2001). One easy way to populate the ensemble is to perturb the parameters for a given physical parameterization scheme such as the PPMP.

This study represents one step towards developing a strategy for the optimal design of a convection-allowing ensemble prediction system. Ideally, members of an ensemble should represent random draws from the probability distribution of truth for a given forecast system and each member should be equally possible to represent the truth. The design of PPMP and MMP ensemble may violate such rules. The parameters

used in PPMP ensemble were not randomly perturbed. The MMP ensemble is composed of different microphysics schemes with each scheme having its own bias. In addition, only the uncertainty in QPF associated with varied microphysics was tested, but uncertainty in QPF from model error can also come from errors in other physics parameterizations such as the boundary-layer, radiation, and surface-atmosphere exchangegb. Since only QPF was examined, it cannot be determined whether either approach is more skillful at predicting other meteorological fields such as surface temperature, relative humidity, wind speed, or geopotential height at other levels in the troposphere. Thus future studies should also examine the impacts of varied microphysics on other fields. The specific conclusions may also be dependent on the specific choices of the MP schemes used in the ensemble, and to the way parameters are perturbed. Future studies should investigate ways to use variations in these parameterizations to better design convection-allowing model ensembles, as their use will be more strongly desired in the future for operational weather forecasting purposes.

3.6 Appendix – microphysics vocabulary

Except for the Ferrier+ scheme, all of the MP schemes used in this experiment use similar assumptions for the PSDs of the various classes. They all use the following model:

$$n_x(D_x) = N_{0x} {D_x}^{\alpha_x} e^{-\lambda_x D_x} \quad (3.1),$$

where subscript x refers to one of the water species (r – rain, g – graupel, i – cloud ice, s – snow, h – hail, c – cloud water; the subscript will be dropped hereafter), n is the volumetric concentration of spherical particles of diameter D to D + ΔD, N_0 is the intercept parameter, λ is the slope parameter, and α is the shape parameter. Note that

when α = 0, this reduces to the inverse-exponential distribution. The p^{th} moment of the PSD is given by

$$M(p) = \int_0^\infty D^p \, n(D) dD = \frac{N_0}{\lambda^{\alpha+p+1}} \Gamma(\alpha + p + 1) \quad (3.2),$$

where Γ is the gamma function and $\Gamma(Z) = (Z-1)!$ when Z is an integer. Three moments would be required to fully characterize this PSD without making any assumptions. Typically the three moments include the mixing ratio – which is proportional to the 3^{rd} moment of the PSD, number concentration – the 0^{th} moment of the PSD, and reflectivity factor – the 6^{th} moment. Each moment requires its own set of conservation equations for each species, which increases the complexity of the scheme and also increases computation time. However, use of a greater number of moments has been shown to increase the realism of simulations of squall-lines at convection-allowing resolutions (Morrison et al. 2009). Four schemes in this study – Thompson, Morrison, M-Y, and WDM6 – use two moments for some species. All are double-moment for rain, and two (Morrison and M-Y) are double-moment for graupel as well. Species mixing ratio and number concentration are the two moments used in these schemes. While the other schemes are only single-moment, other moments can be derived using manipulations of (3.1) and (3.2). For instance, the number concentration can be determined by

$$N_t = \left\{ [N_0 \Gamma(\alpha+1)]^d \left[\frac{q \rho_a}{c} \frac{\Gamma(\alpha+1)}{\Gamma(\alpha+4)} \right]^{\alpha+1} \right\}^{1/\alpha+d+1} \quad (3.3),$$

where ρ_a is air density and c and d are related through the mass-diameter relationship: $m(D) = cD^d$ and q is the species mixing ratio. In this chapter we retrieved the intercept parameter for the double-moment schemes using (3.3). However, the meaning of the

intercept parameter becomes non-physical when the shape parameter is nonzero. Therefore, the "normalized" intercept parameter, which is the intercept parameter for an exponential distribution with the same liquid water content and mean volume diameter (Testud et al. 2001), is used instead. The total surface area of particles is proportional to M(2):

$$Surface\ area = \pi M(2) = \frac{\pi N_t(\alpha+2)(\alpha+1)D_m^2}{\left(\frac{\Gamma(\alpha+4)}{\Gamma(\alpha+1)}\right)^{2/3}} \quad (3.4),$$

where it has been assumed that α is an integer and d = 3, which is true for both rain and graupel, and D_m is the mean mass diameter. The mean mass diameter – the size that splits the PSD into two equal halves – is given by

$$\overline{D} = \left(\frac{\rho_a}{c_x}\frac{q}{N_t}\right)^{1/3} \quad (3.5).$$

Figures

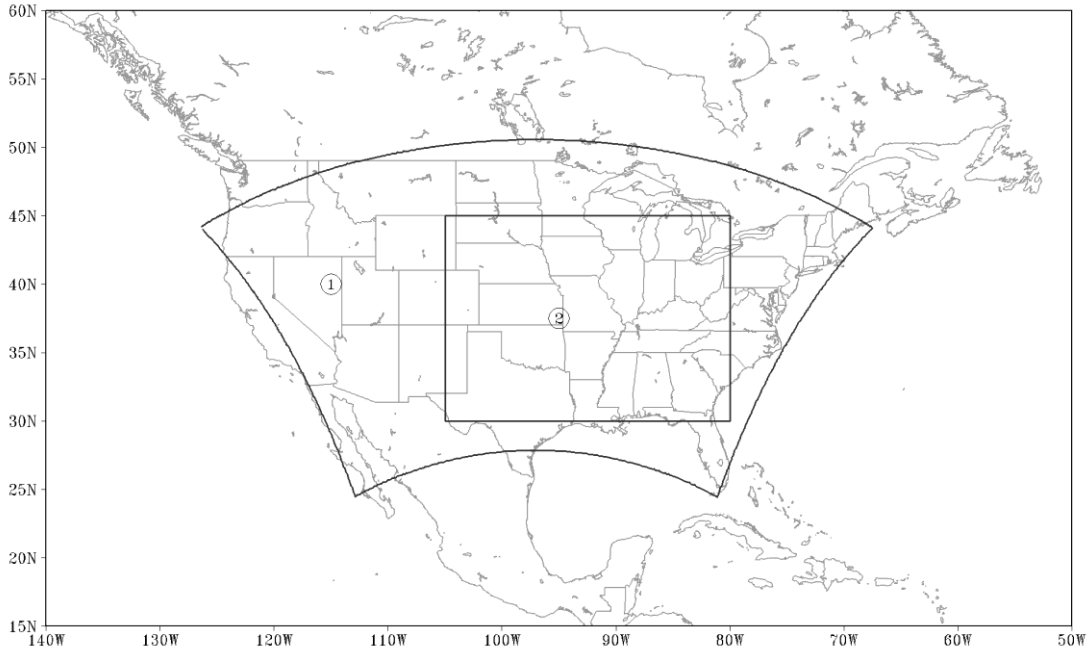

Fig. 3.1. Model domain (circled “1”) and verification domain (circled “2”) used in this study.

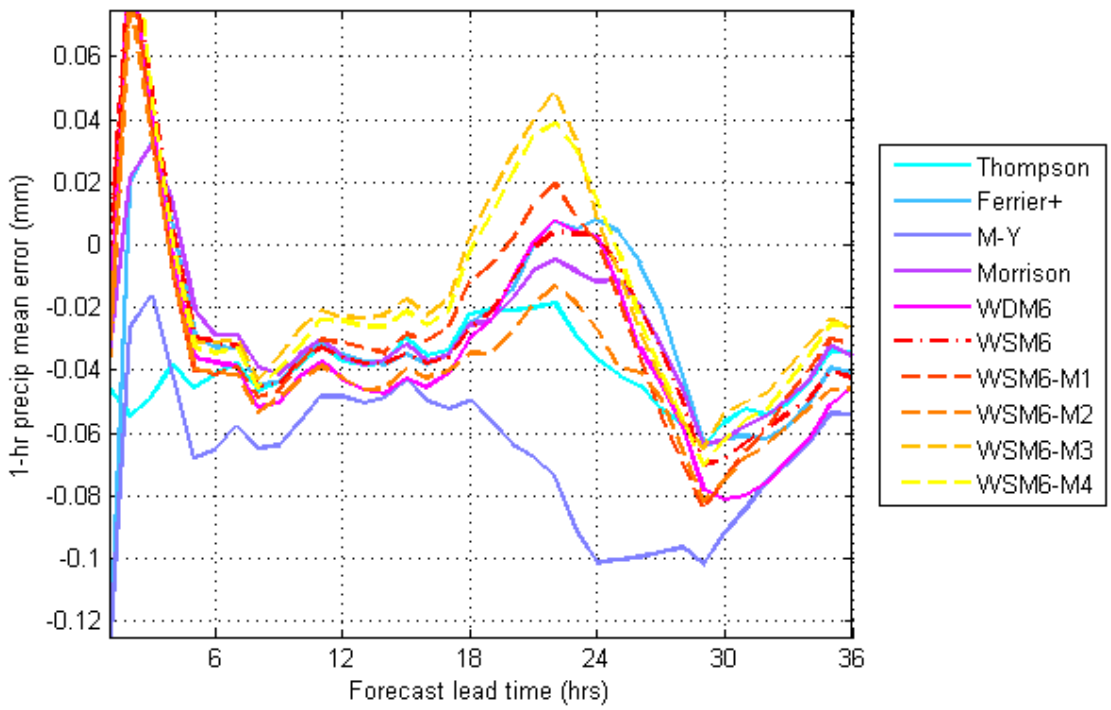

Fig. 3.2. Domain- and case-average mean error of 1-hr QPF for individual members.

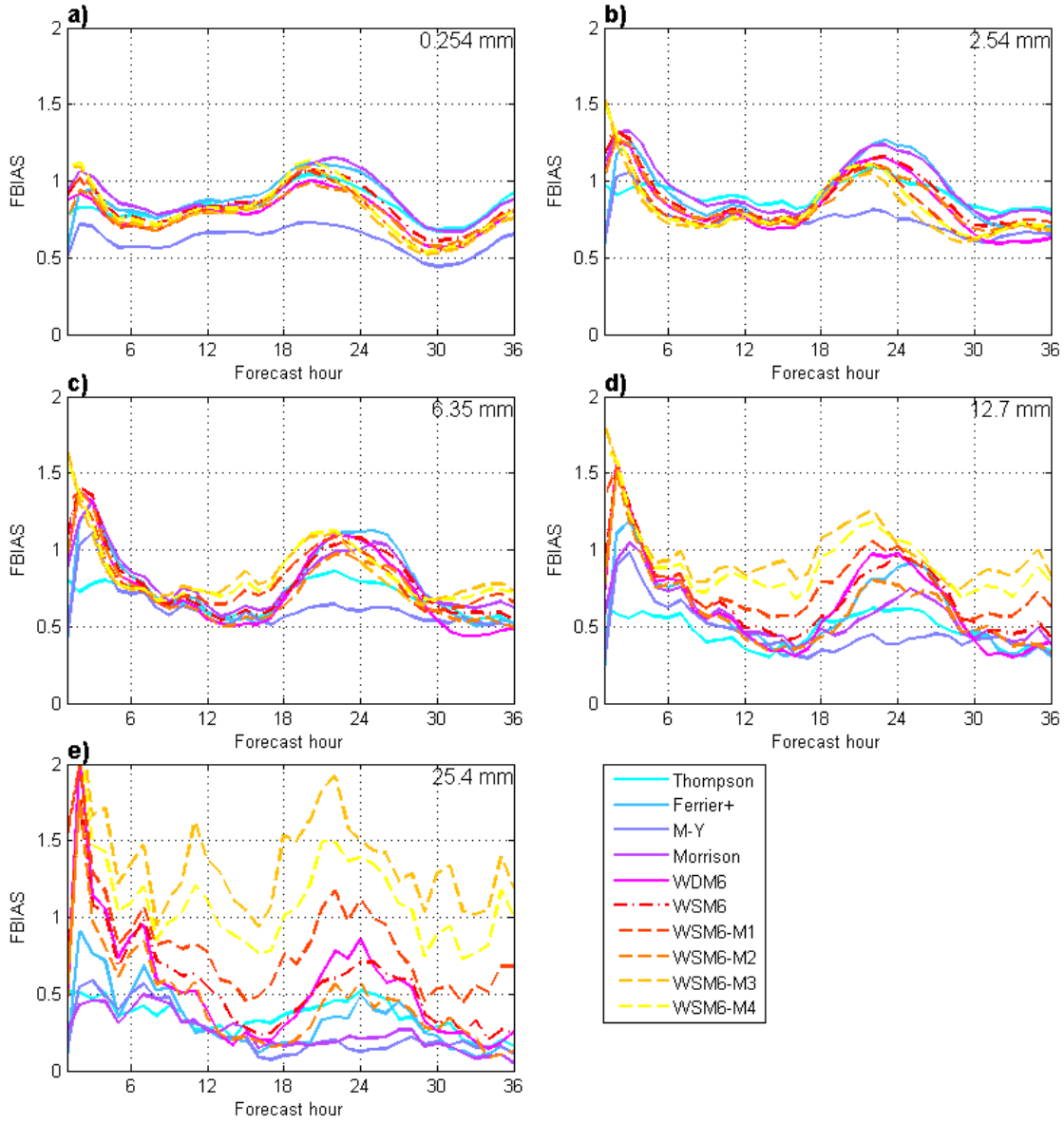

Fig. 3.3. Same as Fig. 3.2 except for frequency biases at the various thresholds indicated.

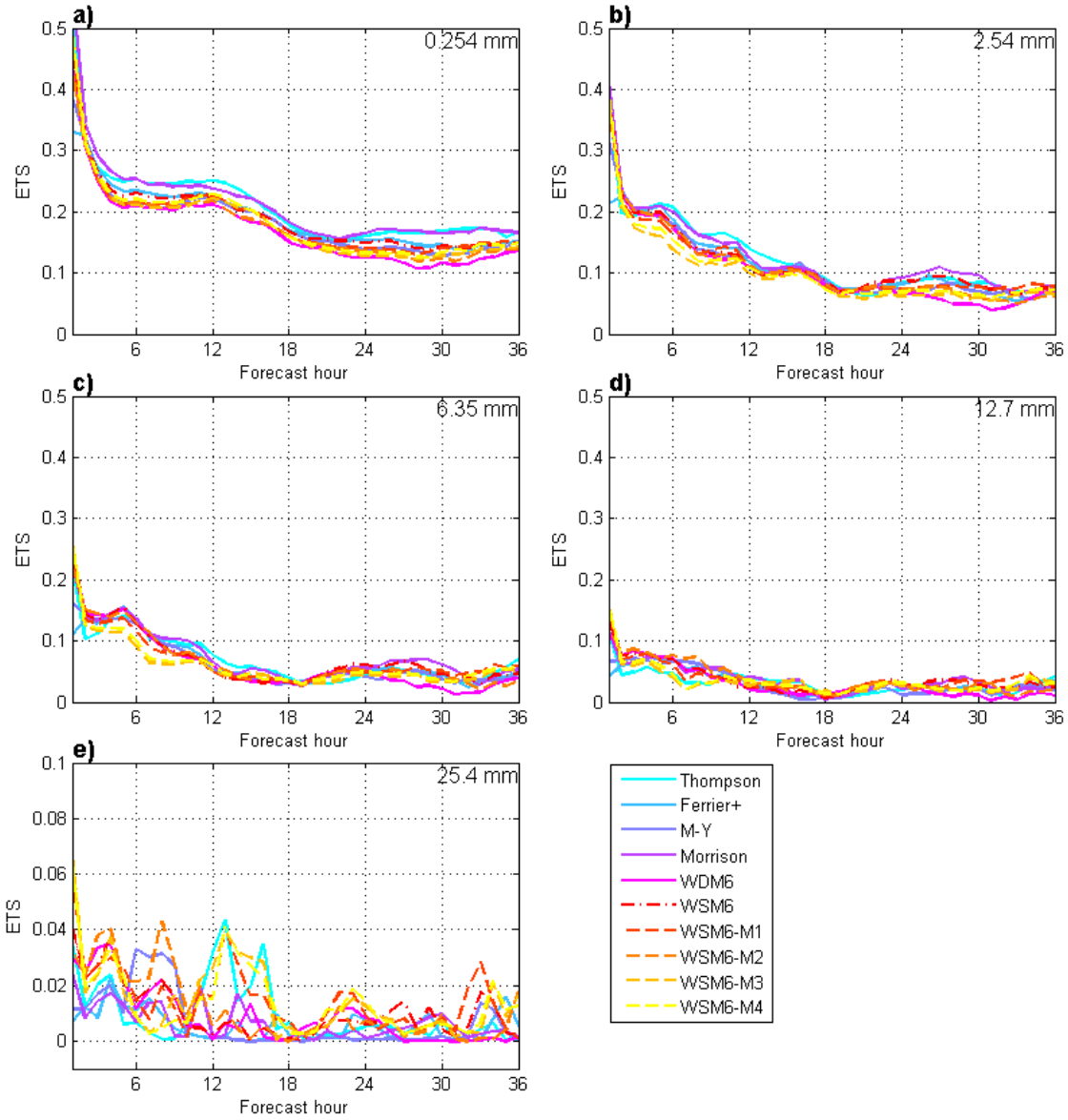

Fig. 3.4. As in Fig. 3.3 except for ETS.

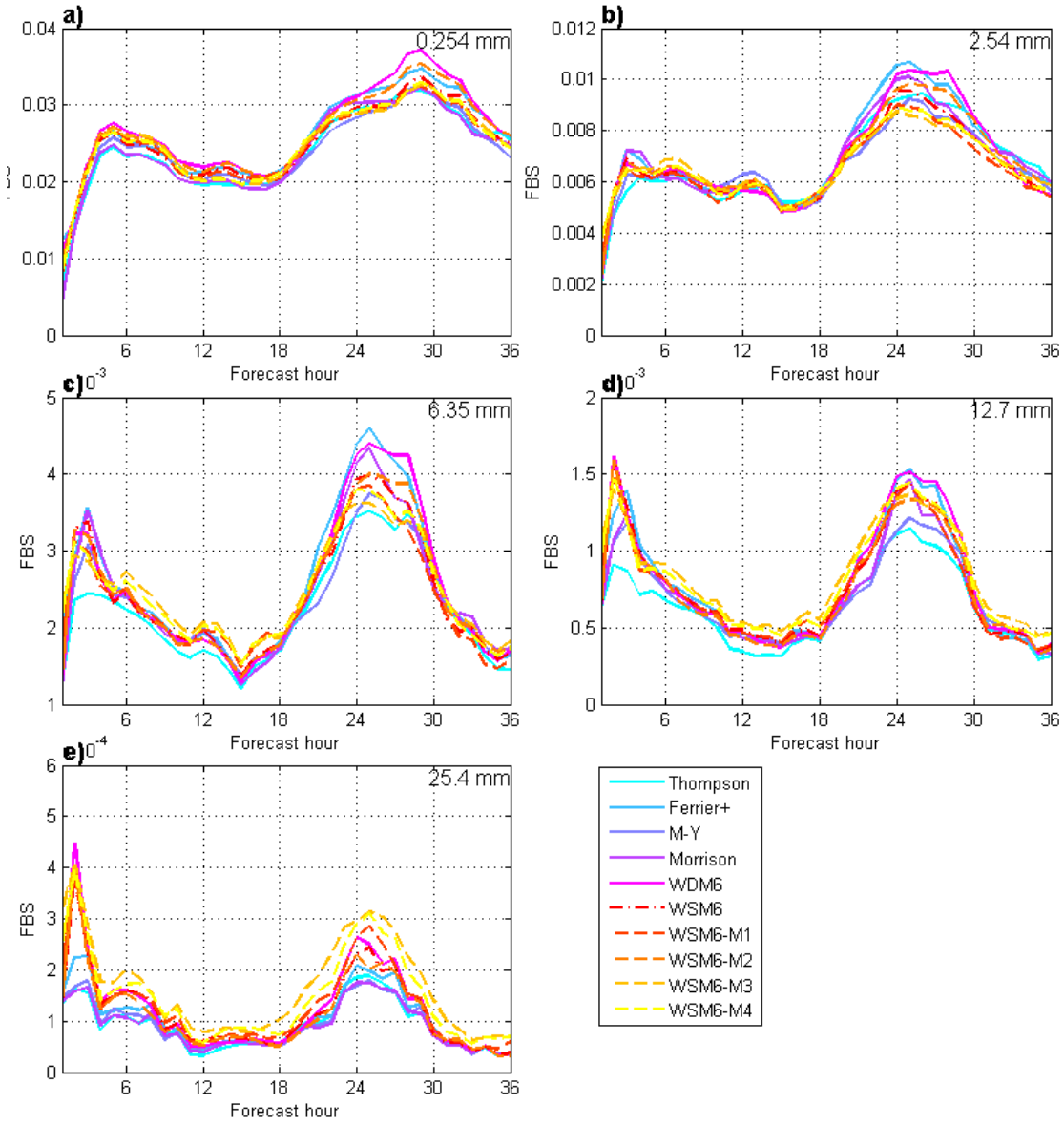

Fig. 3.5. As in Fig. 3.3 except for FBS.

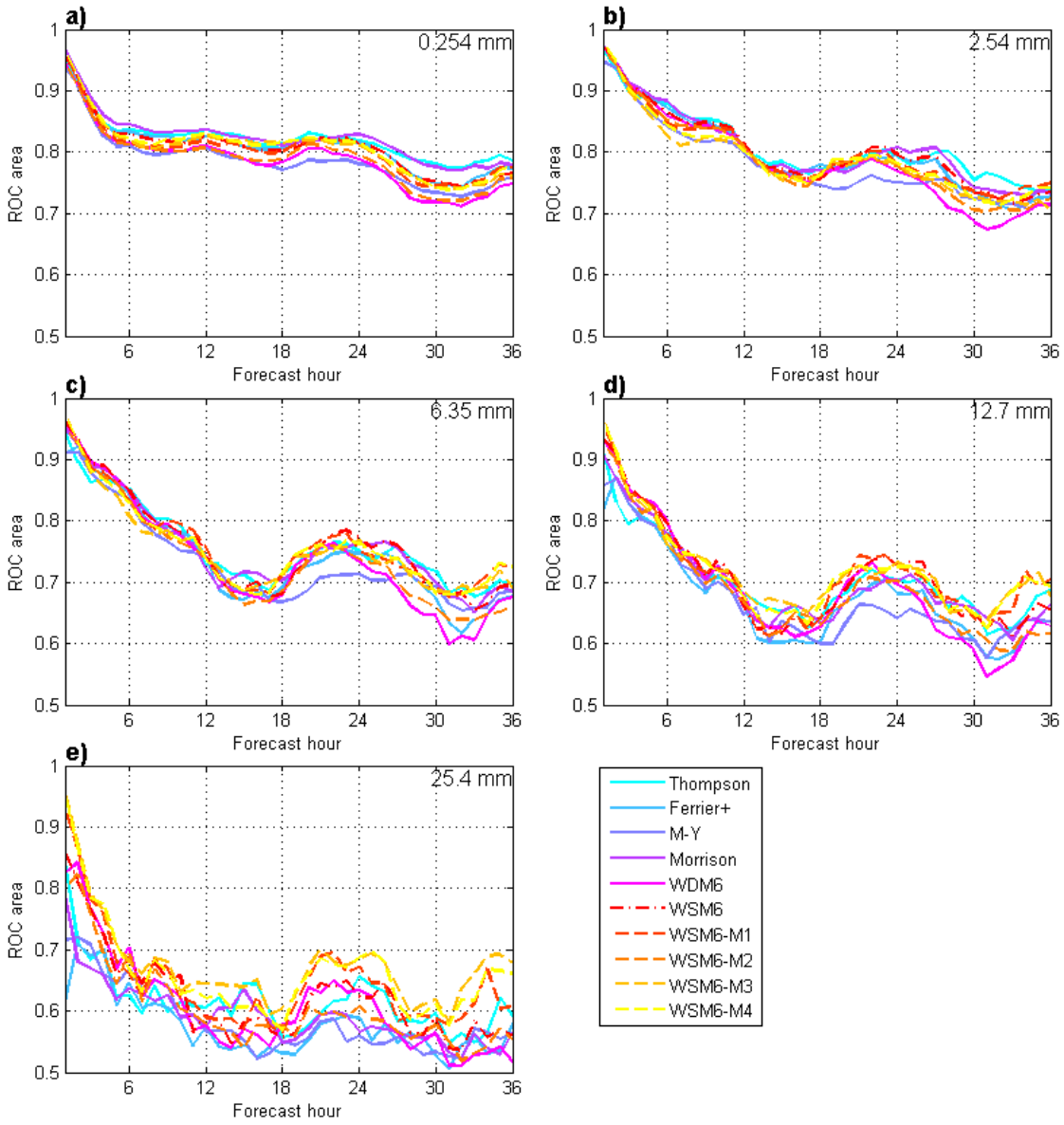

Fig. 3.6. As in Fig. 3.3 except for area under the ROC curve.

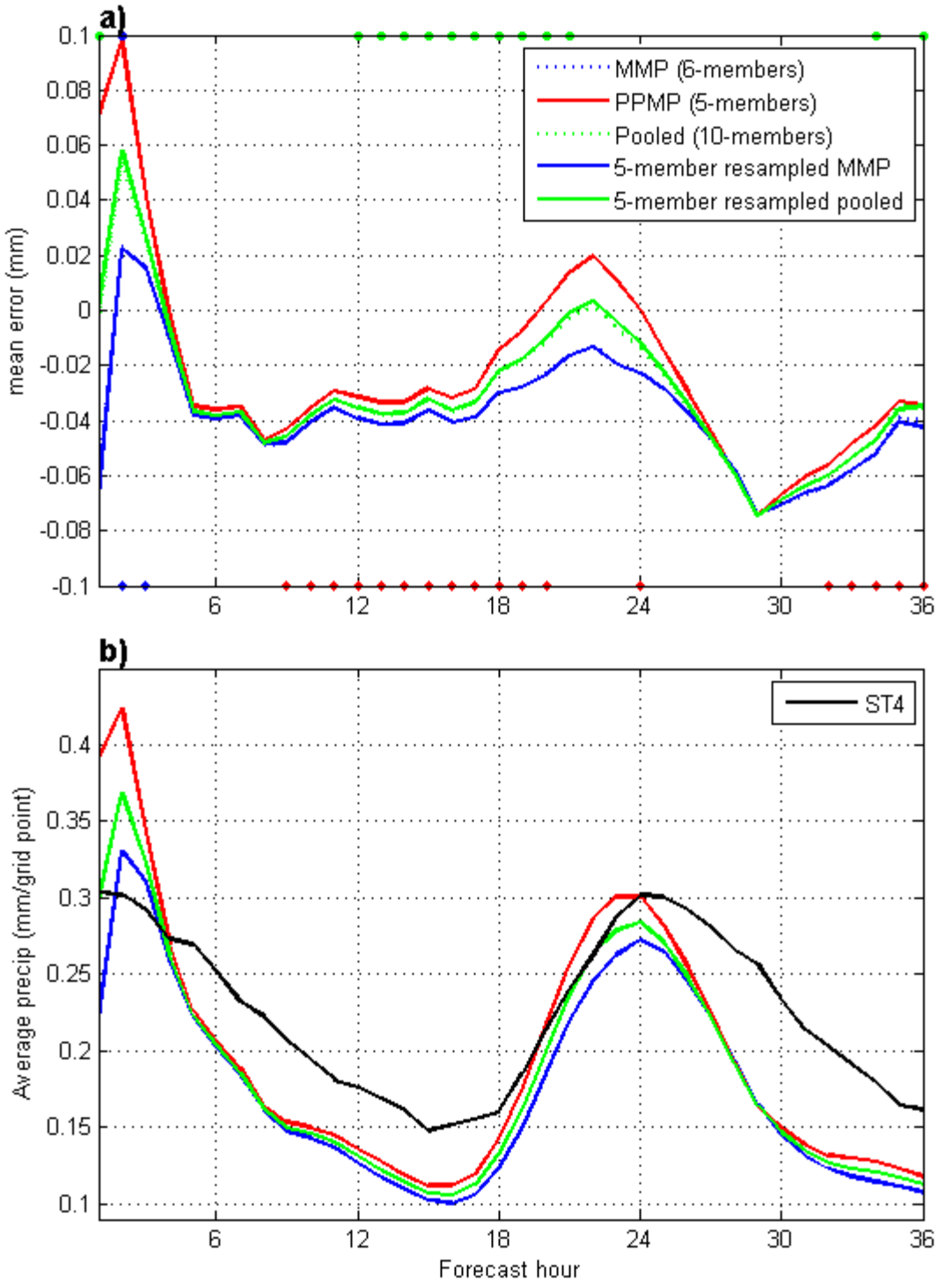

Fig. 3.7. (a) Domain- and case-average mean error of 1-hr ensemble mean QPF. Diamonds across the bottom indicate forecast hours at which the difference between the MMP and PPMP ensembles was significant, with the color indicating which ensemble was superior, whereas filled circles across the top indicate forecast hours at which the difference between the MMP and pooled ensembles was significant. (b) Domain-average 1-hr ensemble mean QPF and ST4 QPF (black line).

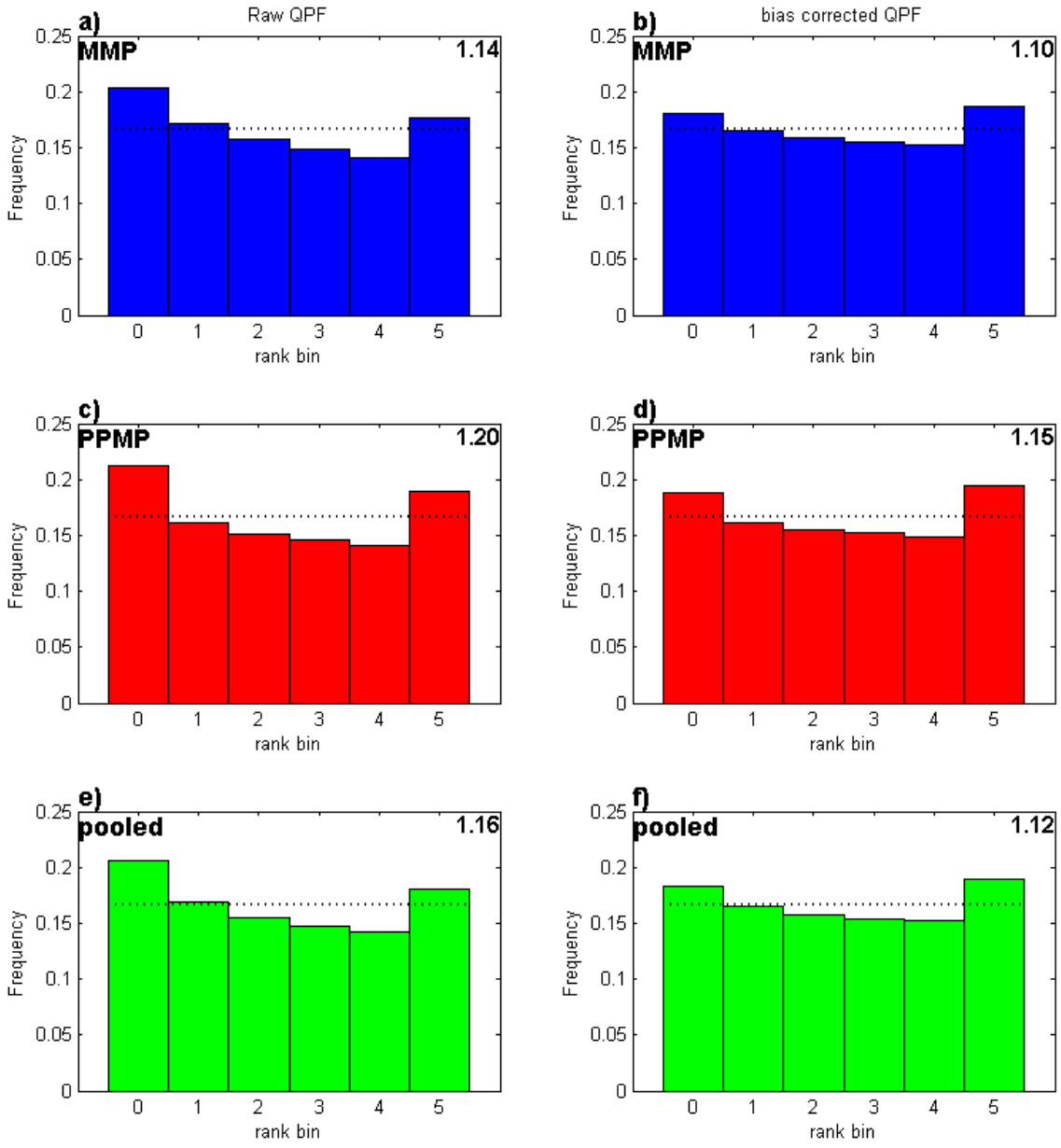

Fig. 3.8. Rank histograms for 1-hr QPF integrated over all forecast hours. (a) and (b) – MMP ensemble; (c) and (d) – PPMP ensemble; (e) and (f) – pooled ensemble. The parameter ext, defined as the ratio of the proportion of values in the extreme ranks to the proportion of values in the extreme ranks of a flat histogram, is displayed in the upper right of each panel. A rank histogram of a perfectly dispersive ensemble is indicated by the dotted black line in each panel.

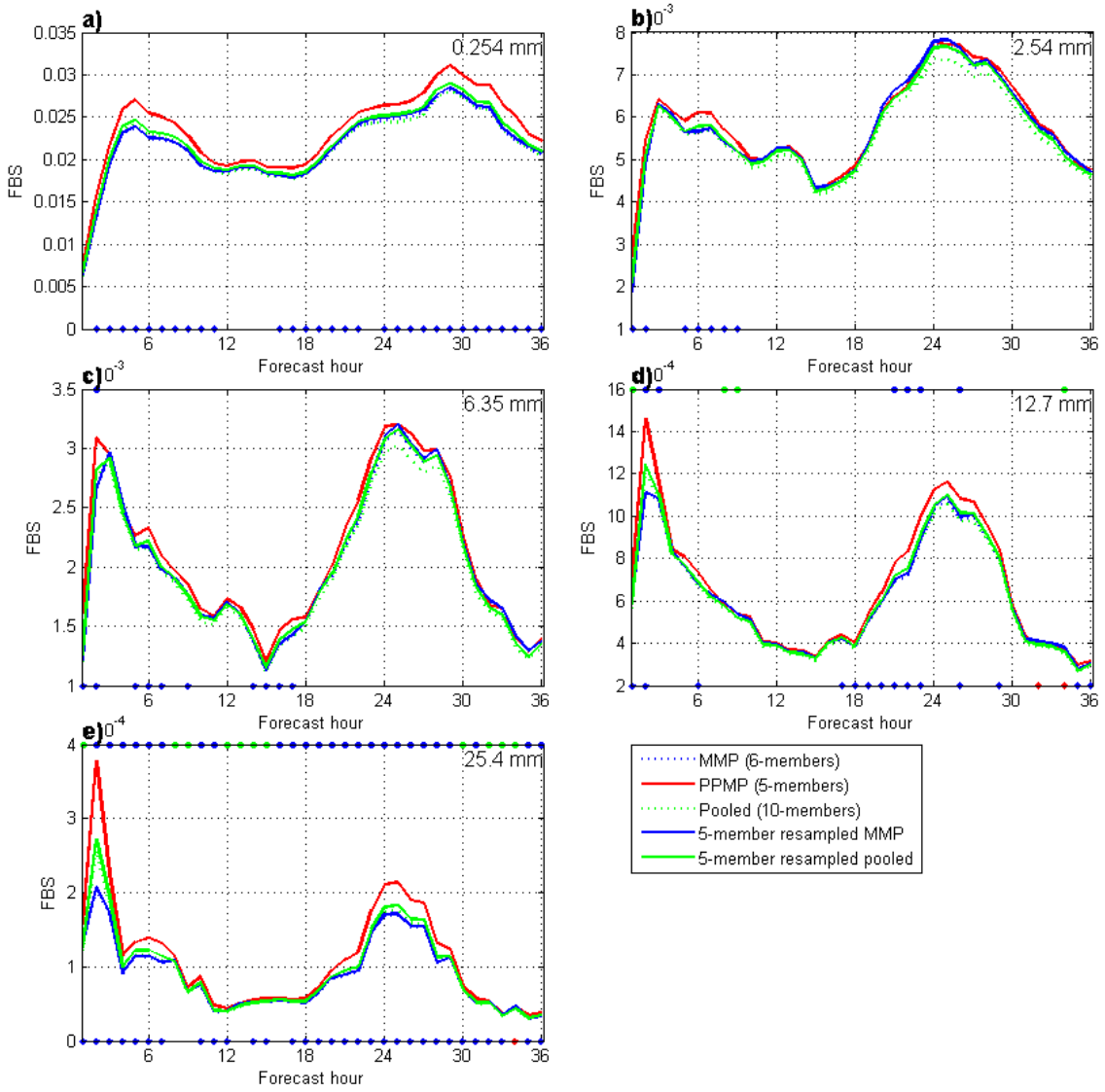

Fig. 3.9. FBSs for 1-hr ensemble mean QPF for the various thresholds indicated in the upper right of each panel. Significant differences are indicated as in Fig. 3.7.

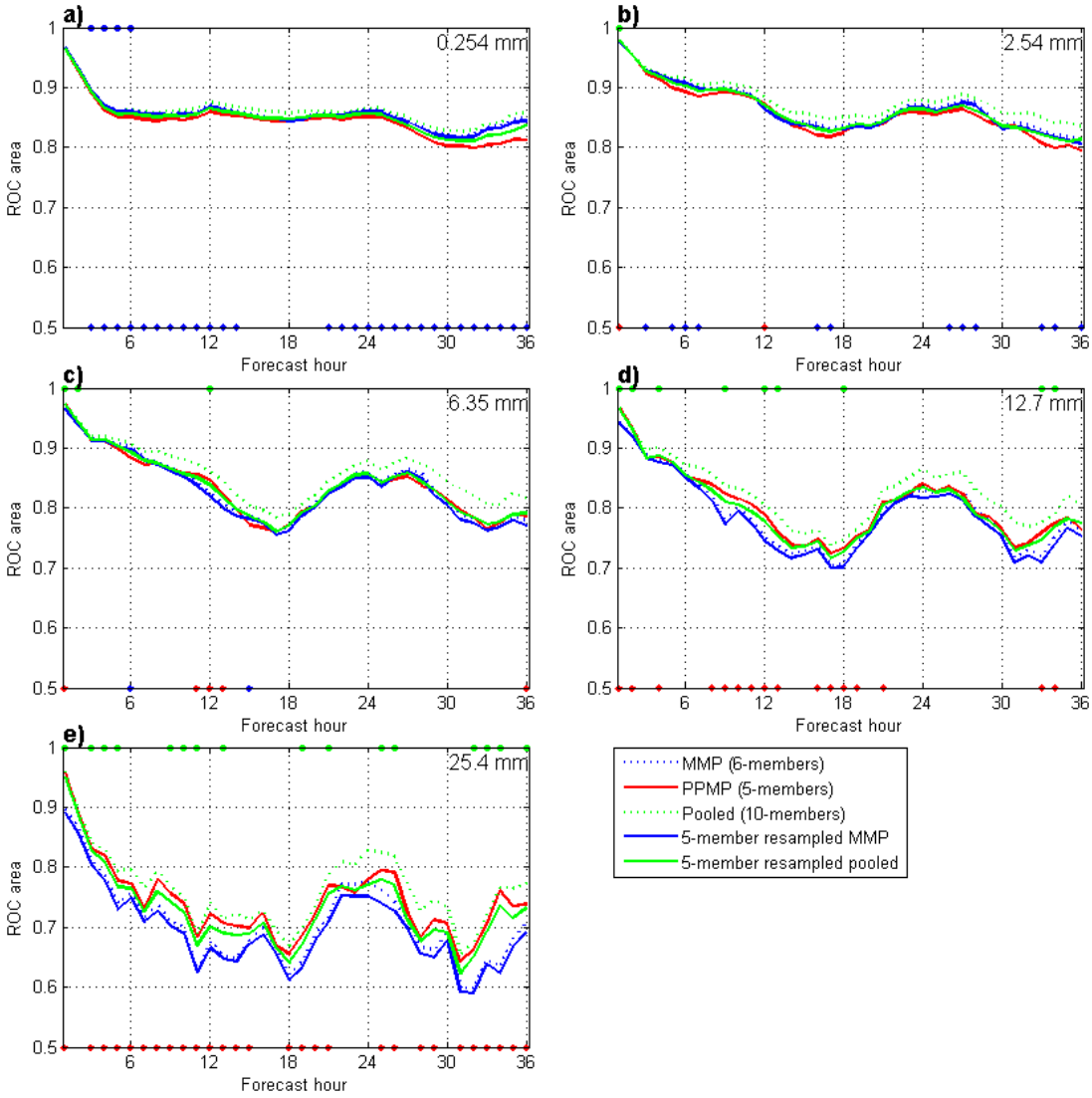

Fig. 3.10. Same as Fig. 3.9 except for area under the ROC curve.

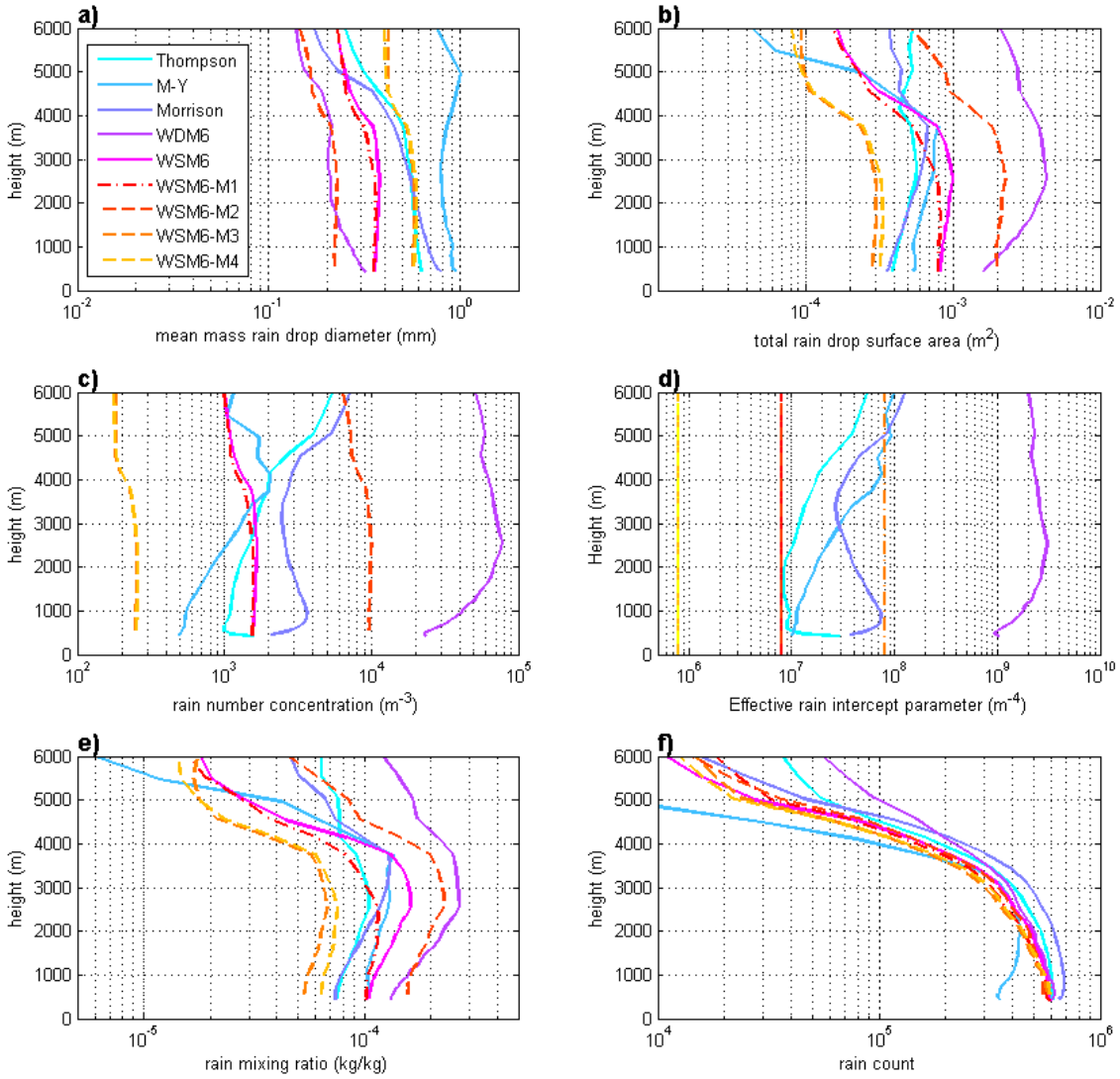

Fig. 3.11. Vertical profiles of area-average microphysics parameters: a) mean mass raindrop diameter (mm); b) raindrop surface area (m^2); c) rain number concentration (drops m^{-3}); d) Normalized rain intercept parameter (m^{-4}); e) rain mixing ratio (kg kg^{-1}); f) number of grid points at which the rain mixing ratio exceeded 10^{-6} kg kg^{-1}.

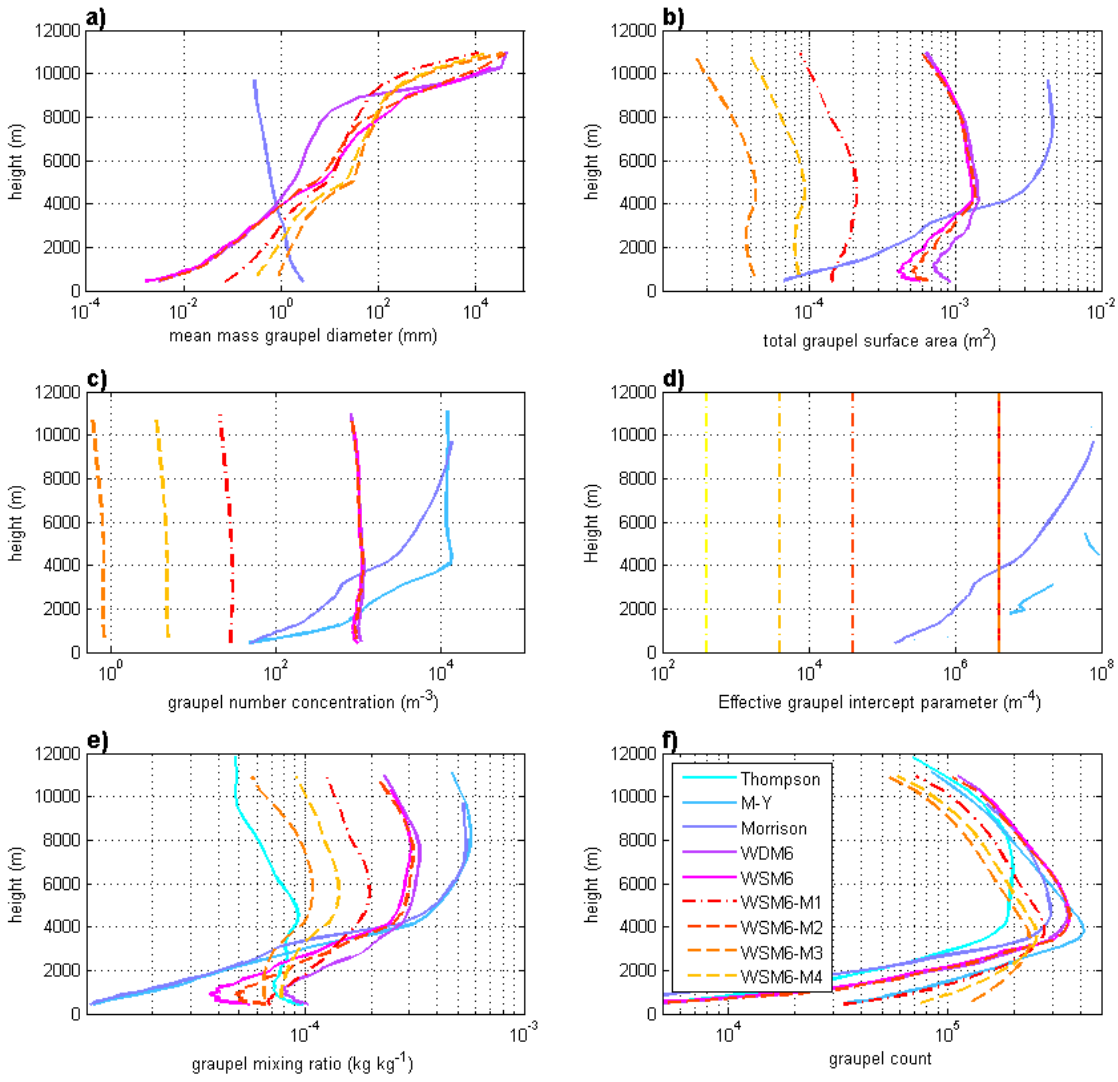

Fig. 3.12. Same as in Fig. 3.11 except for graupel.

[4]Chapter 4. Impact of a Stochastic Kinetic Energy Backscatter Scheme

4.1 Introduction

Current methods used to represent model error in a convection-allowing ensemble include multi-parameter, wherein fixed parameters within a physics parameterization scheme are varied (Hacker et al. 2011; Yussouf and Stensrud 2012; Chapter 3), mixed physics, where separate microphysics and boundary layer schemes are used (Johnson et al. 2011ab; Xue et al. 2011 and references therein; Chapter 3), and multi-model, where separate dynamic cores are used (Ebert 2001; Wandishin et al. 2001; Kong et al. 2009; Candille 2009; Johnson and Wang 2012; and Du et al. 2014). The Storm Prediction Center also uses a multi-model ensemble in which each member is a convection-allowing forecast provided by either the National Centers for Environmental Prediction or the National Severe Storms Laboratory (see www.spc.noaa.gov/exper/sseo). This chapter focuses on a different method of representing model error in an ensemble – stochastic perturbations using a stochastic kinetic energy backscatter scheme.

Research into the use of stochastic perturbations in ensemble forecasting is motivated by the results from prior research showing the benefit of using random perturbations. Buizza et al. (1999), for example, showed that merely including multiplicative random perturbations to the physical tendencies using simple spatiotemporal correlations was sufficient to increase ensemble spread and improve probabilistic precipitation forecasts. This method was based on the notion that the physical parameterizations handle subgrid-scale processes which are inherently random.

[4] This chapter is published as Duda, J. D., X. Wang, F. Kong, M. Xue, and J. Berner, 2016: Impact of a stochastic kinetic energy backscatter scheme on warm season convection-allowing ensemble forecasts. *Mon. Wea. Rev.*, **144**, 1887–1908.

The parameterizations take large-scale flow as input and thus are considered an ensemble average impact from subgrid-scale processes. The random perturbations therefore account for the variability in the subgrid-scale processes. Mason and Thomson (1992), on the other hand, used a SKEB scheme in a large eddy simulation to improve near-surface flow. Similarly, Shutts (2005) developed a cellular automaton stochastic backscatter scheme (CASBS) which inserted random perturbations into the model, but structured differently than the scheme in Buizza et al. (1999), and based on a different justification. The purpose of CASBS was to include a subgrid-scale process missing from global NWP models. The scheme injected kinetic energy (KE) into the model domain to counteract excessive energy dissipation coming from numerical diffusion and interpolation, mountain and gravity wave drag, and deep convection. Not only did CASBS correct the KE spectrum of the European Centre for Medium-Range Weather Forecasting (ECMWF) model, it also improved the spread and skill of 500 hPa geopotential height forecasts. Without CASBS the model failed to correctly simulate mesoscale circulations conforming to the observed $k^{-5/3}$ power law (Nastrom and Gage 1985). Berner et al. (2009) built off the work of Shutts (2005) and developed a spectral stochastic kinetic energy backscatter scheme (SSBS) which was implemented in the ECMWF ensemble prediction system in 2011 (ECMWF, 2012). The SSBS scheme was later modified for use in the Weather Research and Forecasting (WRF) model by Berner et al. (2011), who determined that the SKEB scheme gave superior ensemble mean forecasts of many fields compared to an ensemble using only physics variations. Their work was performed using a horizontal grid spacing of 45 km.

The use of SKEB schemes in operational EPSs has increased recently. Similar versions of the SSBS scheme have been introduced into the Canadian EPS (Charron et al. 2010), the Met Office Global and Regional EPS (Tennant et al. 2011), and the United States Air Force Weather Agency mesoscale ensemble (Hacker et al. 2011). The impact of SKEB has been overwhelmingly beneficial, including increased spread with a maintained or reduced root-mean square error (rmse) of the ensemble mean, and improved probabilistic forecasts of upper-level winds, temperatures, heights, and precipitation.

Prior research into the effectiveness of a SKEB scheme on probabilistic forecasts has been limited to global or otherwise coarse-grid scale EPSs. It remains to be determined how useful or valid such a scheme is for a convective-scale EPS. Such study is still limited. For example, Romine et al. (2014) compared ensemble forecasts at 3 km grid spacing using two stochastic perturbation methods, the SKEB scheme and the stochastically perturbed parametrization tendencies scheme (Buizza et al. 1999; Palmer et al. 2009). They found the SKEB scheme to provide a balance between increased ensemble spread and forecast bias change. The focus of this chapter is to compare ensemble forecasts using the SKEB scheme to a mixed-physics ensemble which is typically used in convection-allowing ensemble forecast system design. The following questions are investigated: is the stochastic error representation method (SKEB in this case) compatible with a mixed-physics approach in a convective-scale forecast? What is the impact of including SKEB perturbations on top of the typically used mixed-physics method in convection-allowing ensemble forecasts? A 4-km WRF ensemble including a

portion of the United States and featuring warm season cases is adopted to achieve this goal.

The rest of this chapter is organized as follows. The SKEB scheme is described briefly in section 4.2. The experiment design is described in section 4.3. Results follow in section 4.4. A summary and conclusions follows in section 4.5.

4.2 The SKEB scheme

4.2.1 Mathematical formulation

Stochastic parameterizations are developed to represent model error and can generate ensemble spread by perturbing the forecast trajectory. For this purpose, a SKEB scheme is employed that adds stochastic, small-amplitude perturbations to the rotational component of horizontal wind and potential temperature tendency equations at each time step. The scheme is briefly outlined here; full details can be found in Berner et al. (2011).

Let ψ(x,y,t) be a 2D-streamfunction-forcing pattern expressed in 2D-Fourier space:

$$\psi \ \ (x,y,t) = \sum_{l=-L/2}^{L/2} \sum_{k=-K/2}^{K/2} \psi_{k,l}(t) e^{2\pi i(\frac{kx}{X}+\frac{ly}{Y})} \qquad (4.1).$$

Here, $\psi_{k,l}$ denotes the spectral coefficient of the perturbation field with *k* and *l* the (K-1) and (L-1) wavenumber components in the zonal *x*- and meridional *y*-direction in physical space and *t* denotes time. The representation in spectral space allows for control of the spatial length scales of the perturbations as a function of wavenumber. Subsequently, the rotational wind components are obtained by differentiation, the pattern transformed back to gridpoint space, and the perturbations added to the

momentum tendency equations. The perturbations to the potential temperature tendencies are generated analogously. The WRF implementation allows for a 2D or 3D perturbation pattern to be generated. Here, we follow Berner et al. (2011) and use the same 2D pattern to perturb the tendencies in all vertical levels.

To introduce spatial and temporal correlations, each spectral coefficient $\psi_{k,l}$ is evolved as a first-order autoregressive process:

$$\psi_{k,l}(t + \Delta t) = \alpha\psi_{k,l}(t) + \sqrt{\alpha - 1}g_{k,l}\varepsilon(t) \quad (4.2),$$

where α is the linear autoregressive parameter determining the temporal decorrelation time, $g_{k,l}$ is the wavenumber-dependent noise amplitude, and ε a Gaussian white-noise process with mean 0.0 and variance 1.0. The noise amplitude $g_{k,l}$ determines the variance spectrum of the forcing and is given by the power law, $g_{k,l} = bn^p$, where n is the wavenumber, p is an assumed constant slope, and b is the amplitude defined as in eqn. (4) of Berner et al. (2009), chosen so that a fixed amount of KE is injected into the flow each time the forcing is applied. The autoregressive parameter determines the decorrelation time τ of the pattern, $\tau = \Delta t/(1 - \alpha)$, where Δt is the model time step. In principle, each spectral coefficient can be associated with a different decorrelation time, but for practical reasons, the same decorrelation time is chosen for all spectral coefficients. The SKEB scheme was originally motivated by the notion that upscale and downscale cascading energy resulted in net forcing for the resolved flow from unresolved scales (Shutts 2005) and the perturbation amplitude was proportional to the instantaneous dissipation rate. Here, the simplifications in Berner et al. (2011), where it is assumed the dissipation rate is constant in space and time, are used. Since the forcing

is no longer state-dependent the perturbations can be considered as additive noise with prescribed spatial and temporal correlation.

4.2.2 Tuning the scheme

The autoregressive parameter α (related to the decorrelation time of the forcing field), the slope of the power spectrum for the perturbations, and the amplitude of the perturbations are adjustable parameters which can be tuned for a specific application. Sensitivity tests were conducted to determine if the default settings of Berner et al. (2011) – obtained for forecasts with a horizontal resolution of 45 km – are also optimal for the much higher horizontal resolution used here. The cases selected, 13 April 2012, 14 June 2012, and 23 June 2013 (none of which are part of the experiment period; section 4.3), featured warm season precipitation episodes exhibiting a wide variety of convective modes and magnitude of large-scale dynamic forcing, thus allowing for a general tuning of the scheme over a range of scenarios. This approach also avoids issues related to in-sample tuning to generalize the results.

Berner et al. (2009) informed the slope of the forcing spectrum, *p*, using coarse-graining cloud-resolving model output (Shutts and Palmer 2007). Rather than following this strategy and coarse-grain output from large-eddy simulations, the slope of the power spectrum herein is determined empirically. For this purpose the default value of the spectral slope was perturbed in the positive and negative directions by 20%, 40%, and 80%. The decorrelation time was perturbed similarly. Additionally, the sensitivity to the amplitude of the perturbations was tested by varying the backscatter dissipation rates for wind and temperature, respectively, by several orders of magnitude. We did not test every combination of parameter values. Instead we evaluated an arbitrary set

that focused on changing only one parameter at a time, although we did test a few sets of coupled parameter perturbations. This method was not intended to be comprehensive, and it is possible that other combinations of values may yield yet better ensemble statistics and may represent a better tuning of the scheme.

Sensitivity tests using larger wind and temperature amplitudes for the perturbations generated drastically increased ensemble spread relative to the default values. However, as these perturbation amplitudes increased, the member forecasts looked increasingly different from verifying precipitation analyses (not shown) and hence were subjectively judged to be degraded. Therefore we chose the default values of the wind and temperature perturbation amplitudes. An example of the structure of the perturbation tendency fields for u-wind and temperature is shown in Fig. 4.1.

While ensemble performance was not obviously superior for any single combination of decorrelation time and spectral slope values, examination of ensemble spread helped to indicate a tendency for a 40% reduction in decorrelation time coupled with a 40% increase in the spectral slope to produce the best agreement between spread and rmse of the ensemble mean for several synoptic scale fields such as temperature, geopotential height, wind, and moisture (Figs. 4.2 and 4.3). While improvement of probabilistic quantitative precipitation forecasts (PQPF) at the convective scale is emphasized, precipitation forecasts were not strongly sensitive to the choice of decorrelation time or spectral slope (not shown). Therefore, we used 6480 s for the decorrelation time and 2.567 for the spectral slope for both wind and temperature perturbations. This combination did not always result in the best precipitation forecasts according to Brier score of 1-hr accumulated precipitation at various thresholds, but it

did not always result in the poorest precipitation forecast either (not shown). Independently, Ha et al. (2015) performed limited sensitivity tests of SKEB scheme parameters in cycled simulations with WRF at a resolution of 15 km and determined that the default parameters performed well. This together with our findings points to the generality of the parameter settings in WRF, at least for application to forecasts in the mid-latitudes.

Only one set of tunable parameter values of the SKEB scheme was used for all experiments involving the use of SKEB. In other words, the same values were used in the SKEB scheme for each combination of physics options used in the ensembles tested. It is likely that the optimal parameters for the SKEB experiment may not be the same as those actually used given the potential dependence of SKEB parameters to the choice of physics. Therefore the comparison between the SKMP and MP ensembles may not maximize the value of adding SKEB on top of mixed physics. However, the conclusion of added value from including SKEB on top of mixed physics should not change qualitatively.

4.2.3 Impact on WRF KE spectra

The original motivation for using a SKEB scheme was to counteract excessive dissipation in the ECMWF model, in part caused by the use of a semi-Lagrangian time stepping scheme (Shutts, 2005). The ECMWF ensemble is a global ensemble and is typically not run at non-hydrostatic, convection-allowing resolutions. At lower horizontal resolution the KE spectrum of NWP models typically does not capture the observed transition from the k^{-3} spectrum characterizing the synoptic scale to the shallower $k^{-5/3}$ spectrum in the mesoscale (Nastrom and Gage, 1985). The ECMWF

implementations of SSBS were able to capture this transition by simulating under-represented mesoscale variability (Shutts, 2005; Berner et al., 2009). It is an area of active research if the existence of a $k^{-5/3}$ spectrum in a NWP model is necessary for reliable ensemble predictions with forecast target times of few days, since it is associated with faster error growth (e.g., Lorenz, 1969; Rotunno and Snyder, 2007).

Following Skamarock (2004) we computed KE spectra from WRF simulations at horizontal resolutions of 1 km and 4 km (Fig. 4.4). At both resolutions the KE spectra are characterized by a $k^{-5/3}$ slope which drops off above wavenumbers of 0.02 m^{-1} and 0.001 m^{-1} for the 4 km and 1 km resolutions, respectively. The slope in the simulations at 1 km continues through the meso-γ scale, which indicates that the drop in the tail of the spectrum in the 4 km simulation is due to numerical truncation error rather than a characteristic of the circulation pattern. For the tuning parameters chosen the simulations using the SKEB scheme do not show any appreciable difference in KE structure from simulations that do not use the SKEB scheme which confirms that the scheme does not introduce artificial energy. Arguably, it would have been desirable to tune the scheme so that the KE spectra at coarser resolution resemble more those at 1 km, but given that WRF has already a $k^{-5/3}$ spectrum at the chosen resolution, the benefit might be small.

4.3 Experiment setup

To compare two methods of accounting for model error, SKEB and mixed-physics, and to evaluate the impact of adding SKEB on top of the multiple physics approach, three ensembles were constructed. Two ensembles contained mixed physics (microphysics, boundary layer, and land-surface model) that only differed in use of the

SKEB scheme. The ensemble that did not use the SKEB scheme is hereafter called the MP ensemble, whereas the ensemble that did is hereafter called the SKMP ensemble. The MP ensemble resembles the common mixed-physics design of the storm scale ensemble forecast system produced by CAPS at the University of Oklahoma for the SPC/NSSL HWT spring forecasting experiment (see, for example, Xue et al. 2011). Additionally, to test whether stochastic error representation alone is sufficient for use in an ensemble compared to mixed-physics error representation alone, a third ensemble, the SK ensemble, was constructed that contained no physics diversity; different random number seeds were used to supply diversity in the SKEB scheme. The configuration of each ensemble is shown in Table 4.1. Seven members comprised each ensemble. The choice of seven members represents a balance between computational resources, availability of various physical parameterization schemes, and adequate representation of the forecast probability distribution. This size is reasonably adequate to produce precipitation forecasts at a spatial scale of 50 km (section 4.4.3) that are statistically indistinguishable from a larger ensemble that would better populate a probability distribution (Clark et al. 2011). Since the focus of the study is on representation of model error, neither initial condition nor lateral boundary condition perturbations were used.

Table 4.1. Configuration of the ensembles. The last column indicates the integer value of the random number seed used to generate pseudo-random numbers in the FORTRAN code that the WRF is built on. The SK ensemble used the same physics configuration as member 1, but the random number seed varied among the members as indicated.

Member	**Microphysics**	**PBL**	**LSM**	**seed**
m1 (control)	Morrison	YSU	Noah	2
m2	Ferrier	MYNN2.5	RUC	3
m3	WSM6	YSU	Noah	4
m4	Thompson	MYJ	Noah	5
m5	MY	ACM2/QNSE*	RUC	6
m6	WDM6	MYJ	RUC	7
m7	NSSL	QNSE	Noah	8

*The PBL scheme for member 5 was switched from ACM2 to QNSE starting with the case initialized at 0000 UTC 8 May due to difficulties resulting from the interaction between the ACM2 PBL scheme and the other physics options in that member.

The Advanced Research WRF (ARW) dynamics core (Skamarock et al. 2008) of the WRF model, version 3.4.1, was used to conduct the simulations. The model domain encompassed a large portion of the central and eastern U.S. (Fig. 4.5) where deep convection is climatologically favored during the late spring and early summer, the period during which the experiment was conducted. We tested 31 cases spanning May 2013. There were a number of active severe weather days in the model domain during this period, so the results of this study should be representative of the overall ability of the SKEB scheme in a convective-scale EPS for warm season events. Each case was initialized at 0000 UTC and integrated for 30 hours. The grid spacing for all members was 4 km with 20 s model time step. NAM model analyses valid at 0000 UTC were used as the initial conditions, and NAM model forecasts from 0000 UTC for each case were used as the lateral boundary conditions.

Verification was performed on a number of fields, including temperature, wind, height, and various moisture variables at multiple atmospheric pressure levels, as well as 1-hr accumulated precipitation. Rapid Refresh (RAP; Brown et al. 2011; Weygandt

et al. 2011; rapidrefresh.noaa.gov) analyses were used as verifying data for upper-air fields. METAR and mesonet observations obtained from the Meteorological Assimilation Data Ingest System (MADIS; https://madis.noaa.gov/) were used to verify 2-m temperature and dewpoint and 10-m winds. Only observations passing quality control checks were used for verification. A motivation behind verifying surface fields using observations rather than a RAP analyses is representativeness of surface values. The fine-scale detail in the WRF can be better verified using observations than by using a 13-km grid spacing model analysis (RAP; which does not resolve features smaller than 26-km in wavelength). Even though observing stations are spread farther apart than the RAP model grid points, they are still capable of capturing small scale features such as individual thunderstorms at certain locations. The signal from a single sample would certainly be wiped out when performing a gridded analysis. However, the WRF model can capture features as small as 8 km in size. For this reason, it was not required for observations to pass spatial continuity quality control checks to be included in the verification. This freedom increases the chances of the model being rewarded for forecasting a minimally-resolved thunderstorm and associated cold pool in the right place at the right time such that the corresponding observation also sampled the storm. Verifying precipitation data were provided by the National Mosaic and Multi-Sensor Quantitative Precipitation Estimation (QPE) project (NMQ) produced by the National Severe Storms Laboratory (Zhang et al. 2011). The NMQ QPEs (native grid at 0.01° resolution) were regridded to the verification domain using bilinear interpolation. Verifications were performed using one-hour accumulated precipitation. The NMQ QPEs have been used in earlier studies to verify storm-scale precipitation forecasts

(e.g., Johnson and Wang 2012; Johnson et al. 2013; Chapter 3). PQPFs were constructed without calibration or bias correction as the number of members in which a threshold value was exceeded.

4.4 Results

4.4.1 Skill of physics packages

The WRF-ARW offers a large set of different physics packages (mainly microphysics and PBL), and many of those schemes were used in the MP and SKMP ensembles. The skill of individual packages was examined first to facilitate the comparison between the SK, MP, and SKMP ensembles.

The rmse of the MP ensemble members for a large number of fields is shown in Fig. 4.6. Especially for upper tropospheric winds and heights (Figs. 4.6a-d), the scores were clustered tightly; only member m5 stood out as a poorer physics package for these fields. For lower tropospheric fields (Figs. 4.6e-l), there was more diversity in the rmse among the members. After about forecast hour 12 or so, two members, m1 and m3, which both used the YSU PBL scheme (Hong et al. 2006b) and Noah land-surface model (Ek et al. 2003), tended to perform better than the other members. This was especially apparent in near-surface fields such as wind, temperature, and dewpoint (Figs. 4.6j-l). In this regime (warm season over mid-latitude plains) the YSU scheme appears to be a better PBL scheme as verified here. The difference between members m1 and m3 was the choice of microphysics scheme; member m1, on which the SK ensemble is based, used the Morrison microphysics scheme (Morrison et al. 2009), noted as one of the better multi-moment microphysics schemes in Chapter 3, whereas member m3 used the simpler WRF single moment 6-class microphysics scheme (Hong

and Lim 2006a). These members were not as skillful for 1-hr precipitation at forecast hours 6-21, and member m3 was less skillful than member m1 after forecast hour 14. However, member m1 was somewhat less biased than most other members in some fields, although it did not always have the smallest bias (not shown). Members m5 and m6 were frequently among those with the highest rmse, especially after forecast hour 18 (Figs. 4.6e-l). Finally, there was a notable clustering by land-surface model in the 2-m temperature and dewpoint fields (Figs. 4.6k,l). Such stark contrast in rmse suggests the forecast distributions for these fields was likely bimodal in some cases. Future work will investigate ways to better account for land-surface model uncertainties to better populate the forecast distribution and to determine if such a bimodal distribution reflects the underlying truth error distribution or is an artifact associated with LSMs scheme behavior.

Overall, the bias and rmse characteristics for the individual members indicates that the SK ensemble was based on a relatively skillful set of physics parameterizations and thus could be expected to provide forecasts competitive with those from the MP and SKMP ensembles, which use mixed physics.

4.4.2 Ensemble spread-error agreement and dispersion

4.4.2.1 Against RAP analyses

4.4.2.1.1 Spread, RMSE, and rank histograms

We first examine the spread-error relationship of the ensembles for several fields. The ensemble spread, averaged over the 31 cases and the verification domain, is shown for several fields in Fig. 4.7. The addition of the SKEB scheme added a large

amount of spread to the upper-tropospheric fields in the SKMP ensemble. For fields such as hgt500 and u500__ (see Table 4.2 for field abbreviations), the amount of added spread exceeded 100% at forecast hour 36. Fig. 4.4b shows that, for the u-wind component, spread was added at nearly all but the finest scales, with relatively more diversity added at the largest scales where the perturbation amplitudes were also the largest. There was also increased spread in the SKMP ensemble at fields in the lower troposphere, although not as much as was added above. There was even an increase in spread in moisture fields (pwat__ and accppt) despite those fields not being directly perturbed. The spread in the SKMP ensemble was higher than that in the MP ensemble at all forecast hours after forecast hour 1, and the difference in spread between the MP and SKMP ensembles generally increased with time throughout the 30-hour forecast, but especially after forecast hour 6 or so, after which time model spin-up was complete. Additionally, the spread in the SK ensemble was also much larger than that in the MP ensemble for upper-tropospheric fields. At middle and lower tropospheric levels, the SK ensemble had lower spread than the MP ensemble until about forecast hour 5, after which the order was reversed.

Table 4.2. Abbreviations for field names used for verification.

Field name	**Description (units)**
hgt500	500 hPa geopotential height (m)
v850__	850 hPa v-wind component (m s^{-1})
v500__	500 hPa v-wind component (m s^{-1})
u500__	500 hPa u-wind component (m s^{-1})
u250__	250 hPa u-wind component (m s^{-1})
sph850	850 hPa specific humidity (g kg^{-1})
pwat__	precipitable water (mm)
tmp850	850 hPa temperature (K)
tmp500	500 hPa temperature (K)
accppt	1-hr accumulated precipitation (mm)

Given the general under-dispersive nature of many ensembles (e.g., Chapter 3), increased spread is an attractive result. However, increased spread does not necessarily mean the forecast error was sampled more appropriately. A large spread can result from incorrectly sampling forecast errors, which could lead to degradation of the ensemble mean forecast or member forecasts that are extremely different from one another and therefore lead to a degraded probabilistic forecast. The forecast quality is first examined via the rmse of the ensemble mean, also shown in Fig. 4.7. For most fields there was not a numerically large difference in the rmse among the three ensembles. However, for several fields, the rmse of the SKMP ensemble was lower than that of the MP ensemble at a large number of forecast hours (e.g., tmp500, v850__, sph850, pwat__, tmp850, and accppt), and that difference is statistically significant using a t-test at $\alpha = 0.05$. The fields showing the biggest decrease in rmse were concentrated in the lower troposphere, suggesting the SKEB scheme is quite effective in perturbing fields at lower levels when also coupled with physics perturbations, an observation also made by Hacker et al. (2011) and Berner et al. (2011) for their studies at 45 km grid spacing. Since the magnitude of the wind and temperature tendencies for SKEB perturbations is not

dependent on height, the relative magnitude of the perturbations is larger in the lower troposphere, which may have played a role in the greater improvement there. The decreased rmse in moisture fields such as pwat__ and sph850 is particularly interesting since they are not directly perturbed by the SKEB scheme. This decrease in rmse could come from improvement in precipitation and thunderstorm processes, which are directly impacted through the lower-tropospheric wind and temperature perturbations. There are a few fields for which the rmse of the ensemble mean of the SKMP ensemble was higher compared to the MP ensemble. However, the degradations were limited to winds and heights at 500 hPa and above. The increase in rmse of the hgt500 field in particular could be related to changes in the number of thunderstorms present in the forecasts caused by lower tropospheric wind and temperature perturbations. An individual thunderstorm can strongly perturb the height field through non-hydrostatic vertical accelerations.

The rmse of the ensemble mean in the SK ensemble was commonly higher than that in the MP ensemble at early lead times in most fields (black dots across the top of each panel in Fig. 4.7). However, in the middle of the forecast period, and for lower tropospheric fields like tmp850 and v850__, the SK ensemble had a lower rmse than the MP ensemble. In the tmp850 and sph850 fields, the SK ensemble had a lower rmse than MP ensemble generally after forecast hour 12. The SK ensemble also had a lower rmse than the MP ensemble for pwat__ between forecast hours 16 and 25. Also in the tmp850 field the SK ensemble had a lower rmse than even the SKMP ensemble for forecast hours 15-29 and in the v850__ field for forecast hours 15-22.

The increased dispersion in large-scale fields is further supported through examination of rank histograms. For nearly every large-scale field examined, the rank histograms at nearly every forecast hour were flatter in the SKMP ensemble than in the MP ensemble (Fig. 4.8). It should be noted that observation error was not incorporated into any verifications in this study. Error in the observation data sources is either unknown or undocumented. Therefore it is inappropriate to make claims regarding proper ensemble dispersion. Instead we can only discuss the differences in dispersion among the ensembles. It should also be noted, however, that given the broad similarities between this study and that of Berner et al. (2015), where the order of performance of various methods of representing model error was not a function of inclusion of observation error, it is not expected that the increased spread, flatter rank histograms, and lower rmse of the SKMP ensemble over the MP ensemble is conditional on inclusion of observation error.

The rank histograms for the SK ensemble were also generally flatter than those of the MP ensemble after the first few forecast hours, but not as flat as those of the SKMP ensemble, again suggesting that this method of accounting for stochastic error is at least as effective as a mixed-physics approach after the added perturbations have had time to accumulate and create diversity among the members. For accppt the rank histograms were not noticeably flatter in the SKMP ensemble than the MP ensemble, although the positive bias in that field makes dispersion characteristics less pertinent.

4.4.2.1.2 Case study

The increased ensemble spread and member diversity can be illustrated via some atmospheric fields from a representative case. We chose the case initialized at 0000

UTC 19 May 2013 as it contained a severe weather event associated with a mesoscale feature (a dryline) forced by a synoptic-scale upper-level trough. The impacts of the perturbations on scales ranging from synoptic to storm scale can be seen. First we examine the 500 hPa height field (Fig. 4.9). Even after the model spun up convection across portions of western Kansas and Oklahoma in the early forecast hours (not shown), the 5760 m height contours in the MP ensemble at later forecast hours show little diversity in areas near and upstream of that convection, which had propagated into eastern Kansas at the valid time in Fig. 4.9. Compared to the same height contours in the SK and SKMP ensembles, it is clear that the perturbations in the SKEB scheme have generated some synoptic-scale diversity. The SK and SKMP ensembles have larger area-averaged ensemble standard deviations (upper right of each panel in Fig. 4.6) than the MP ensemble. While there are still displacement errors (bias) relative to the RAP analysis of that contour in all ensembles, there are members in the SKMP ensemble for which the contour is more accurately placed due to the increased diversity in the ensemble.

The impacts of increased diversity on mesoscale aspects of the forecast are illustrated using precipitable water in Fig. 4.10. The 25-mm contour delineates the dryline extending generally southwestward from central Oklahoma across central Texas. In the late morning to mid-afternoon before convective initiation, the dryline surged eastward before settling at its most eastward location late in the afternoon (not shown). During one particular afternoon hour (1700 UTC, Fig. 4.10), each ensemble placed the dryline too far east in northern Texas and across Oklahoma. However, there was additional diversity in the SKMP ensemble compared to the MP ensemble in which

one or two SKMP ensemble members contained a more westward dryline than in the MP ensemble, closer to the RAP analyzed dryline location. The MP ensemble was more biased and overconfident on the location of the dryline in eastern Oklahoma, whereas the SKMP ensemble gave a more reasonable uncertainty estimate, having members that varied more on the longitudinal placement of the dryline.

Water vapor mixing ratio at 2 m illustrates increased spread at the convective scale (Fig. 4.11). A contribution from the mesoscale variability in the location of the dryline combined with a contribution of diversity on the convective scale resulted in larger ensemble spread in the SKMP ensemble compared to that in the MP ensemble along the dryline in Oklahoma and Texas. There was a small region of enhanced ensemble spread in northwest Oklahoma near a bulge in the dryline. The valid time of the data in Fig. 4.11 is one or two hours before convection first developed near that dryline bulge. The pattern of 1-hr accumulated precipitation from each member in the SKMP and MP ensembles (not shown) suggests there is more variability in the location, coverage, and intensity of the storms that developed near this dryline bulge in the SKMP ensemble compared to the MP ensemble. This difference in variability of precipitation is likely related to the variability in 2-m mixing ratio through changes in buoyancy of surface-based parcels. While there is also mesoscale variability in the placement of the dryline and the bulge in the SK ensemble, there is generally very little spread in 2-m mixing ratio near the dryline bulge compared to that in the SKMP and MP ensembles. The 1-hr accumulated precipitation fields in the SK ensemble also appear very similar among members, thus corroborating the lower diversity compared

to the SKMP and MP ensembles (see section 4.4.2.1.1 for a discussion on the lack of diversity in the SK ensemble).

4.4.2.2 Against observations from MADIS

Verifications of surface variables against METAR and mesonet observations (from MADIS) are shown in Figs. 4.12 and 4.13. Similar to other fields verified against RAP analyses, the spread was larger (a few percent to as high as 25%) in the SKMP ensemble than the MP ensemble for all surface fields by forecast hour 30. The increase in spread was accompanied by almost no change in the rmse of the ensemble mean for 2-m temperature and dewpoint. For 10-m winds the rmse of the ensemble mean in the SKMP ensemble was lower than that in the MP ensemble for all forecast hours, and the difference is statistically significant generally after forecast hour 6. Rank histograms (not shown) for 2-m temperature and dewpoint and 10-m wind components were slightly flatter in the SKMP ensemble than the MP ensemble, but the difference was not as remarkable as for fields verified against RAP analyses. Brier scores for exceedance of certain values of dewpoint and wind speed (Fig. 4.13) also suggest the SKMP ensemble produced better probabilistic forecasts at nearly every forecast hour and threshold compared to the MP ensemble. In particular, the SKMP ensemble had lower Brier scores than the MP ensemble at high dewpoint thresholds, especially between forecast hours 18-24, which correspond to the mid-late afternoon, a time when convective available potential energy is likely to be at its daily maximum and convective inhibition is likely to be at its daily minimum. Therefore, the SKMP ensemble may provide an improved forecast of the thermodynamic environment in a large-scale environment overall supportive of convective storms.

The SK ensemble spread was much lower than the MP ensemble spread, especially at earlier forecast hours, and especially for 2-m temperature and dewpoint (Fig. 4.12). The ensemble spread difference in the 10-m wind component fields was smaller than in the 2-m temperature and dewpoint fields. The lack of physics diversity in the SK ensemble likely negatively impacted the ensemble forecasts in these fields. This was also seen in Berner et al. (2011). The rmse of the ensemble mean was significantly lower in the SK ensemble compared to the MP and SKMP ensembles for 2-m temperature and dewpoint generally between forecast hours 12 and 24 and significantly higher elsewhere. The SK ensemble had a significantly lower rmse than the MP and SKMP ensembles for 10-m wind components throughout the forecast. The SK ensemble had lower Brier scores than both of the SKMP and MP ensembles for light to moderate 10-m wind speeds and also for all but the highest 2-m dewpoint thresholds during the mid-late afternoon (Fig. 4.13). The physics package used for the SK ensemble is chiefly responsible for the superior near-surface wind, temperature, and dewpoint forecasts (section 4.4.1). This result suggests the potential of simplifying the ensemble design in the future by selecting the best physics parameterization scheme and using a stochastic method to sample stochastic model errors which relieve the need of maintaining multiple physics schemes that are not necessarily independent from each other.

4.4.3 PQPF skill

The spread of 1-hr accumulated precipitation in the SKMP ensemble was greater than that in the MP ensemble after the first few forecast hours (Fig. 4.7). Additionally, the rmse of the ensemble mean of the SKMP ensemble was lower than that of the MP

ensemble at most forecast hours, with the SKMP ensemble being significantly better from forecast hours 7-24. Each ensemble had a positive bias of around 0.01-0.05 mm during the second half of the forecast period, and the SKMP ensemble was slightly more biased than the MP ensemble (not shown). The SK ensemble had less spread than the MP ensemble from forecast hours 1-12 and 21-30. The rmse of the ensemble mean was similar between the two ensembles during these periods. In the period of forecast hour 13-19, however, the SK ensemble had increased spread and decreased rmse compared to the MP ensemble. However, this time period corresponds to the morning and midday time when precipitation is less common, so the inference that the SK ensemble provides better PQPFs during this period may not be robust. The SK ensemble was less positively biased at forecast hours 20-25.

In spite of the bias, several grid-point and neighborhood-based probabilistic and traditional verification measures suggest PQPFs were improved by the use of the SKEB scheme in the SKMP ensemble. Considering Brier scores (Fig. 4.14), the SKMP ensemble had a lower Brier score than the SKMP ensemble at nearly all forecast hours for light and moderate accumulation thresholds (0.254 and 6.35 mm). With a lower climatological occurrence of precipitation exceeding 25.4 mm per hour, there was less difference in the Brier score between the SKMP and MP ensembles, and due to a relatively larger standard error, the differences were less likely to be significant. The SK ensemble also had lower Brier scores than the MP ensemble during the middle portion of the forecast period. The duration in which the SK ensemble had lower Brier scores generally decreased with increasing threshold except for at the 25.4-mm threshold,

where the SK ensemble had significantly lower Brier scores than the MP ensemble over nearly the entire forecast range.

The fractions skill score (FSS; Roberts and Lean 2008) was used as a neighborhood-based verification (see section 2.4) metric. A neighborhood radius of 48 km (12 grid points as in Johnson and Wang 2012; Romine et al. 2014 used a similar neighborhood radius of 50 km for precipitation verification) was used for all neighborhood based scores presented as in Chapter 3. There was very little distinction in FSSs between the SKMP and MP ensembles at the lightest threshold (Fig. 4.15). The SKMP ensemble had slightly higher FSSs during the middle part of the forecast (forecast hours 9-24) for moderate and heavy rain thresholds, but the FSSs of the SKMP ensemble were slightly lower than those of the MP ensemble generally after forecast hour 24 at all thresholds. The SK ensemble generally had lower FSSs than the MP ensemble except after forecast hour 23 at the lightest threshold (0.254 mm) and a few sporadic moments in the range of forecast hours 12-18 at the other thresholds. The difference in FSSs between the SK and MP ensembles at heavy rain thresholds was especially large, providing further evidence of the need to incorporate physics uncertainty into a convective-scale ensemble for heavy precipitation forecasting.

A neighborhood-based version of the Receiver Operating Characteristic (ROC) curve was also calculated (Mason 1982; see sections 2.3.2 and 2.4). The area under the ROC curve (ROC area) is shown in Fig. 4.15 as a function of 1-hr precipitation threshold for a few select forecast hours. Each ensemble produced skillful QPFs for light rain thresholds after spin-up issues settled. Depending on forecast lead time, ROC area tended to peak at either the very lightest threshold or the light-moderate (2.54-mm

or 6.35-mm) thresholds and decreased steadily with increasing threshold. For example, at forecast hour 17 (the time of the minimum in domain average precipitation) the highest ROC areas occurred at the lightest threshold and decreased steadily through the highest threshold, whereas at forecast hour 25 (the diurnal peak in precipitation), the highest ROC areas occurred at the 2.54- and 6.35-mm thresholds and decreased more slowly towards both higher and lower thresholds. The shape of the plot at forecast hour 21, during a sharp increase in domain average precipitation, contained features similar to those at both forecast hours 15 and 27. As a function of forecast lead time, skill generally increased until the late part of the forecast with some oscillation leading up to the peak around forecast hour 26-28, corresponding to evening on the next day. The exception is at the lightest threshold (0.254 mm), where skill peaked at forecast hour 16 and slowly declined afterward (not shown). In general ROC areas agreed with FSSs in that the SKMP ensemble had higher scores than the MP ensemble mostly during the middle portion of the forecast (forecast hours 6-24). Outside of that range, the ensembles had approximately the same skill. Reduced overall sample size likely explains the noisy pattern at the higher thresholds.

4.5 Conclusions

As a step towards improving the design of convection-allowing EPSs, the impact of a stochastic kinetic energy backscatter scheme was evaluated for a set of warm season cases over a large portion of the continental U.S. Three seven-member ensembles were constructed for the testing. The SK ensemble contained no physics diversity among the members, but the SKEB scheme was employed. Diversity in this ensemble came from the random seed used to generate the pseudo-random numbers.

The MP ensemble was a mixed-physics ensemble containing variations in the microphysics, planetary boundary layer, and land-surface model parameterizations. The SKEB scheme was not active in the MP ensemble. The SKMP ensemble had the same mixed-physics configuration as the MP ensemble with the SKEB scheme turned on. These ensembles were designed to answer the following questions regarding convective-scale ensemble forecasting:

1) Can a stochastic error representation scheme (SKEB in this case) add meaningful ensemble information and improve the forecast distribution?
2) Is the stochastic error representation method compatible with a mixed-physics approach? If so, does the combination of these methods further improve probabilistic forecasts on the convective-scale?

Each ensemble member had 4 km grid spacing (no convection parameterization was used) and was initialized at 0000 UTC, running for 30 hours to give a complete day 1 forecast of next-day severe weather and heavy precipitation. Both large-scale fields such as temperature, height, and winds above the boundary layer, as well as 2-m temperature, 2-m dewpoint, 10-m wind components, and 1-hr accumulated precipitation were verified using both grid-point and neighborhood probabilistic verification metrics.

The SKEB scheme is designed to (1) correct for insufficient kinetic energy near the grid scales of a forecast model and (2) add spread to the ensemble. The SKEB scheme injects kinetic energy into the model at all scales through additive perturbations to the rotational wind and temperature fields. The NWP model used in this study does not appear to suffer from excessive kinetic energy dissipation in the mesoscales, and thus (1) was not of major concern. For a reasonably tuned SKEB scheme, a positive

impact on ensemble spread and probabilistic forecasts was found. Neither robust nor comprehensive tests of the parameters for optimal tuning of the scheme were performed; optimal tuning of the scheme for use at the convective-scale is left for future work.

The SKEB scheme was successful in accomplishing (2). Marked gains in ensemble spread were noted in nearly every field verified, especially large-scale fields. Spread was even increased in fields that were not directly perturbed (i.e., specific humidity, dewpoint, and precipitation), although the increase in spread in those fields was reduced compared to the increase in other fields that were directly perturbed. The increase in spread is also confirmed through examining the rank histograms. Histograms in the SKMP ensemble were flatter than those from the MP ensembles. The increased spread was accompanied by a reduction in the rmse of the ensemble mean. Some of these reductions were statistically significant.

Quantitative precipitation forecasts were also improved. Since the SKEB scheme does not correct individual grid point errors in deterministic forecasts nor does it perturb the moisture field directly, the connection between the SKEB scheme and precipitation is convoluted and indirect, occurring through changes in stability of air parcels on the convective scale as well as through changes in the wind field that provide forcing for convection initiation and affect ongoing storms. The perturbations are very small, so the changes are subtle but can accumulate over time periods long enough to impact the evolution of the near-surface wind and temperature field enough to affect the initiation or maintenance of convection. It is difficult to determine the precise factors that impact the change in PQPF skill from the use of the SKEB scheme given the

random nature of the perturbations. For some individual cases ensemble forecasts are bound to be improved through a better representation of the uncertainty of the atmospheric state.

The performance of the SK ensemble was competitive with the other ensembles despite the lack of physics diversity. It contained almost as much spread as the SKMP ensemble at many forecast hours and in many fields (thus exceeding that from the MP ensemble) and also had flatter rank histograms than the MP ensemble. The rmse of the SK ensemble mean was also similar to that of the SKMP and MP ensembles. There were even forecast hours and thresholds at which precipitation skill scores in the SK ensemble were better than either of the MP or SKMP ensembles. However, in agreement with Berner et al. (2011), the spread of the SK ensemble was much lower in the boundary layer compared to the MP ensemble. Thus it seems the best choice is to combine the uncertainty in the physical processes impacting temperature, wind, and moisture in the boundary layer by using mixed physical parameterizations with the uncertainty in the dynamics and in other unparameterized subgrid-scale processes by using the SKEB scheme. Hacker et al. (2011) and Berner et al. (2015) also found the combination of a SKEB scheme and physics diversity to give the best forecasts at and below 700 hPa for convection-parameterizing resolutions. Additionally, similar to Chapter 3, the performance of the SK ensemble is likely sensitive to the physics parameterization options used (Morrison microphysics, YSU PBL, and Noah land-surface); this combination was shown to be more accurate than most of the other combinations used in the MP ensemble, and the SKEB scheme parameters were tuned for this particular combination and were not changed when used with other physics

combinations. Since the optimal parameters for the SKEB scheme may be dependent on the choice of physics, the comparison between the SK and SKMP ensembles may not maximize quantitatively the added value of including the SKEB scheme on top of the mixed-physics approach.

This study is among the first to examine the effect of combining stochastic methods with traditionally used mixed-physics methods for convection allowing ensemble design. Given the resources needed to maintain various physics schemes, future research is still needed to explore to what extent a mixed-physics method is needed in the presence of a stochastic method. It is also acknowledged that the conclusion may also be dependent on the diagnostic and verification methods adopted.

The experiment presented in this chapter did not incorporate initial or lateral boundary condition perturbations to isolate the impact of the model-error representation on the ensemble forecasts. Such perturbations could further broaden the forecast probability distribution and reduce or eliminate poor forecasts (present in this study but not discussed) caused by inadequate initial and lateral boundary conditions. Future work should incorporate such initial and lateral boundary condition error representation using advanced ensemble based data assimilation (Johnson et al. 2015) with model-error representation to further improve convective scale ensemble forecasts.

Figures

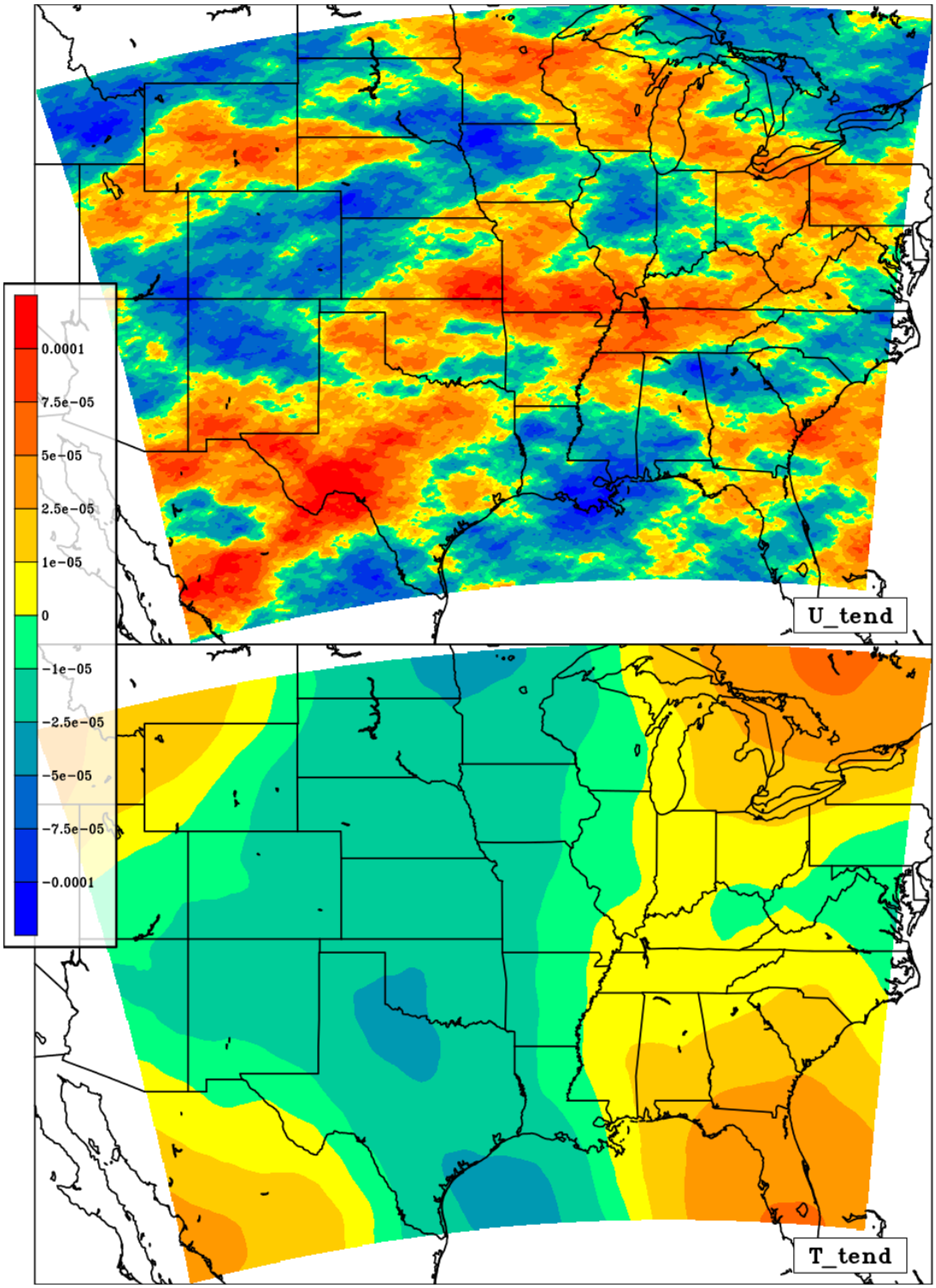

Fig. 4.1. Example forcing tendencies for the u-wind (top; m s^{-2}) and temperature (bottom; K s^{-1}) fields.

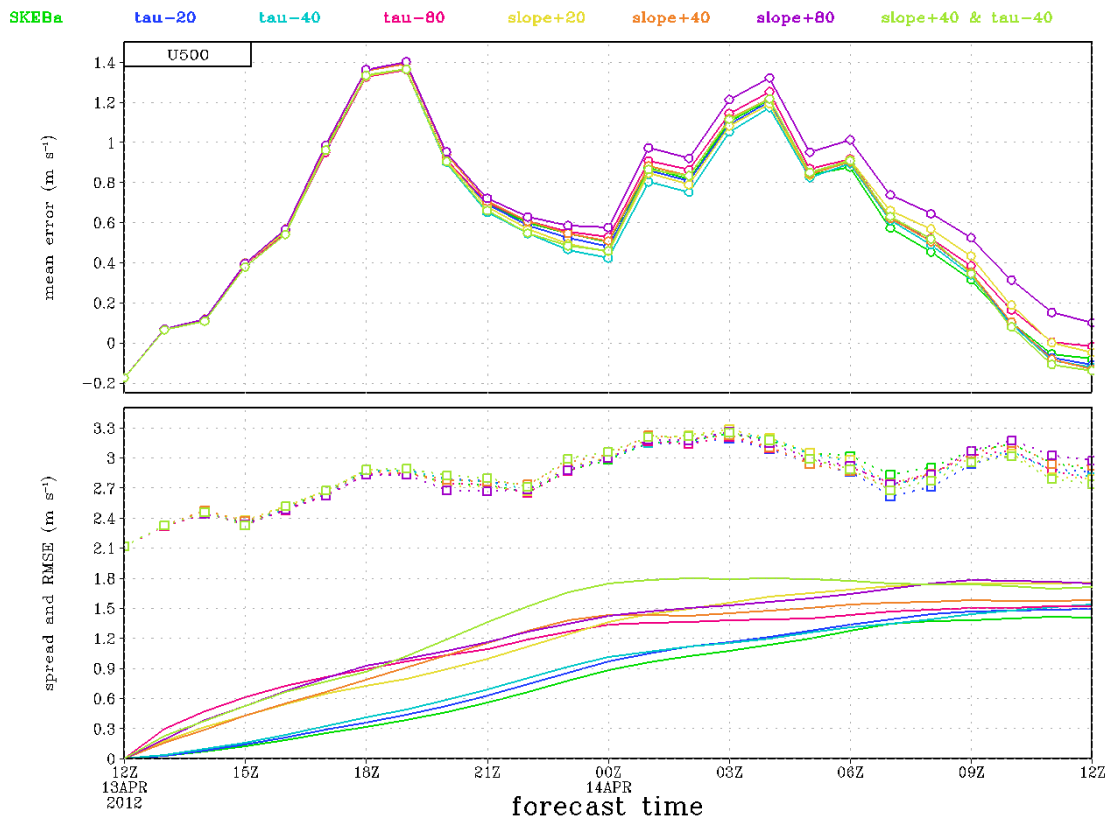

Fig. 4.2. Bias (top), RMSE (bottom dotted), and spread (bottom solid) of the u-wind component at 500 hPa verified against RAP analyses for a test case initialized at 1200 UTC 13 April 2012. SKEB scheme settings are colored according to the key at top. Numbers indicate the percentage perturbation from the default values of the SKEB scheme (3 hr for decorrelation time tau, 1.83 for the spectral slope).

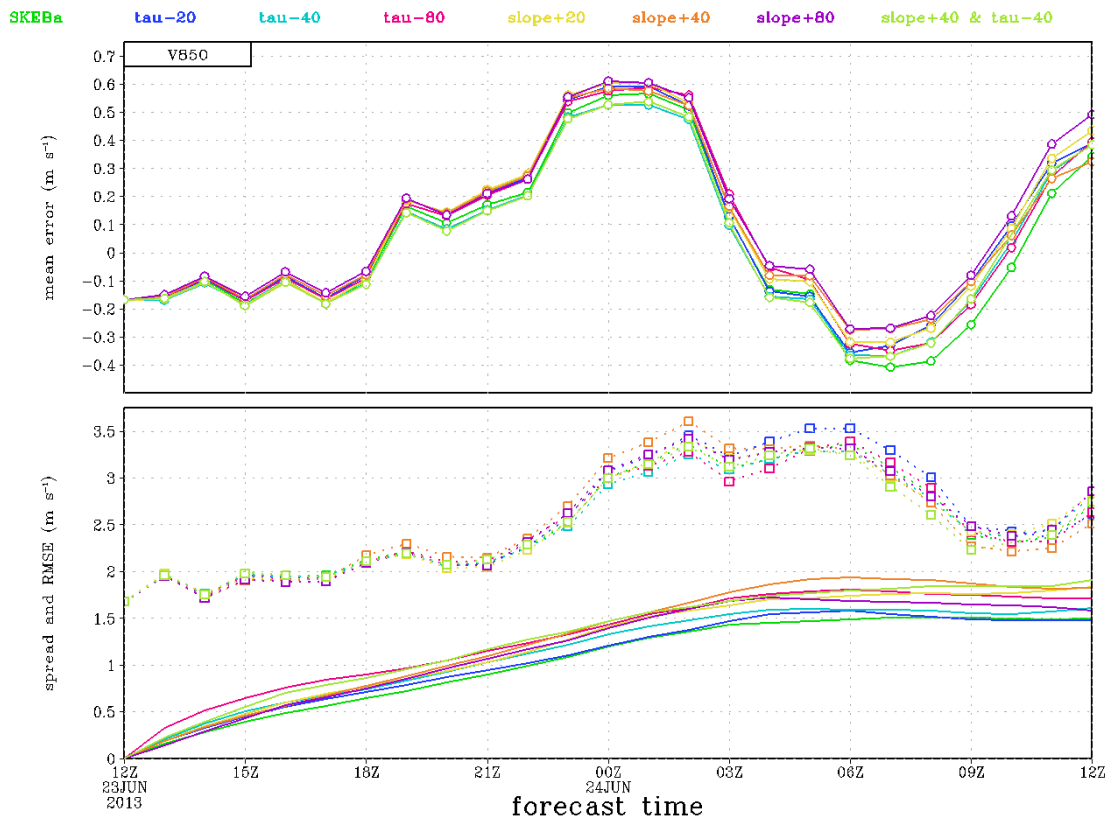

Fig. 4.3. Same as Fig. 4.2 except for the v-wind component at 850 hPa for a test case initialized at 1200 UTC 23 June 2013.

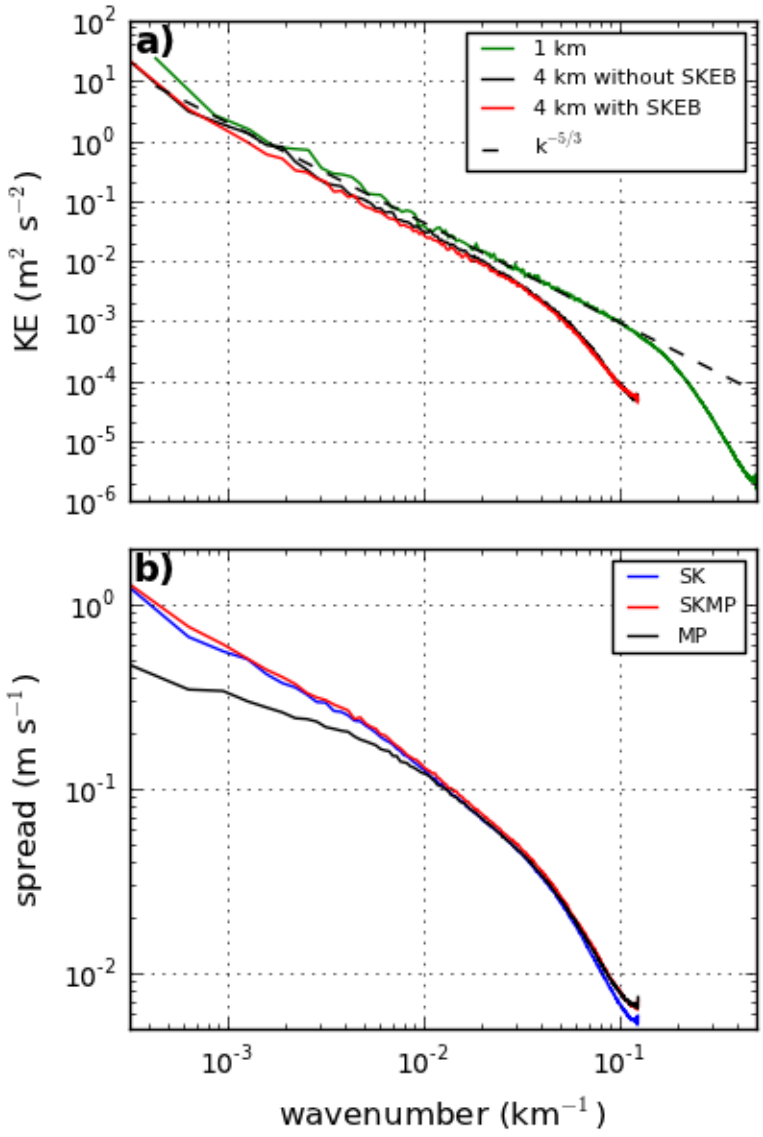

Fig. 4.4. (a) Kinetic energy spectra from WRF simulations with grid spacings of 4 km and 1 km (24 hour forecasts from different initializations). The solid red and black spectra are from otherwise identical WRF simulations at 4 km grid spacing where one uses the SKEB scheme and the other does not. A reference $k^{-5/3}$ slope is included in dashed black. (b) Spectral decomposition of u-wind spread from one case, also a 24 hour forecast.

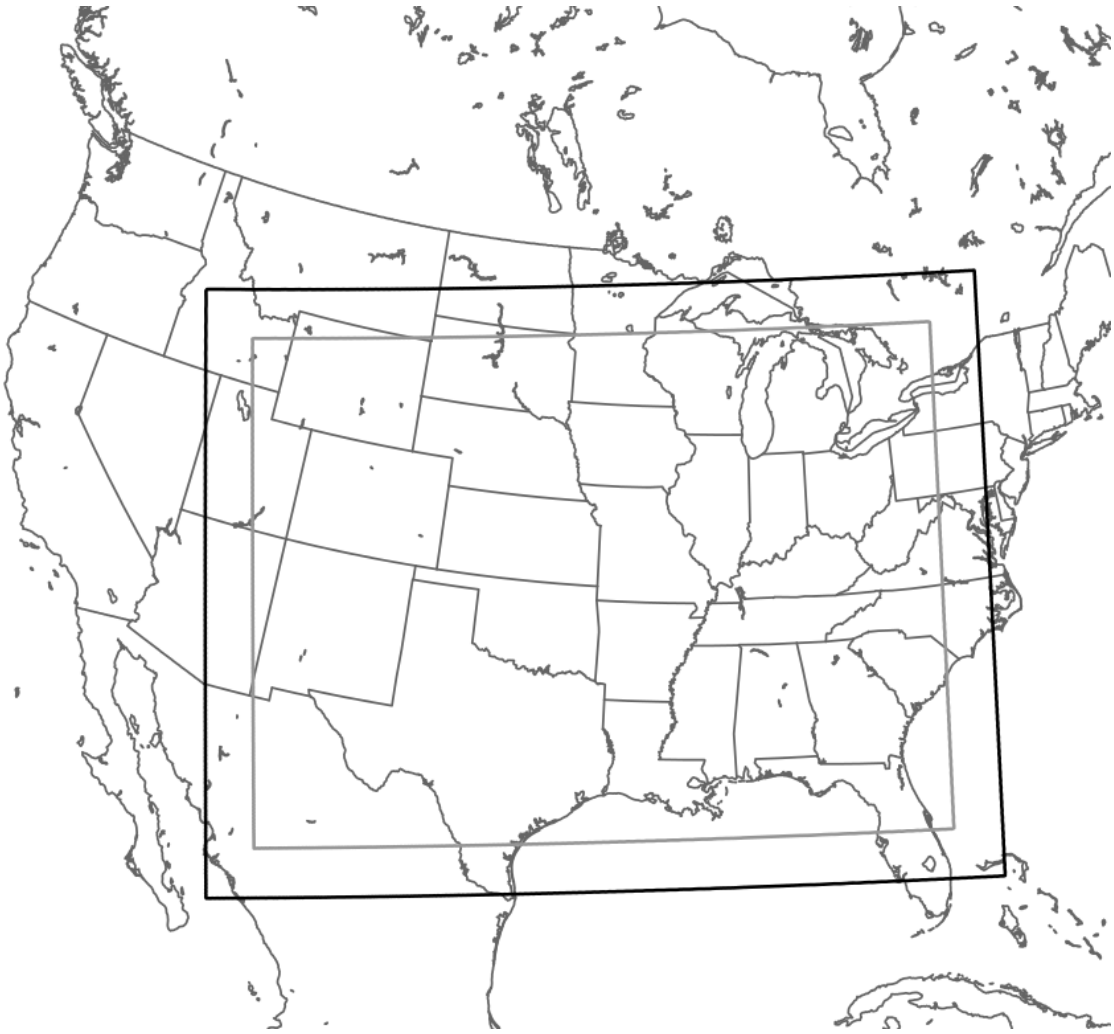

Fig. 4.5. Model (thick black) and verification (thinner gray) domains.

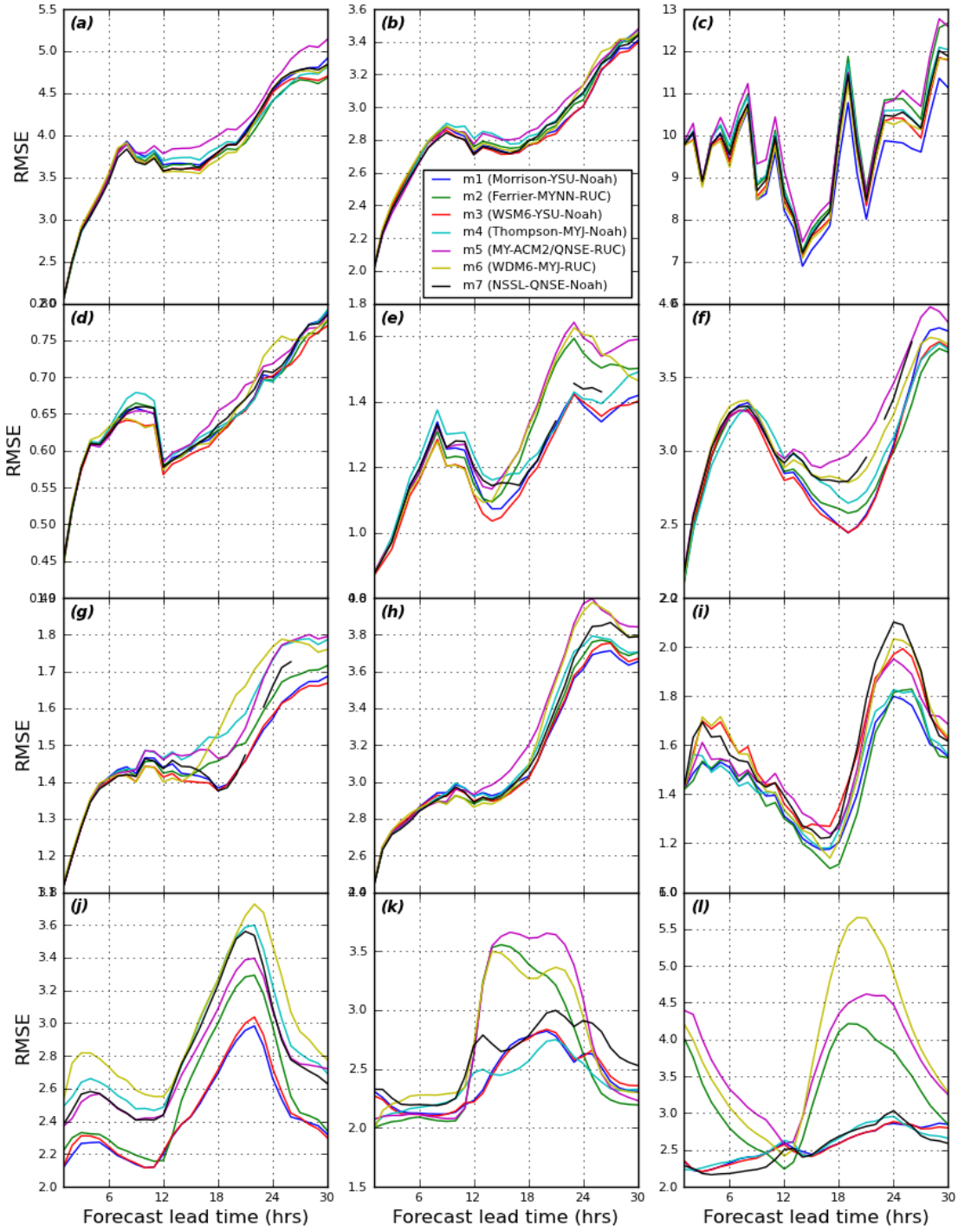

Fig. 4.6. Member rmse for (a) u-wind at 250 hPa (m s^{-1}), (b) u-wind at 500 hPa (m s^{-1}), (c) 500-hPa geopotential height (m), (d) temperature at 500 hPa (K), (e) temperature at 850 hPa (K), (f) v-wind at 850 hPa (m s^{-1}), (g) specific humidity at 850 hPa (g kg^{-1}), (h) precipitable water (mm), (i) 1-hr accumulated precipitation (mm), (j) u-wind at 10 m (m s^{-1}), (k) temperature at 2 m (K), and (l) dewpoint at 2 m (K).

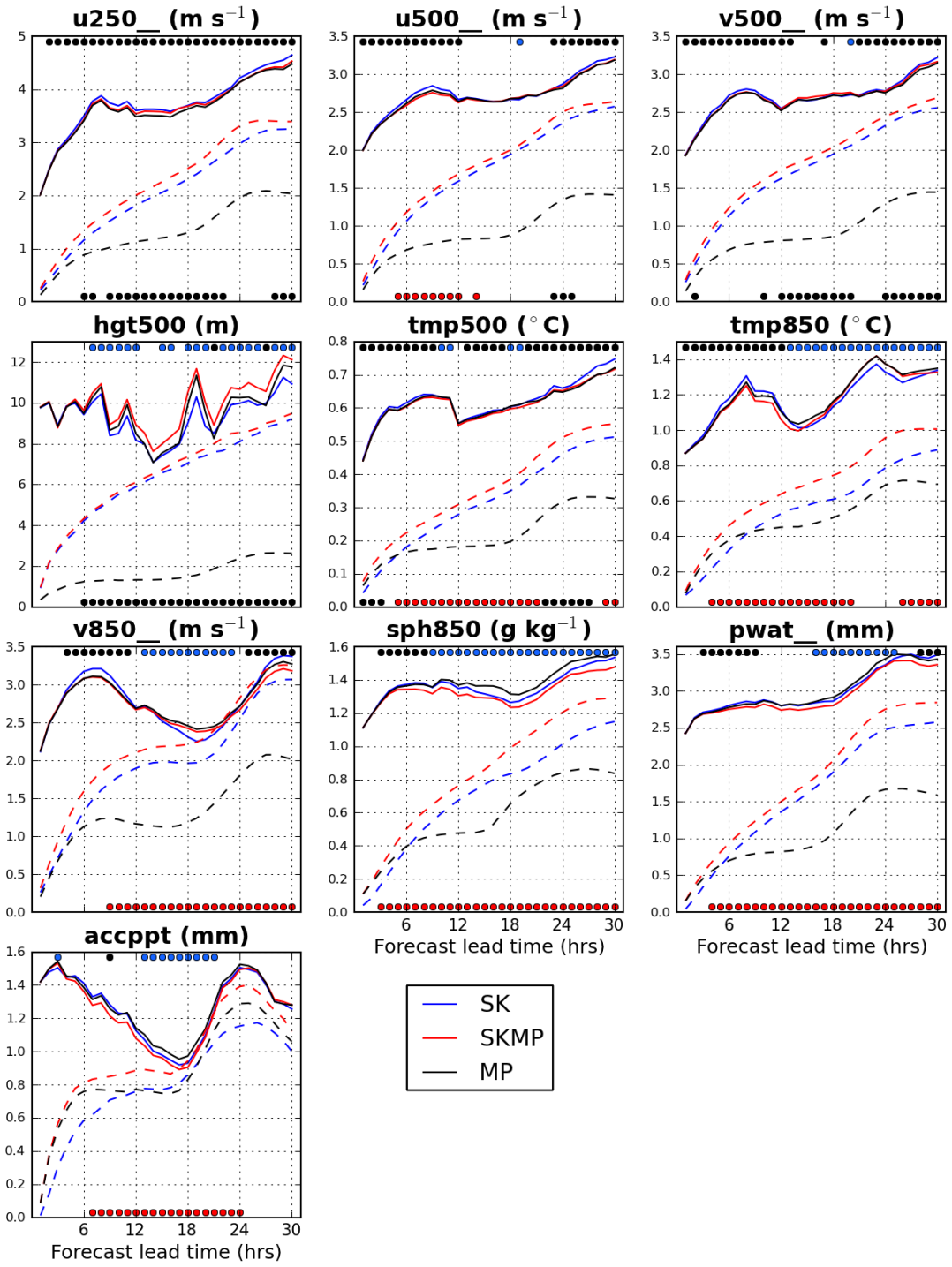

Fig. 4.7. Verification domain-average ensemble mean rmse (solid) and ensemble spread (dashed). Red dots across the bottom indicate forecast hours at which the rmse of the SKMP ensemble was statistically significantly lower than that of the MP ensemble, whereas black dots indicate the opposite. Similarly, across the top of each panel, blue dots indicate when the SK ensemble had a significantly lower rmse than the MP ensemble, whereas black dots indicate the opposite.

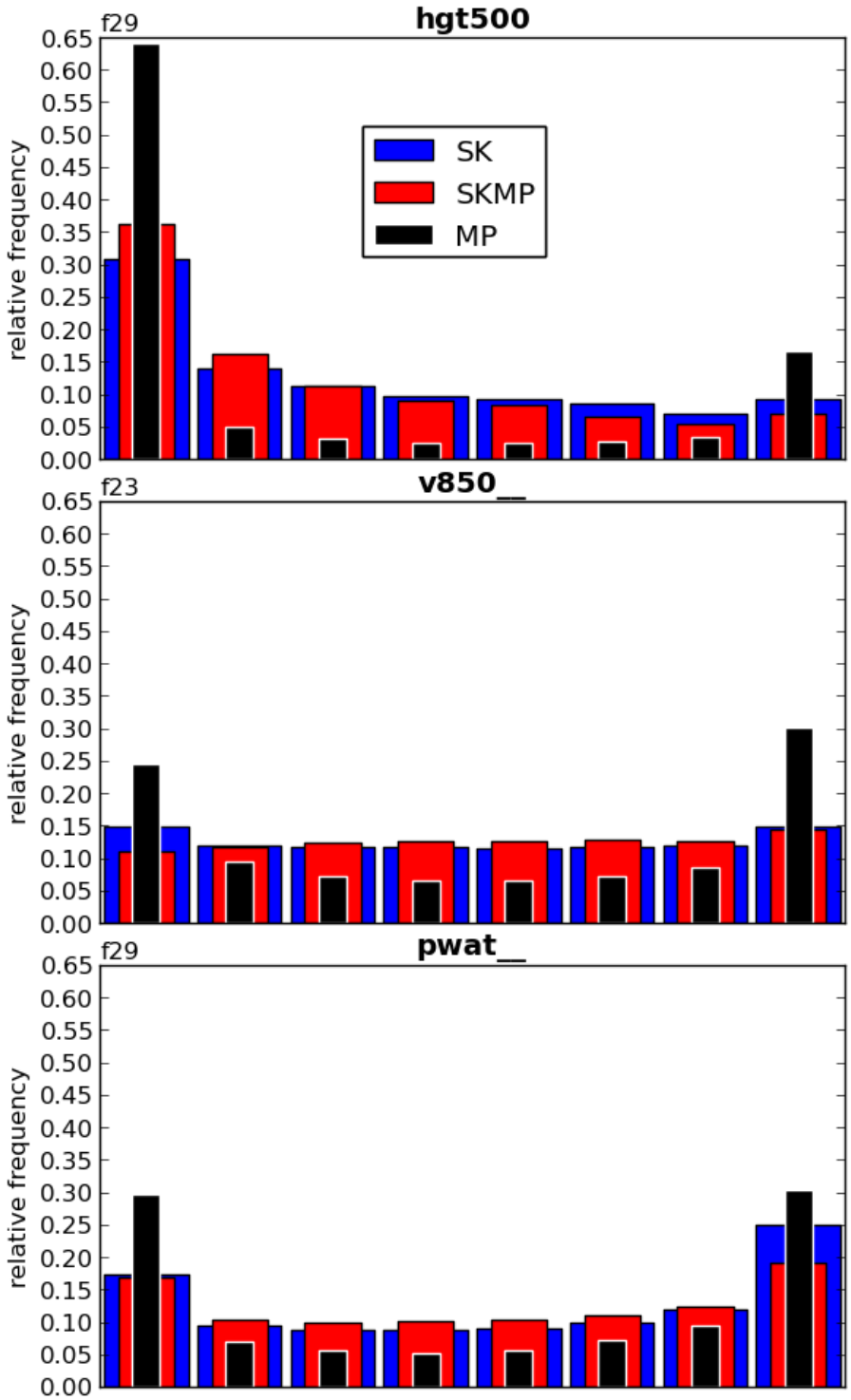

Fig. 4.8. Rank histograms for 500 hPa geopotential height at forecast hour 29 (top), 850 hPa v-wind at forecast hour 23 (middle), and precipitable water at forecast hour 29 (bottom).

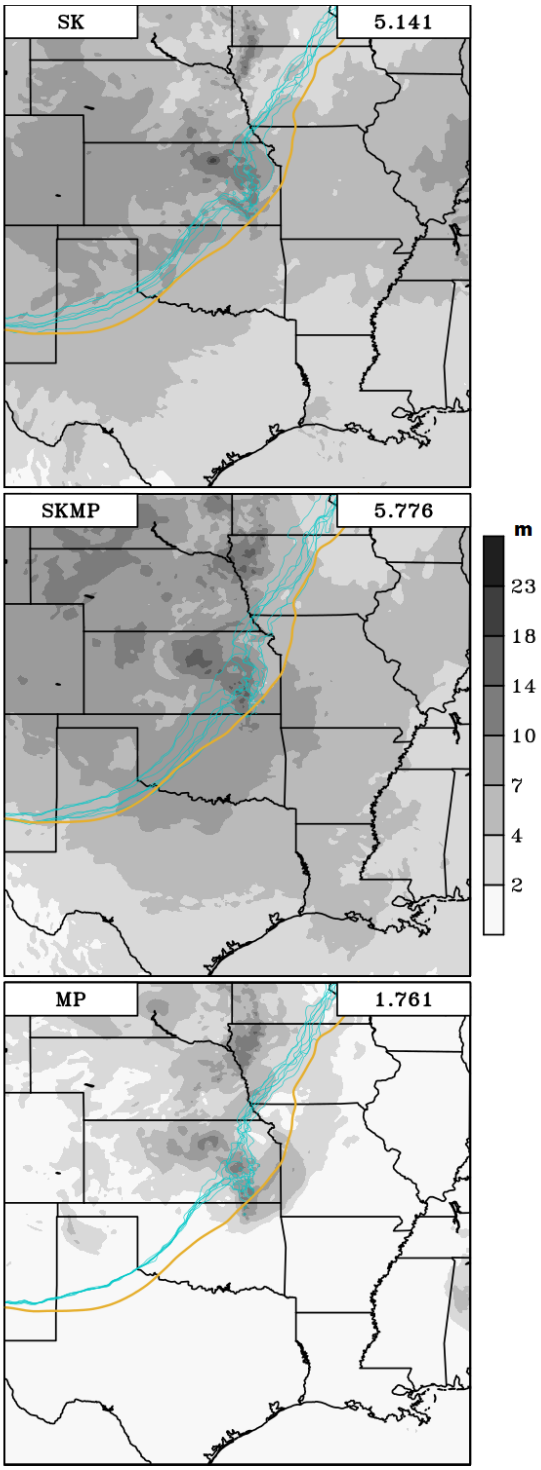

Fig. 4.9. Ensemble standard deviation (shaded, m) of 500-hPa geopotential height valid 1100 UTC 19 May 2013. Individual member 5760-m height contours (light blue) and the analyzed 5760-m height contour from a RAP analysis (gold) are also shown. Area-averaged ensemble spread is indicated in the upper right of each panel.

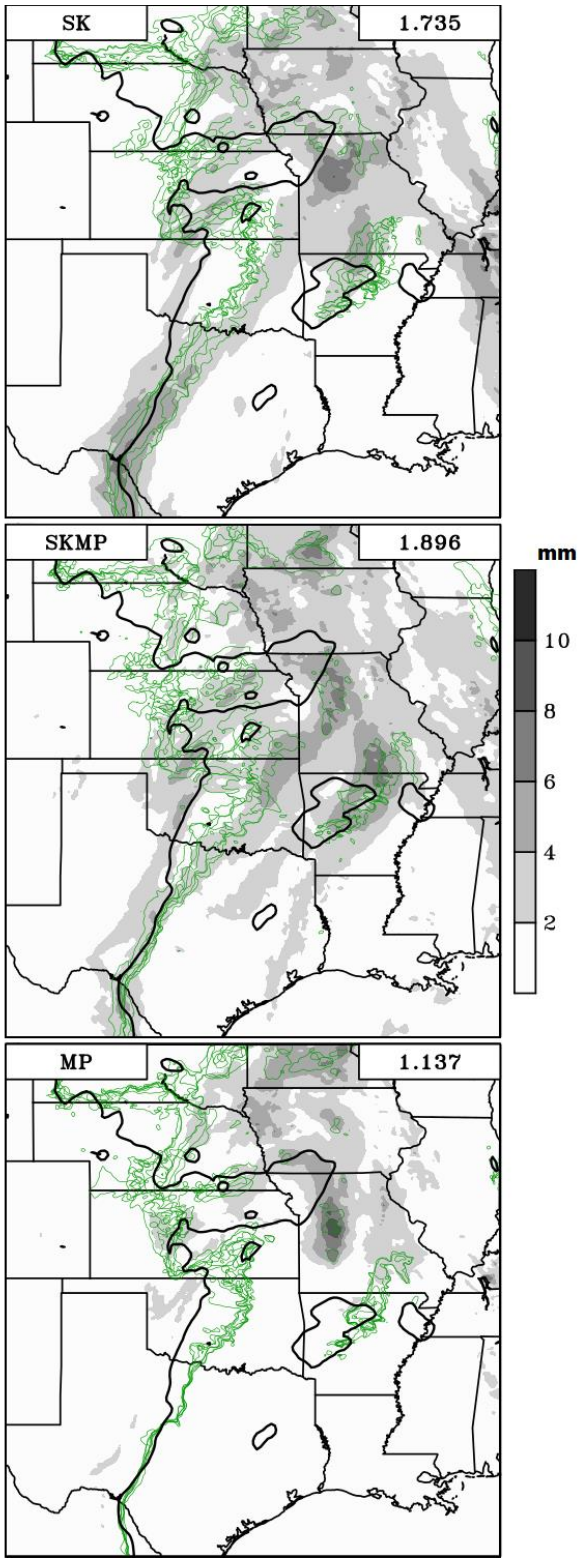

Fig. 4.10. As in Fig. 4.9 except for precipitable water (mm, 25 mm contour is displayed) valid 1700 UTC 19 May 2013. Individual member contours are in green, whereas the RAP analyzed contour is shown in the thick black line.

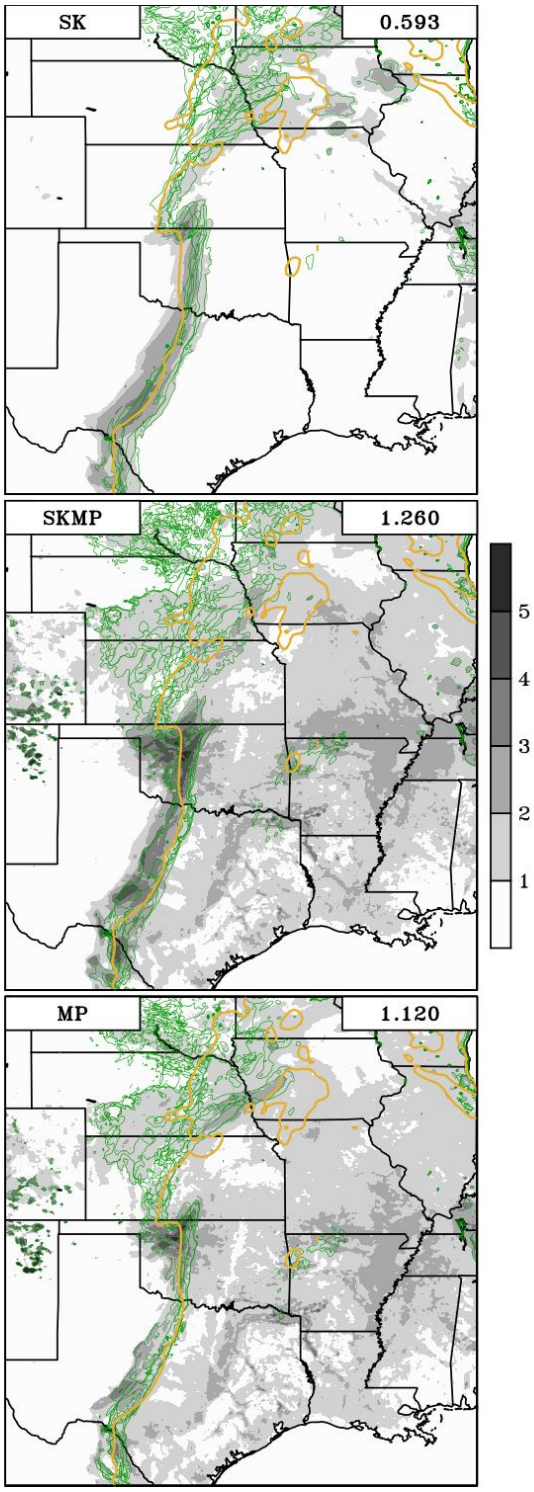

Fig. 4.11. As in Fig. 4.9 except for 2-m water vapor mixing ratio (g kg^{-1}, 12 g kg^{-1} contour is displayed) at 1900 UTC 19 May 2013. Individual member contours are in green, whereas the RAP analysis contour is the thick gold line.

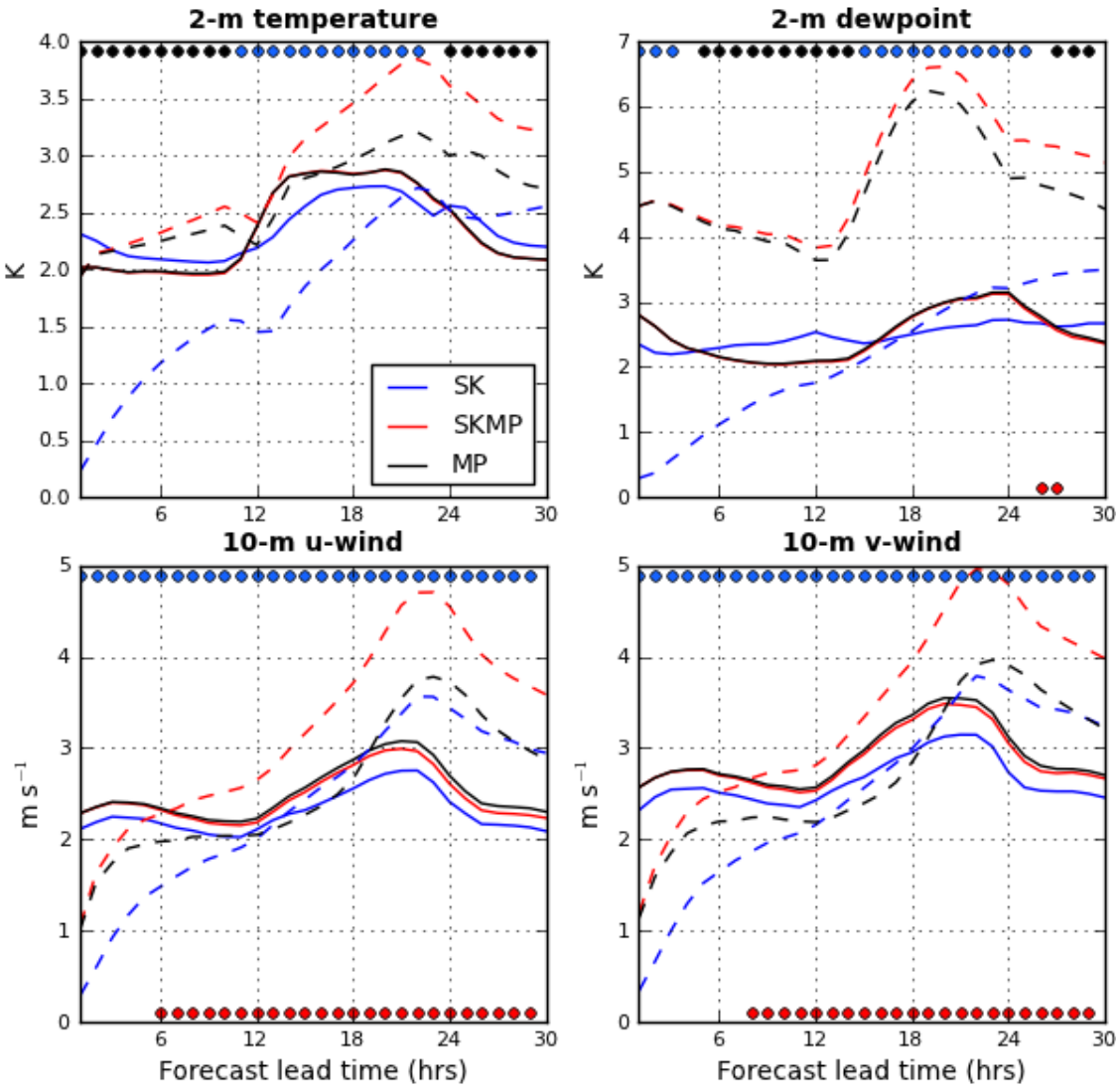

Fig. 4.12. Same as Fig. 4.7 except for the indicated fields verified against METAR observations.

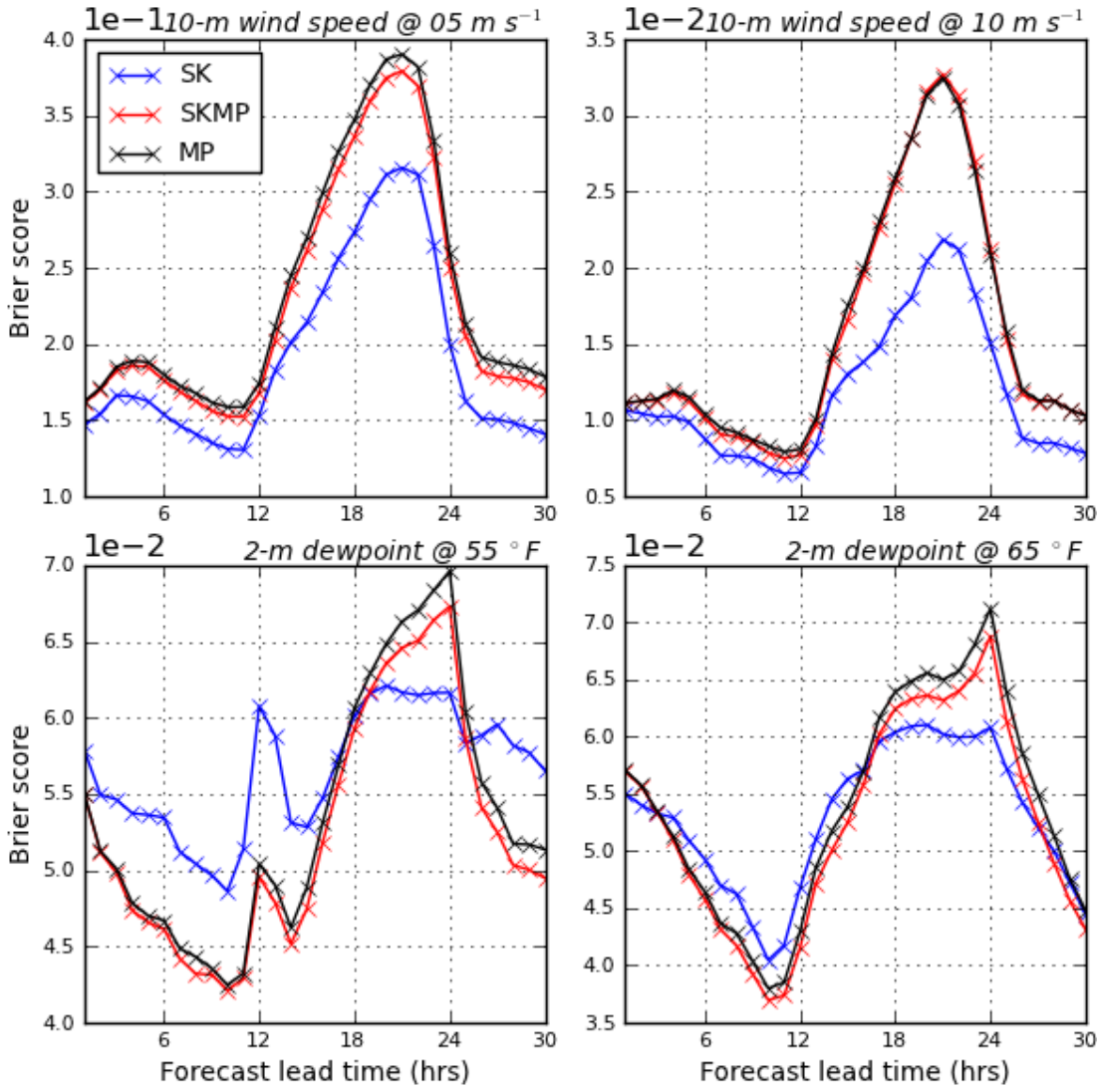

Fig. 4.13. Brier scores for (top row) 10-m wind speed, (middle and bottom rows) 2-m dewpoint forecasts at the indicated thresholds.

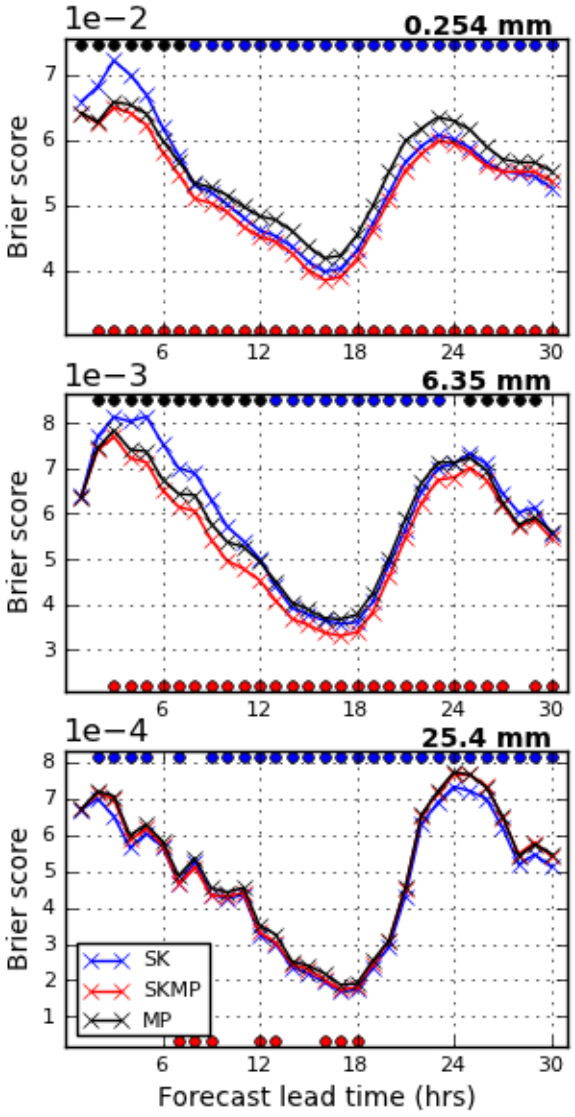

Fig. 4.14. Brier scores for the indicated 1-hr accumulation thresholds. Colored dots represent statistically significant differences as in Fig. 4.7.

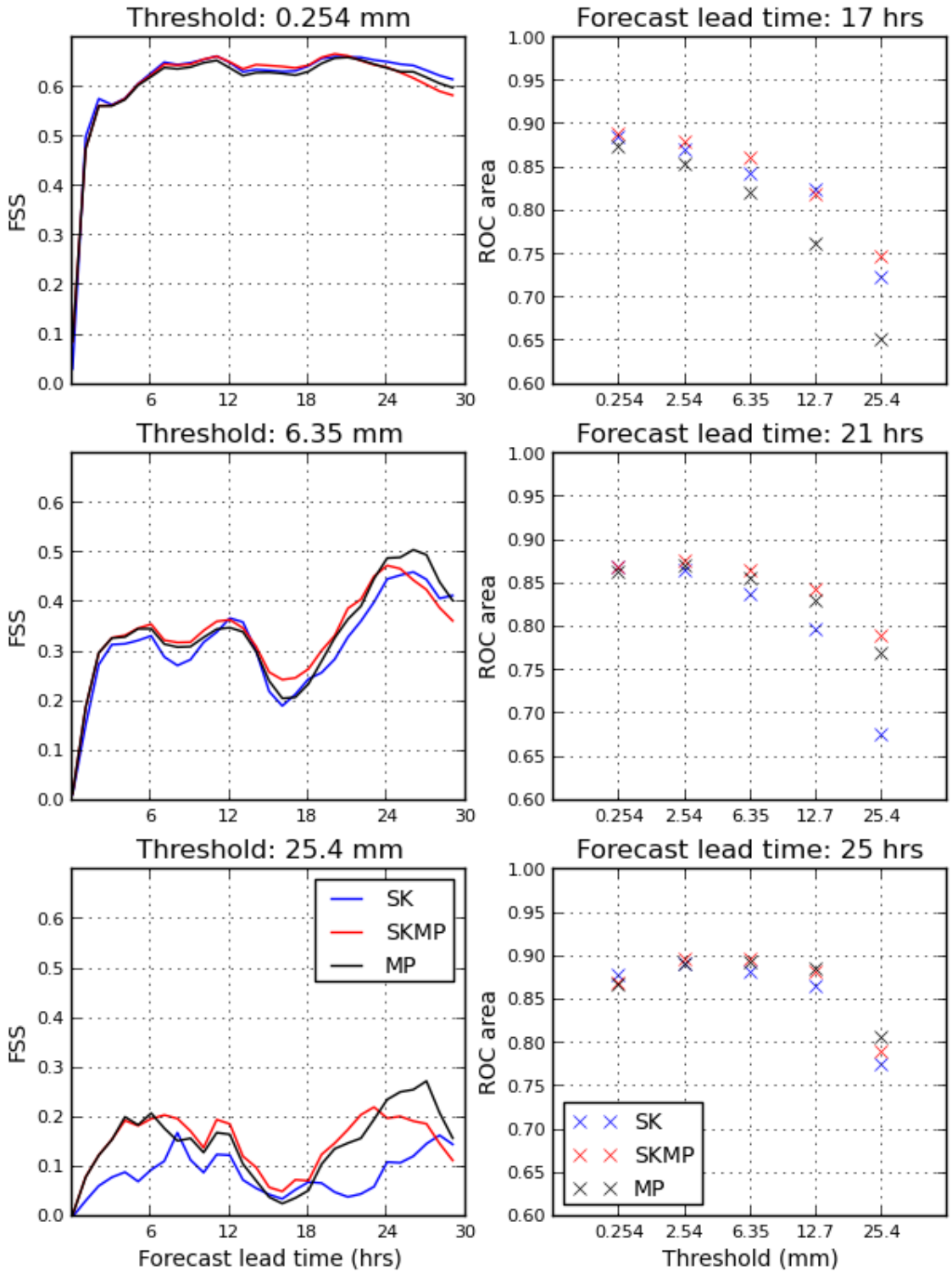

Fig. 4.15. (left column) Fractions skill scores for the indicated 1-hr accumulation thresholds; (right column) area under the ROC curve at the indicated forecast hours for various 1-hr accumulation thresholds. No statistical significance testing was performed on the FSSs or ROC areas.

Chapter 5. Sensitivity of Convection Forecasts to Land-Surface Model and Implications for Ensemble Design

5.1 Introduction and Motivation

The state of the land surface and associated surface-atmosphere exchange processes exert a strong degree of control over the spatial and temporal patterns of deep moist convection via the impacts from turbulent sensible and latent heat fluxes (e.g., Anthes 1984; Rabin et al 1990; Clark and Arritt 1995; Pielke 2001; and Segele et al. 2005). Additionally, a correct specification of the land-surface state, i.e., greenness vegetation fraction, leaf area index, soil texture, and land use is also critical in accurately forecasting deep moist convection, especially during the warm season when the ground surface is exposed (i.e., no snow or ice cover) and vegetation is green and photosynthetically active (Kurkowski et al. 2003; Robock et al. 2003; Godfrey et al. 2005; Miller et al. 2006). Therefore, correct representation of surface-atmosphere exchange processes in land-surface models (LSM) coupled to numerical weather prediction (NWP) models is critical for accurate forecasts of thunderstorms and heavy precipitation.

Much of the prior research on the sensitivity of forecasts of deep moist convection to surface-atmosphere exchange processes have indicated that the initial soil moisture state exerts the most influence. Sutton et al. (2006), for example, showed that the amount of diversity in convection-permitting forecasts of precipitation due to use of different initial soil moisture analyses can rival that from using varying convection parameterizations in coarser simulations. Aligo et al. (2007) illustrated the level of sensitivity of convective-scale forecasts to soil moisture perturbations and found some skill in probabilistic precipitation forecasts based on an ensemble using only those

perturbations. Trier et al. (2008) determined that the variation in initial soil moisture state exerted more influence on forecasts of precipitation than choice of LSM. Finally, while poor coverage of observations is a major source of uncertainty in soil moisture analyses, more dense observations will not eliminate uncertainty, as soil moisture has been shown to be highly spatially heterogeneous even within a small area (Basara 2001).

While the importance of an accurate initial soil moisture state for convection forecasts cannot be ignored, there is additional uncertainty in the formulation of physical processes related to energy fluxes. Chen et al. (1997) compared various methods for parameterizing surface sensible heat exchange in the LSM used by NCEP for their operational mesoscale model and noted sensitivity to the formulation of the stability term and the thermal roughness length in calculating turbulent sensible heat flux. Numerous other studies have also identified sensitivity of precipitation forecasts to the specification of an empirical constant used to calculate the thermal roughness length (Zilitinkevich 1995; Chen et al. 1997), which is used in the formulation of the exchange coefficient in the sensible heat flux calculation (Marshall et al. 2003; LeMone et al. 2008; Chen et al. 2010; Trier et al. 2004, 2011). These studies have shown not only sensitivity, but that various fixed values of this constant may result in better agreement between simulated and observed heat fluxes and better precipitation verification scores than others depending on the scenario. Numerous studies have also suggested that the computation of latent heat flux from plant transpiration contains many uncertainties. Perhaps the most important is the resistance term, whether stomatal resistance or generic canopy resistance, which governs how effectively plants release water through

their leaves into the environment and how effectively that water can be carried out of the canopy and into the lower atmosphere (Chen and Dudhia 2001; Jackson et al. 2003; Godfrey and Stensrud 2010; Kumar et al. 2011). It should be noted that implementation of a more mixed or comprehensive set of variations based on these uncertainties in an ensemble forecast framework has not been attempted.

Among the major physics components in convection-allowing NWP forecasts, which include cloud microphysics, boundary layer, radiation, and surface-atmosphere exchange/LSM, the LSM component remains one for which uncertainty is only minimally accounted for, even in ensembles that sample model error by using mixed physics (see Clark et al. 2009, 2011, Kong et al. 2007, and Xue et al. 2008, for examples of experimental convection-allowing ensemble configurations). Even in ensembles that include mixed LSM parameterizations, the variability is usually limited to one of two schemes, the frequently used Noah LSM (Chen and Dudhia 2001; Ek et al. 2003; these two papers have been cited nearly 2600 times), and the Rapid Update Cycle (RUC) LSM (Smirnova et al. 1997, 2000, 2015; total citations among these papers is over 200), which is sensible considering these schemes are used operationally. However, large uncertainties exist within these and other LSMs which have not been accounted for in convection-allowing ensembles. In this study, an effort to sample LSM uncertainty and investigate the sensitivity of forecasts of convection to perturbations to LSM-related parameters is documented. The implications of adding such LSM perturbations to other physics perturbations in convection-allowing ensemble forecast systems are also discussed.

The remainder of this chapter is organized as follows. A brief exposition on the uncertainties within the LSM component and the role it plays in the WRF model is given in section 5.2. The specific LSMs used in this study are introduced and briefly described in section 5.3. The experimental setup is described in section 5.4. Two case studies illustrating the diversity and sensitivity of forecasts of deep moist convection to various physics perturbations are detailed in sections 5.5 and 5.6. The broader implications of accounting for the LSM uncertainties in convection-allowing ensemble forecasts are discussed in detail in section 5.7. A summary and conclusions follows in section 5.8.

5.2 The role of the land-surface model in the WRF

In the WRF model, the LSM is partially coupled to the surface layer parameterization scheme in that some of the duties in the surface-atmosphere exchange process are shared between the two, with arrays being passed from one to the other where needed. In some NWP models, the two separate schemes may actually be combined into a single scheme.

The LSM component has two main roles in WRF simulations: 1) to calculate and pass heat and moisture flux values to the atmospheric component via the planetary boundary layer (PBL) scheme, and 2) to evolve the soil state and any additional parameterized vegetation state. The chief prognostic variables unique to the LSM component include soil temperature, soil moisture (soil liquid water is used when sub-surface ice is present), and skin temperature. If a LSM parameterizes vegetation, then related arrays such as canopy water content and leaf-area index may also be prognosticated. When snow or ice is present, LSMs typically also integrate snow/ice

areal cover, depth, and water content. The surface layer scheme handles the near-surface wind, temperature, and moisture. In the WRF, surface layer schemes are also designed to calculate turbulent energy fluxes in the event that a LSM is not used (so that any changes to the near-surface atmosphere can be fed back into the rest of the atmosphere).

5.2.1 Formulation of land-surface models

Land-surface models use the surface energy budget as a guide. The surface energy budget is expressed as

$$R_{net} = H + LE + G \quad (5.1),$$

where R_{net} is the net radiation, the difference between the incoming and outgoing components of the shortwave and longwave radiation (several of which are calculated in the radiation parameterization scheme), H is the turbulent sensible heat flux, which transports heat into the atmosphere, LE is the latent heat flux, which carries moisture away from the surface, and G is the ground heat flux, which transports heat into the soil. In essence, the land-state evolution is driven by the net radiation and how it is partitioned into these three basic energy components. LSMs in the WRF do not strictly close the energy budget equation. Rather, each term is calculated independently, with the assumption that the equation should be approximately true. The LSMs in the WRF contain subroutines that decide whether to terminate model integration if the residual becomes too large. One particular scheme dumps the residuals into the history file. Cursory examinations of the residuals suggest that they are of O(1 W m^{-2}) but are highly variable in space and time. Residuals occasionally exceed 10 W m^{-2}. While LE represents water vapor flux at the surface, in LSMs in the WRF it acts primarily as a

diagnostic value; the moisture flux (not coupled to the latent heat of vaporization) is passed to the PBL scheme so moisture can be distributed throughout the atmosphere.

5.2.2 Uncertainty in the LSM

The construction of LSMs contains many uncertainties. Such uncertainties include the formulation of the energy fluxes, the numerical methods used to approximate the governing equations of heat and water transport within the soil, and soil and vegetation state parameters. Most of the attention paid to LSM uncertainty in this work regards the first source of uncertainty. Sources of uncertainty in the sensible and latent heat fluxes are of chief importance.

5.2.2.1 Sensible heat flux

A general formula for sensible heat flux is

$$H = \rho C_p \frac{T_s - T_a}{r_h} \quad (5.2),$$

where ρC_p serves to convert between radiative transfer and kinematic units, T_s represents the surface (skin) temperature, T_a represents the near-surface temperature, and r_h is resistance to sensible heat flux. The uncertainty in the formulation of H rests in the formulation of the resistance term. Resistance to sensible heat flux is the result of properties of the atmosphere within the thin layer immediately in contact with the ground surface where molecular processes dominate over macroscopic processes. It is thought that sensible heat flux effectively originates not at the surface, but at a small elevation above the surface. Similar to the roughness length for wind, the thermal roughness length, z_{0h}, is defined as the height at which heat flux originates, with the resistance coming from the layer between the ground surface and z_{0h}. Thermal

roughness length is effectively impossible to measure, and must be calculated. A number of formulations of z_{0h} have been proposed. One popular formulation (Zilitinkevich 1995) used in the LSM in the operational NAM model is to assume z_{0h} is related to the momentum roughness length by

$$\frac{z_{0m}}{z_{0h}} = \exp(\mathrm{kC}\sqrt{Re^*}) \quad (5.3),$$

where z_{0m} is the momentum roughness length, k is the von-Karman constant, Re^* is the stress Reynolds number, and C is an empirical parameter whose value is unknown. Chen et al. (1997) suggest a standard value of 0.1, but values ranging from 0.01 to 2 have been used in the literature (Zilitinkevich 1995; Marshall et al. 2003; Trier et al. 2004, 2011; LeMone et al. 2008, 2010). Varying C with location, based on land cover and/or soil texture, has also been proposed, but has not been shown to be clearly superior (Trier et al. 2011). Still other formulations of C and z_{0h} itself have been offered (e.g., Chen and Zhang 2009; Chen et al. 2010). Therefore, there is much support in the literature on the uncertainty of the calculation of the thermal roughness length. Many of the above cited papers compared energy flux forecasts to observed values at a limited number of points in an attempt to implement improvements to a given LSM. Accounting for this uncertainty in a convection-allowing ensemble has not been attempted.

5.2.2.2 Latent heat flux

Latent heat flux can be formulated similarly to sensible heat flux as

$$LE = \rho L_v M \frac{q_{v,g} - q_{v,a}}{r_v} \quad (5.4),$$

where L_v is the latent heat of vaporization, M is moisture availability or soil moisture stress factor, $q_{v,g}$ and $q_{v,a}$ represent moisture terms at ground level and in the near-surface atmosphere, and r_v is the resistance to latent heat flux. This formula is more useful in illustrating the underlying physics, and is not always practical since the term $q_{v,g}$ is ambiguous. There are methods that approximate this formula or make assumptions about what $q_{v,g}$ represents. However, it is more common in contemporary LSMs to partition the total latent heat flux into three components: 1) bare soil evaporation, 2) canopy water evaporation, and 3) plant transpiration. The total latent heat flux is then calculated as the sum of the three components. The formulations for these three components are highly varied among existing LSMs. In addition, the calculation of the resistance is very complicated in many schemes, and there are many different ways to formulate the resistance calculation.

5.3 Description of LSMs used

Four land-surface models were used in this study. The LSMs vary considerably in structure, complexity, and in the formulation of physical processes. However, each is complex enough to provide reasonable convective-scale forecasts.

5.3.1 Noah

The Noah LSM is currently used in the operational NAM and is the oldest LSM studied here. However, the Noah LSM also has a long and well documented history of updates, improvements, and performance evaluation, and has grown considerably in its complexity since its initial development in the 1980s. The scheme originated at Oregon State University as a combination of a soil model with a potential evaporation scheme and a canopy model (Mahrt and Ek 1984; Mahrt and Pan 1984; Pan and Mahrt 1987)

with further refinements added by Noilhan and Planton (1989). The most recent full update (including documentation of updates) is provided by Ek et al. (2003), although other improvements have been investigated by Godfrey and Stensrud (2010), Chen et al. (2010), and Kumar et al. (2011). It is unknown whether these improvements have been implemented in the operational version of the scheme. They have not been implemented in the version used in this study.

The Noah LSM contains four soil layers with centers at depths of 5, 25, 70, and 150 cm and a soil bottom at 2 m below ground. The model contains a single combined bare soil/canopy/snow layer. It uses 19 soil texture classes from the STAS dataset and 27 land-use categories from the U.S. Geological Survey (USGS).

5.3.2 Rapid Update Cycle (RUC)

The RUC LSM (Smirnova et al. 1997) was designed to be used with the RUC model (now replaced by the Rapid Refresh model). The overall construction of the RUC LSM is similar to that of the Noah LSM. The following differences from the Noah LSM are noted. Either 5 or 8 soil layers can be used, with 5 being the default value. With five soil layers, the layer centers are at depths of 0.5, 10, 30, 100, and 230 cm with soil bottom at 3 m below ground. The model uses a modified version of the 19 soil texture categories from the STAS dataset, as well as a modified version of the 27 land use categories from the USGS plus an additional category for lakes. One particularly important difference between the Noah and RUC LSMs is in the parameterization of resistance to plant transpiration. The Noah uses the "F-factor" method from Nolihan and Planton (1989), whereas the RUC uses values from a lookup table.

5.3.3 Pleim-Xiu (PX)

The PX LSM is described by Gilliam and Pleim (2010) and references therein. It is meant to be coupled to a specific PBL parameterization scheme (the Asymmetric Convective Model, version 2; ACM2). It was designed for retrospective simulations using a soil nudging algorithm and data assimilation and has also been used in air quality studies. Gilliam and Pleim (2010) note that it has not been tested in a forecast mode and with soil nudging turned off, both of which are the case in this study.

The PX LSM contains two soil layers at 0-1 cm and 1-100 cm below the surface. Therefore, the soil model can be considered to be a force-restore type. The scheme uses a similarly complex formulation as the Noah scheme for latent heat flux and was included in this study based on its availability within the WRF. It uses 16 soil categories with values derived from Noilhan and Planton (1989) and Jacquemin and Noilhan (1990). Plant transpiration is computed independent of vegetation or land use type.

5.3.4 Noah-MP

The multiparameterization-Noah LSM – or Noah-MP – was introduced by Niu et al. (2011). It is based on the Noah LSM, but with a large number of expansions. The soil structure and number of soil texture and land use categories is identical to that of the Noah LSM. Many of the expansions were added to make the new scheme more appropriate for climate-scale simulations. However, many expansions are also relevant to forecasting warm-season diurnally forced convection. Particularly important changes include the addition of a vegetation canopy separate from the ground surface, a semi-tiling approach where fluxes over the bare soil and vegetated portions of the grid box are calculated separately as well as within in the vegetation canopy, and adding

calculations to account for gaps in canopy coverage. A dynamic vegetation model was also added to the scheme. While unique among the features of the LSMs studied here, the dynamic vegetation model was thought to be unlikely to cause perceptible changes over the course of convective-scale model integrations of length on the order of a day or so, and was not used. The namesake of this scheme comes from the fact that many important physical processes within the model, such as computation of stomatal resistance and exchange coefficients, were expanded to allow for multiple formulations to be used. In total, multiple formulations exist for 12 different processes in the version of the Noah-MP scheme used here. For most processes, there are only two choices, but for some processes, up to four options were available. Of the 12 processes that contain multiple options, five were considered important to convective-scale forecasts. These options are detailed in section 5.4.

5.4 Experimental setup

Based on the uncertainties within available LSMs in the WRF, a series of perturbations was developed. These perturbations were then implemented in a set of 10-member ensembles using the Advanced Research WRF (WRF-ARW; Skamarock 2008), version 3.6.1 with 4 km grid spacing.

One perturbation method is to C in (5.3), hereafter the CZIL parameter. Ensemble members were generated by sampling from the following set of values: {0.1, 0.25, 0.5}. It should be noted that due to the separation of duties between the coupled LSM and surface layer schemes in the WRF, exchange coefficients for heat and moisture are not computed in all but the Noah-MP LSMs. Rather, they are computed within the surface layer scheme code and passed to the LSM scheme code. In particular,

the surface layer scheme used in the control member, the Mellor-Yamada-Nakanishi-Niino (MYNN; Nakanishi and Niino 2009), was used to make the perturbations to CZIL. The choice of using MYNN rather than a different scheme was a combination of opportunity and choice. Some of the surface layer schemes used for the physics perturbations in this study do not use CZIL to calculate z_{0h} or do not use z_{0h} to calculate the exchange coefficient. Of those that do (MYNN, MYJ, and YSU), the MYNN scheme was chosen arbitrarily.

A second perturbation methodology was to use the various multiparameterization options in the Noah-MP scheme. To date, no studies exist documenting the performance of the Noah-MP scheme in convection-allowing forecasts, neither deterministically nor in an ensemble framework by using different choices of processes. This study offers a first attempt to document the performance of the Noah-MP scheme in convection-allowing ensemble forecasts. The various multiparamterization options do not require coupling, so ensemble members were generated by comprehensively sampling from the available options for each process while making sure no members used identical sets of options. The options can be set in the WRF namelist, making for a very simple method of generating ensemble members without having to recompile the model. There are 18 combinations of these options. Details on specific options can be found in Niu et al. (2011), but the perturbed options are summarized briefly here. The set of options used in various ensemble members is shown in Table 5.1.2. Namelist option opt_crs refers to a method used in computing plant stomatal resistance: Ball-Berry and Jarvis methods are available. Namelist option opt_sfc refers to a method for computing the exchange coefficient for heat: the method

used in the operational Noah scheme (eqn. 5.3) and a method based on more general Monin-Obukhov similarity theory is available. Namelist option opt_btr refers to a method of computing a soil moisture stress factor for computing stomatal resistance: the method used in the Noah scheme using soil moisture and two methods using matric potential (but different functional forms) are available. Namelist option opt_rad refers to a method for accounting for solar radiation interacting with the vegetation canopy: a modified two-stream method, a two-stream method applied to the entire grid cell, and a two-stream approximation applied only to the vegetated portion of the grid cell are available. Finally, namelist option opt_tbot refers to heat flux at the bottom of the soil (at 2 m depth): zero heat flux or heat flux assuming an annual mean temperature at 8 m depth are the available options.

The third perturbation strategy was to use multiple LSMs. The multitude of differences including soil structure, soil texture and land use classification, and formulation of physical processes among the LSMs provides an excellent opportunity to test the effectiveness of using a mixture of LSMs in an ensemble configuration.

Three ensembles were constructed. A control ensemble, referred to as the fixed-LSM, hereafter FLSM, ensemble contains no LSM perturbations. It uses the Noah LSM with CZIL = 0.1 (the default value). Physics perturbations are applied by using a mixture of microphysics, PBL, and surface layer parameterizations (Table 5.1.1). In contrast, the mixed-LSM, hereafter MLSM, ensemble applies the LSM perturbations. To consider the effectiveness of just the LSM perturbations alone, a third ensemble, the LSM-only (hereafter LSMO) ensemble, was constructed using the LSM perturbations but with fixed microphysics, PBL, and surface layer parameterizations.

Table 5.1.1. Description of member physics. PBL scheme abbreviations that have not been defined in the text are as follows: MYJ = Mellor-Yamada-Janjić, YSU = Yonsei University, ACM2 = Asymmetric Convective Model version 2. The asterisk denotes a suggested coupling by scheme designers. Text in italics indicates the configuration of the LSMO members. FLSM members are configured as shown in the bold text, each using the Noah (CZIL=0.1) LSM. MLSM members are configured as shown in each row. The control member has the configuration of the top row.

Microphysics	**PBL**	**LSM**	**FLSM name**	**MLSM name**	**LSMO name**
Morrison	**MYNN**	***Noah (CZIL=0.1)***	MYNN-MO	MYNN-MO-Z01	Z01
NSSL	**MYNN**	*Noah (CZIL=0.25)*	MYNN-N	MYNN-N-Z25	Z25
Ferrier	**MYNN**	*Noah (CZIL=0.5)*	MYNN-F	MYNN-F-Z50	Z50
Thompson	**MYJ**	*Noah-MP1*	MYJ-T	MYJ-T-MP1	MP1
WDM6	**YSU**	*Noah-MP2*	YSU-WD	YSU-WD-MP2	MP2
WSM6	**ACM2**	*Noah-MP3*	ACM-WS	ACM-WS-MP3	MP3
Thompson	**YSU**	*Noah-MP4*	YSU-T	YSU-T-MP4	MP4
Morrison	**YSU**	*Noah-MP5*	YSU-MO	YSU-MO-MP5	MP5
Milbrandt-Yau	**MYJ**	*RUC*	MYJ-MY	MYJ-MY-RUC	RUC
Milbrandt-Yau	**ACM2***	*PX**	ACM-MY	ACM-MY-PX	PX

Table 5.1.2. Noah-MP LSM namelist options.

LSM	opt_crs	opt_sfc	opt_btr	opt_rad	opt_tbot
Noah-MP1	1	1	1	3	2
Noah-MP2	2	2	2	2	2
Noah-MP3	1	2	3	1	2
Noah-MP4	2	2	1	3	1
Noah-MP5	1	1	2	2	1

The effectiveness of the LSM perturbations against the other physics perturbations was evaluated using a small number of case studies. The case studies featured diurnally driven deep moist convection either strongly forced by large-scale processes, by mesoscale features, or with weak large-scale forcing. In each case severe weather was reported, but the degree of severity and the spatial coverage, as well as storm mode, differed among the cases. Two case studies will be discussed in detail to highlight the sensitivity of convection forecasts to the various LSM perturbations as well as differentiate the impacts of the LSM perturbations from those of the microphysics and PBL perturbations. These cases were preferred since a relevant portion of the model domain in each case in which severe convection occurred was sampled by the Oklahoma and West Texas Mesonets, which offer easily accessible archived observations with high spatial and temporal density as well as observations of soil fields. Data from both networks were used for qualitative analysis and verification in both cases. The model domain was centered in different locations to best contain the events of interest in each case, but covered most of the CONUS.

5.5 Case study 1: 13-14 June 2010

5.5.1 Overview and observations

This case was part of a multi-day outbreak of severe weather, with organized convection occurring daily and severe weather occurring nearly every day between 10-15 June across the Great Plains. The particular event of focus in this case initiated as tornadic supercells in the early afternoon of 13 June in the Texas and Oklahoma panhandles, with continuous redevelopment leading to upscale growth into a substantial mesoscale convective system (MCS) that impacted much of Kansas, Missouri, and Oklahoma through 1200 UTC on 14 June. A few dozen reports of severe hail, wind, and tornadoes resulted from this event (see the Storm Prediction Center severe weather events archive at http://www.spc.noaa.gov/exper/archive/events/ for more information).

A high-amplitude synoptic scale trough was present to the west of the Plains. Slowly moving, it was present for several days leading up to the event and lifted northward, only glancing the Plains, on 13-14 June. Nonetheless, coupled with very high moisture content for early June across the Plains, the trough provided a conducive environment for daily thunderstorms, including sufficient vertical wind shear for supercells. The organized convection in this case was largely the result of mesoscale driven processes, however: the main forcing for initiation and redevelopment was the intersection of an outflow boundary (OFB) from the previous day's convection with a dryline (Fig. 5.1). Widespread severe convection across the central and southern high Plains persisted through the overnight and into the morning of 13 June, leaving a well-defined OFB that sank as far south as far northern Oklahoma and the northern Texas panhandle before slowly lifting during the afternoon. Meanwhile, a dryline formed

across west Texas and mixed east during the afternoon. A bulge in the dryline formed, with the nose near the Lubbock, Texas latitude. The dryline and OFB met in the northern Texas panhandle. The presence of these two boundaries means there were three air masses present: a cT air mass behind the dryline and two mT airmasses – one east of the dryline and south of the OFB, and the other north of the OFB, cooler than the mT air mass to its south, but still warm.

As the PBL grew and destabilized through the early afternoon, lift along the triple point was sufficient to initiate convection near Spearman, Texas between 1700 and 1800 UTC (not shown). More robust development occurred just east of there over the next few hours, and by the late afternoon there was a line of intense cells across northeast portions of the Texas panhandle extending across the eastern Oklahoma panhandle. Some of these storms produced tornadoes, primarily between 2100 and 2200 UTC. After that, continuous development along and to the northeast of the triple point combined with new development along the lifting OFB across northwest Oklahoma resulted in gradual upscale evolution to an MCS. The continuous redevelopment near the OFB was likely aided by continuous inflow of undisturbed very moist and warm air from the south.

5.5.2 Control member forecast

In a general sense, the event was forecast similarly on the mesoscale by each member of each ensemble. Therefore, a detailed evaluation of the forecast will only be presented for the control member. Substantial differences from the control member forecast will be noted in later sections. Simulations were initialized at 0000 UTC 13 June 2010 and run for 36 hours.

The control member forecast the previous day's overnight convection sufficiently well. It placed the weakly organized convection slightly south of where it occurred, but the motion, morphology, and duration were well forecast. Despite the displacement of convection, the control member forecast the OFB close to where it was observed across Oklahoma. To the west, in the Texas panhandle, the OFB was forecast slightly south of where the observations suggest it was. The exact position is not well known due to the paucity of observations in the Texas panhandle.

The control member forecast of the afternoon evolution of the OFB and PBL was not as good. Observed surface winds just north of the OFB remained strongly backed to the E or ESE through 1900-2000 UTC, gradually veering to the SE by the late afternoon. In the model, surface winds veered much more quickly, reaching SE to SSE in the same area by 1900 UTC. In fact, winds across the entire region were veered from the observations[5], including in the warm sector across the rest of Oklahoma and along the dryline in Texas. Observed surface winds in Oklahoma were generally from 150-180° throughout the afternoon, whereas in the model, near-surface winds struggled to stay backed of about 165° in central and eastern Oklahoma, and closer to 180° in western Oklahoma and nearby portions of Texas, where observed winds remained especially backed (to around 150° or more easterly) through the late afternoon (Fig.

[5] Wind components at the first model level were used to represent near-surface wind vectors in this study. Observed wind vectors are at a height of 10 m above ground level, whereas the first model level is typically around 30 m above ground in the southern Plains. However, wind vectors at the lowest model level are very similar to diagnosed 10-m winds in the model, with speed and direction differences of around 1-2 m s^{-1} and 5 degrees or less except in areas affected by precipitation. Therefore, the differences between model and observed winds in this environment is truly a model error rather than representativeness error. More generally, 10-m winds and 2-m temperature and moisture fields are diagnosed in the surface layer scheme. Due to issues with some members producing unrealistic 2-m temperature and mixing ratio values, these diagnostic fields were not used in analyzing forecast accuracy. Throughout this paper biases in near-surface fields will not be discussed unless they are deemed to be legitimate forecast errors with magnitudes larger than those that would occur due to representativity differences between first model level fields and surface fields.

5.2). Dryline evolution was also poorly forecast; it mixed quickly eastward in the model, reaching close to the boundary between western Oklahoma and the eastern Texas panhandle by 0000 UTC 14 June, around 100 km too far east. There was also a bulge forecast, but it was farther north, nudging up against the triple point region (Fig. 5.3). Additionally, the model developed a double-dryline structure in which there were two south-north oriented gradients of moisture in the PBL separated by distances that varied among members but was approximately 100 km. The western gradient was more prominent in terms of the magnitude of the moisture gradient and the wind shift and appears to be the model's interpretation of the main dryline. East of the eastern gradient, moisture quality was well forecast. Between the gradients, however, moisture gradually lessened from east to west. This structure was not sampled by the surface observations. Not only was this double-dryline forecast not represented in the observations, but CAPE was reduced and CIN increased between the gradients compared to RUC analyzed values (Fig. 5.4). With little easterly component to the winds to advect moisture from the east, air parcels from within the region between the gradients advected drier air towards the triple point and OFB.

The control member forecast initiation of cellular convection between 2000 and 2100 UTC, approximately two to three hours late (Fig. 5.5). Due to only a slight displacement error in the OFB across the Texas panhandle, the location of initiation in the model was close to where it was in the observations, located only a few tens of kilometers to the east. Remnant moisture pooled along the OFB, forming a narrow tongue of higher moisture content that extended west towards the triple point with the dryline. As solar radiation warmed the lower atmosphere, CAPE increased and CIN

decreased, so modest amounts of convergence in the PBL were able to force parcels to their levels of free convection (LFC) along the OFB just northeast of the triple point, resulting in convection initiation.

Splitting storms were evident in forecast reflectivity within a few hours of initiation with the right member exhibiting hourly maximum updraft helicity (integrated between 2 and 5 km above ground level; HMUH) exceeding 50 $m^2 s^{-2}$ for several hours, suggesting it was a supercell (Kain et al. 2008). However, convection failed to redevelop along the OFB. Additional convection initiated later across north-central Oklahoma also forced via convergence from the weakening OFB (Fig. 5.5), with yet additional development farther north in Kansas a few hours later. However, most of this convection dissipated with no meso-alpha-scale organization noted (Fig. 5.5). Eventually, the model developed widespread convection across northern Oklahoma and southern Kansas into the overnight, bringing it eastward. However, the model was several hours too late, and the convection moved in the wrong direction.

5.5.3 Sensitivity in the FLSM ensemble

5.5.3.1 Reflectivity evolution

Changes to both microphysics and PBL schemes appeared to play significant roles in the simulation of the cold pool and OFB associated with the previous overnight convection, as no pairs of FLSM members that shared either microphysics or PBL schemes appeared strongly more similar than any other pair of members in their depictions of the extent and strength of the cold pool or the placement of the OFB. The forecast reflectivity in each member was similar on the mesoscale, as each member forecast a non-linear MCS that developed as the model spun up, persisting through to

the next morning, then dissipating at various rates between 1200 and 1800 UTC 13 June (not shown). There were differences on the storm-scale, however. As an example, members MYNN-MO, MYJ-T, MYJ-MY, YSU-WD, MYNN-F, YSU-T, and YSU-MO depicted development of scattered cellular convection behind the sinking OFB across the Texas panhandle whereas the other members either depicted no additional development on the OFB or depicted a small number of larger, stronger cells.

Sensitivity of initiation is illustrated using simulated composite reflectivity in the mid-afternoon in Fig. 5.6. Discrete cells developed in the Texas and Oklahoma panhandles and southwestern Kansas first just before 1900 UTC in members MYJ-T, MYJ-MY, and ACM-WS. Convection initiated last in that region in members MYNN-MO and MYNN-N, doing so around 2100, although in members YSU-WD, MYNN-F, YSU-T, and YSU-MO, convection was weak and limited in areal extent before 2100. Storms did not develop in the same location in each member, not even when accounting for differences in forcing mechanisms. In general, however, initiation occurred in a region of focused convergence and moisture pooling along or near the OFB/dryline triple point. In members MYJ-T and MYJ-MY the veering of near-surface winds in the late morning north of the original OFB was so rapid that the original OFB along the southern edge of the cold pool dissipated while the northern edge of the cold pool in southern and central Kansas became a focused boundary (Fig. 5.7). Convection initiated simultaneously along this northern boundary and at the triple point in the northeast Texas panhandle in member MYJ-T, while initiation was limited to the triple point in member MYJ-MY. Initiation also occurred right on the triple point in member ACM-MY, but in member ACM-WS it was displaced from the triple point by 75 km in an

area of moisture pooling. The OFB tended to wash out in these members as well. In members MYNN-F and MYNN-N, the original OFB remained focused long enough for initiation to occur both along the OFB and the triple point. In members YSU-T and YSU-MO, both of which used the YSU PBL scheme, the OFB became somewhat ill-defined, and convection initiated in areas with localized strong convergence and where moisture pooling was more noticeable. Initiation in member YSU-WD was similar to that in members MYNN-F and MYNN-N rather than YSU-T and YSU-MO, even though it used the same PBL scheme as the latter members.

Each member generally failed to correctly capture the continuous redevelopment along and northeast of the triple point and the subsequent upscale evolution to an MCS. Each member captured the secondary development of convection across north-central Oklahoma that occurred within an hour or two of the primary development, although member MYNN-N contained only small isolated cells that did not grow upscale, unique behavior among FLSM members, likely related to the NSSL microphysics. Each member forecast an additional wave of development across north-central Oklahoma and south-central Kansas after 0000 UTC 14 June. This development resembled the observed reflectivity evolution, although several hundred kilometers northeast and several hours late. While this new development persisted through the end of the simulation in each member, resulting in an expansive complex of convection across eastern Kansas and portions of Missouri, the simulated reflectivity fields do not suggest a well-defined convective and stratiform precipitation region as seen in observed reflectivity. Member MYNN-N forecast only a small area of moderate convection (based on simulated reflectivity values), whereas members MYJ-MY, YSU-

WD, ACM-WS, YSU-T, and YSU-MO forecast a large area of reflectivity with a protruding reflectivity fine line along the southern and eastern flank late in the simulation, indicative of a surging gust front (Fig. 5.8). The fine line was especially apparent in member YSU-WD.

5.5.3.2 Mesoscale environment and PBL structure

As discussed above, the location of initiation was closely tied to the location of the OFB and triple point. The orientation and latitudinal position of the OFB varied among the members. In members MYNN-MO, YSU-WD, MYNN-F, and MYNN-N, the OFB contained an extended segment oriented west-east across northwestern Oklahoma, whereas in members MYJ-T, MYJ-MY, YSU-T, and YSU-MO, the OFB was aligned more southwest-northeast. Also in members MYNN-MO, YSU-WD, MYNN-F, and MYNN-N (but also MYJ-T and MYJ-MY), a similar magnitude of moisture convergence was maintained along the OFB, whereas the convergence was much more broken or spotty in members YSU-T and YSU-MO. The OFB made it the furthest south in member MYNN-F, in which the Ferrier microphysics may have played a role. Research not presented in Chapter 3 indicated the M-Y and WDM6 microphysics schemes tended to produce colder and more expansive cold pools. However, there was no noticeable tendency for the OFB to be further south in members MYJ-MY, YSU-WD, and ACM-MY, which used those schemes. Given the forecast hour, there was probably a lot of interaction between the microphysics and PBL components to mask any tendencies in areas impacted by significant precipitation.

Location and timing differences may have been related somewhat to thermodynamic differences as well. Mixed-layer CAPE (MLCAPE) was the largest,

exceeding 3000 J kg^{-1} by 1900 in a widespread area along and south of the OFB and east of the dryline, in members MYJ-T and MYJ-MY in which convection developed earlier than in other members. However, MLCAPE was among the lowest in member ACM-WS, only exceeding 2500 J kg^{-1} in a small area of northwest Oklahoma, although that was where initiation occurred. MLCAPE in members MYNN-MO and MYNN-N had intermediate values (Fig. 5.9). Overall, initiation tended to occur in or near the region of largest MLCAPE, suggesting the controlling influence of moisture pooling along boundaries in this case. Decrease of MLCIN did not appear to play a major role in timing or location of initiation, as values decreased below -10 J kg^{-1} along the OFB and dryline by 1800 to 1900 UTC in every member (not shown).

The character of the double-dryline region may have also impacted initiation via the size and magnitude of the source region of parcels with reduced CAPE and increased CIN (section 5.5.2). It also tended to cluster by PBL scheme. Members using the ACM2 scheme were overall drier ahead of the dryline and had a broad and smooth moisture gradient. They also had a straighter south-north orientation to the dryline with no noticeable bulge, and a wider dry tongue that extended further north towards the Oklahoma panhandle. Also in these members there was very little CIN more than a few tens of kilometers away from the dryline, consistent with the notion of the PBL being deeply mixed in these members. In members that used the MYNN and MYJ schemes, on the other hand, The width of the double dryline region varied meridionally and the gradient within the double dryline region was overall more spatially heterogeneous. The dryline also arced more towards the southwest in west central Texas than in members using the ACM2 scheme. Behavior in members that used the YSU scheme was

intermediate between that from members that used the ACM2 and MYNN/MYJ schemes. In some MYJ and YSU members, there was even an indication of a secondary minimum in moisture within the double dryline region (not shown). Since wind direction within the double dryline region was nearly uniformly southerly in all members, then failure of redevelopment along the triple point can be explained using the same logic as for the control member.

Ultimately, it is probably impossible to determine the specific forcing mechanism that prompted initiation given the model configuration. The one-hour model output frequency means many storm-scale processes (e.g., lift by individual PBL eddies or horizontal convective rolls) that are probably very important to initiation were not captured. Even if sufficiently frequent output was available, however, such features responsible for convection initiation will not be fully resolved on a 4-km grid. A grid with spacing of 1 km to 100 m would be needed for adequate analysis (e.g., Xue and Martin 2006; Markowski et al. 2006). Therefore, it would be unreasonable to assume the specific mechanism could even be identified in this domain, or would represent an actual circulation.

Since the only variability in the FLSM ensemble came from microphysics and PBL differences, and since no significant precipitation fell in the warm sector during the daytime (therefore, microphysics diversity played no role in differences between ensemble members), an examination of PBL structure may also elucidate causes of thermodynamic differences among ensemble members. Point soundings taken in multiple locations show the PBL mixed thoroughly in all members, so that assumption will be maintained throughout all future discussion on PBL structure. West of the

dryline, members ACM-WS and ACM-MY generally had the driest, deepest, and warmest PBLs during the daytime, although near-surface potential temperature was very similar among all members (Figs. 5.10a,b,c). Members MYJ-T and MYJ-MY were the most moist near the surface. Members using the YSU scheme were also among the more moist members during the afternoon. Members using the MYNN scheme were similarly dry as those using the ACM2 scheme. PBL growth rates diverged strongly during the afternoon, resulting in maximum PBL heights that differed by several hundred meters (Fig. 5.10a). The rankings of late-afternoon PBL heights do not correlate well with sensible heat fluxes, however (not shown). Since the definition of PBL top differs among the schemes, some of the differences in PBL height in Figs. 5.10a,d,g do not correspond to differences in physical processes. Point soundings in Fig. 5.11 reveal discrepancies between the subjectively analyzed PBL top and the numerically calculated PBL top in some members. In particular, the PBL top in members MYJ-T and MYJ-MY appears to be much closer to that in members MYNN-MO, MYNN-F, and MYNN-N (MYNN scheme), which suggests the numerically calculated PBL top in the MYJ scheme is slightly higher than that analyzed. One feature of note, however, is a shallow secondary mixed layer above the top of the PBL in the late afternoon in members MYNN-MO, MYNN-F, and MYNN-N which may explain a rapid decrease in near-surface mixing ratio just before 0000 UTC 14 June in those members (Fig. 5.10c).

In the warm sector east of the dryline, the average behavior of the members was similar to that west of the dryline. Members ACM-WS and ACM-MY had the warmest, deepest, and driest PBLs (Figs. 5.10d,e,f). Members MYJ-T and MYJ-MY were the

most moist, and there was a range of about 1.5 g kg^{-1} in average mixing ratio between FLSM ensemble members during the afternoon (Fig. 5.10f). Oddly, in member MYNN-N the PBL was noticeably cooler and shallower, almost an outlier among the FLSM ensemble members, although members MYNN-MO and MYNN-F were also among the cooler members. Despite the higher PBL moisture content compared to areas closer to and behind the dryline, a shallow layer with absolutely stable lapse rates remained above the top of the PBL in the warm sector (not shown), thus precluding development of convection, especially given the absence of forcing for ascent.

Behavior north of the OFB was more mixed due to the variable size of the region (Figs. 5.10g,h,i). However, there was still a tendency for members ACM-WS and ACM-MY to be warm and dry, members MYJ-T and MYJ-MY to be moist, and members MYNN-MO, MYNN-F, and MYNN-N to be cool and moist. Members using the YSU scheme were in the middle in terms of heat and moisture content.

5.5.4 Sensitivity in the LSMO ensemble

5.5.4.1 Reflectivity evolution

On the mesoscale, the forecast evolution of the previous day's convection was very similar among the members of the LSMO ensemble, much more similar than among the members of the FLSM ensemble. The mesoscale environment around the time of initiation of the tornadic supercells was also similar among the members. However, there were still noticeable differences among the LSMO ensemble members in terms of timing of initiation and subsequent evolution of convection. For example, members MP2 and MP5 contained a large, strong cell in the northeast Texas panhandle at 2000 UTC while the other members were either lacking in storm coverage or only

had low reflectivities just before storm initiation, which would follow by 2100. However, members MP1, Z50, Z25, and MP4 still only had a single small cell or a cluster of very small cells at 2100. By 2200, however, robust convection had initiated in all members (Fig. 5.12). Splitting storms were evident in all members, and updraft helicity (not shown) suggested the right split rotated, even if only briefly, in each member.

Similar to the evolution in the FLSM ensemble, the LSMO ensemble members failed to maintain the initial cells. In fact, the initial storms had almost completely dissipated in members MP2 and MP5 by 2300, followed by member RUC and to some extent member MP4, by 0100 UTC 14 June. Convection was maintained longer in other members. Secondary development of robust convection also occurred near the north-central Oklahoma/south-central Kansas border shortly after the primary initiation in the Texas panhandle, similar to the FLSM ensemble. However, coverage and longevity of this development was limited, especially in members RUC, Z50, and MP4. Additional redevelopment occurred in some form in all members generally after 0300. However, also similar to the FLSM ensemble, the LSMO ensemble members did not adequately evolve that convection into an MCS. However, many members forecast a large complex of storms in eastern Kansas into western Missouri overnight (Fig. 5.13). In member Z50, in particular, there was evidence of somewhat well-defined convective and stratiform zones, and in members MP2, Z50, and MP5, hints of surging outflow were noted in reflectivity. In member PX, on the other hand, there was little indication of any attempt by the model to develop mesoscale organization of existing convection, and

most of it had dissipated or was dissipating by 0900, a time when a leading convective line/trailing stratiform precipitation MCS was still present in observed reflectivity.

In summary, mesoscale error growth appears to be much slower in the LSMO ensemble than in the FLSM ensemble with respect to evolution of organized convection. However, by late in the simulation, differences had grown sufficiently as to cause noticeable differences in radar representation of meso-alpha-scale convective activity.

5.5.4.2 Mesoscale environment

Likely due to the reduced impact of LSM perturbations on mesoscale evolution during the night, the meso-alpha-scale features important to convection initiation, chiefly the OFB, were very similar among the LSMO ensemble members by the late morning of 13 June. The OFB was present in virtually the same location and with similar structure and orientation in each member throughout the early afternoon (not shown), thus providing most of the explanation for the similarity in location of initiation of convection. However, there were noticeable mesoscale differences in terms of the moisture quality and temperature. In particular, members RUC and PX were noticeably more moist everywhere (Figs. 5.14c,f,i), although the drying west of the dryline in member RUC was considerably greater than in any other members, rendering it one of the drier members by the early evening. Member MP4 was also moist west of the dryline, but not in the warm sector. Members MP4 and PX tended to be cooler with shallower PBLs in both air masses, and member MP2 was also cool (point soundings show this better than air mass averaged quantities, not shown). Member RUC was cooler in the warm sector, but not behind the dryline, for reasons that will be explained

later. Members MP1, MP3, and MP5 had the warmest and driest, and some of the deepest, PBLs in the warm sector in the afternoon (Figs. 5.14d-f). Overall, differences among the members in near-surface mixing ratio and potential temperature of 1-2 g kg^{-1} and 1-2 K, respectively, were common throughout the late morning and afternoon across Oklahoma and Texas. The moisture variability is similar to that from the FLSM ensemble, but the temperature variability is larger in the LSMO ensemble. North of the OFB, members Z01, Z50, and Z25 were the warmest and driest instead of members MP1, MP3, and MP5. Members RUC, MP4, and PX were cool, but member MP4 was dry.

Since the drier members tended to be warmer and vice versa, differences in equivalent potential temperatures (theta-e), which is highly correlated with CAPE, were rather small. Near the OFB, theta-e tended to be largest in members MP1, MP5, and PX (not shown), in which the maximum CAPE prior to initiation also tended to be larger by up to 500-1000 J kg^{-1} in very localized areas (Fig. 5.15). Unlike the FLSM ensemble, MLCAPE exceeded 3000 J kg^{-1} in all members, so there was less variability in CAPE among LSMO members than among FLSM members in this case. Since member MP2 had less CAPE but was one of the first members to develop a strong storm, stability indices alone cannot explain differences in initiation timing. The focus and strength of near-grid-scale forcing probably provides the ultimate explanation for timing differences, but as described in section 5.5.3.2, it is probably impossible to analyze those forcings in this case. Like in the FLSM ensemble, however, there was focused low-level convergence on the near-grid-scale in each member prior to initiation (not shown)

5.5.4.3 Surface energy flux impacts on the lower atmosphere

The primary feedback from the LSM component to the atmospheric component of the WRF occurs through the sensible heat flux and moisture flux. Since these processes are bottom boundary layer conditions for the atmosphere, LSM perturbations would be expected to be the most apparent in the lower atmosphere, namely, within the nocturnal stable boundary layer and the afternoon convective boundary layer. Therefore, the focus of the rest of this section is on surface energy fluxes.

Early afternoon sensible heat flux at is shown in Fig. 5.16. Large differences among the members can be found, especially in western Texas and eastern New Mexico, suggesting much more diversity than in the FLSM ensemble south of the OFB and especially west of the dryline. In this area sensible heat flux is the largest in members MP1, RUC, and MP5. Members MP1 and MP5 use the generic Monin-Obukhov method in the Noah-MP LSM (Niu et al. 2011) to calculate the surface exchange coefficient, and the exchange coefficient for heat is much larger over most of the region in those members compared to most other members, including the other Noah-MP members (not shown). The higher exchange coefficient means that for a given temperature difference between the lowest model level and the surface (skin temperature) more heat is transported from the surface into the atmosphere, which should result in higher near-surface temperatures and warmer and deeper PBLs. There is a stark contrast between members MP1 and MP5 and members MP2, MP3, and MP4, the latter of which use the Chen et al. (1997) method to calculate the exchange coefficient. The exchange coefficient in these latter members is much lower, restricting transport of heat away from the surface. As a result, there is an accumulation of heat at

the surface, and the energy is instead forced to be transported deeper into the soil, which is manifest as larger ground heat flux and warmer soil temperatures (Fig. 5.17d-f). The opposite is true for members MP1 and MP5. Sensible heat flux values are also consistent with the values of CZIL in members Z01, Z25, and Z50. In the MYNN surface layer scheme, as CZIL increases, z_{0h} decreases, so the exchange coefficient decreases, as does sensible heat flux. Member Z50 uses CZIL=0.5 and has the smallest sensible heat flux compared to members Z01 and Z25. Member Z01 uses CZIL=0.1 and has the largest sensible heat flux. Because it also has a higher exchange coefficient, sensible heat flux is among the highest in the LSMO ensemble west of the dryline in member RUC. However, east of the dryline, member RUC is closer to the ensemble mean in terms of sensible heat flux. As a result, there is less vertical mixing in the PBL east of the dryline in member RUC, and it remains one of the cooler and more moist members.

Surface moisture flux is of chief importance in driving variability in moisture content among LSMO members. The magnitude of moisture flux is dependent on many factors, including soil moisture availability, moisture contrast between the surface and the lower atmosphere, amount of condensed water present on the vegetation canopy, and exchange coefficients, but since each LSM parameterizes moisture flux differently, exchange coefficient formulations for a given process like plant transpiration differ among the members. Therefore, differences in total moisture flux are the result of a combination of differences in the formulation and calculation of each of the factors that contribute to it, which can make a comparison among schemes less straightforward. In spite of this difficulty, an attempt to classify and explain the differences will be made.

Moisture flux accumulated over the daytime of 13 June and prior to initiation is shown in Fig. 5.18. There is a large overall increase in moisture flux from west to east across the region, the result of not only a west-east increase of soil moisture, but also of an increase in vegetation density across the region. Across Texas and western Oklahoma moisture flux is especially larger in members RUC and PX with more isolated areas of larger moisture flux in members MP1, MP2, and MP4. Moisture flux is the lowest in member MP3. The streaky nature of the large moisture flux in member RUC is the result of large bare soil evaporation (not shown), with some of the liquid provided by the previous overnight convection. On average, however, bare soil evaporation in member RUC was higher than in any other member regardless of antecedent soil moisture. While bare soil evaporation was also high in member PX, plant transpiration drove much of the higher moisture flux in that member (not shown). The overall low moisture flux in member MP3 was the result of rather low plant transpiration, unique behavior among members that used the Noah-MP scheme. There were large differences in the spatial patterns of both ground evaporation[6] and especially plant transpiration among Noah-MP members, suggesting an effective perturbation strategy by varying opt_crs, opt_btr, and opt_rad, all of which should influence these processes.

A detailed examination of the moisture flux behaviors in the LSMO ensemble members is offered using Fig. 5.19, taken from a point near the Oklahoma Mesonet site at Chandler. There is a significant clustering of total surface moisture flux: members using the Noah scheme and the RUC and PX schemes have the most moisture flux

[6]Because of the semi-tiling approach in the Noah-MP LSM, bare soil evaporation is only one component of the full moisture flux from ground evaporation.

during the daytime, whereas members using the Noah-MP scheme have much less moisture flux. Since there was very little precipitation during the simulation at this location, there was correspondingly little canopy water evaporation, so ground evaporation and plant transpiration provided effectively all of the total moisture flux. The overall large bare soil evaporation in members RUC and PX is apparent, although it should be noted that there were grid points at which these members did not have the largest bare soil evaporation. The exchange coefficient for bare soil evaporation in members PX and RUC was much larger than that for ground evaporation in the Noah-MP members (not shown). Most of the total moisture flux was due to plant transpiration in the Noah members, while the PX LSM also calculated very large plant transpiration. The soil moisture tendency in the PX scheme is likely a result of the soil structure. It uses only two layers to 1 m depth, shallower than the other schemes. Thus it is presumably easier for moisture to move into the atmosphere from the deep layer. The PX scheme also uses different land cover/vegetation categories than the other members, so perhaps that is part of the cause of differences as well. It should also be noted that, although the RUC LSM uses the same soil texture and land use categories as the Noah and Noah-MP LSMs, the threshold values of soil parameters (e.g., wilting, saturation, and soil matric potential parameters) are different in the RUC LSM. An additional simulation was run by swapping these values from those used in the Noah scheme to determine if these parameter differences were the cause of the greatly different moisture flux behavior in member RUC. Only minor changes to moisture flux were noted in the additional run, suggesting that is not the cause.

In two of the Noah-MP members, plant transpiration decreased to nearly zero during portions of the daytime (Fig. 5.19f), caused by stomatal resistance reaching very large values (not shown). Stomatal resistance is in the denominator of the exchange coefficient for plant transpiration in the Noah-MP LSM, so the drop in plant transpiration is due to the drop in the exchange coefficient. In Noah-MP members in which plant transpiration did not approach zero, stomatal resistance instead became very low during the daytime, thus the exchange coefficient became large and transpiration proceeded. The behavior of plant transpiration in specific Noah-MP members varied by location, however. Regardless of location, the exchange coefficient for transpiration tended to be very similar in members MP1 and MP5. The exchange coefficients in members MP2 and MP4 were very similar at some locations, but at other locations they differed greatly. The behavior of the exchange coefficient in member MP3 was intermediate between the other pairs of members. At locations where rain did not fall during the simulation, the exchange coefficient decreased during the day, but not as much as it did in members where plant transpiration ceased. At rainy locations, the exchange coefficients tended to be very similar among all of the Noah-MP members. The exchange coefficient for transpiration in members Z01, Z25, Z50, and PX reached similar, but slightly larger, magnitudes as members MP1 and MP5 during the daytime (not shown). The resultant transpiration in members Z01, Z25, Z50, and PX was thus larger than that in members using the Noah-MP LSM, although that is probably not the only factor explaining the difference in transpiration between sets of members. The exchange coefficient for transpiration was largest in member RUC, but perhaps due to the difference in formulation of plant transpiration in the RUC LSM

(i.e., the use of a plant coefficient from a lookup table), transpiration was more similar to that of the members using the Noah-MP LSM rather than being overwhelmingly larger than all other members, as implied by the large exchange coefficient.

In general, the RUC LSM parameterized the largest bare soil evaporation, although other members, especially those using the Noah-MP LSM, occasionally had similar values. Also, the PX LSM parameterized the largest plant transpiration, although at some locations members using the Noah LSM and some of the members using the Noah-MP LSM had similar or larger values. It is also worth noting that at locations where no precipitation was forecast, soil moisture was almost identical between members, with differences in 5-cm soil moisture content differing among the members by less than 1%. However, forecasts of soil moisture were not always good. At the Chandler mesonet site, for example, there was a bias of around 0.15 m^3 m^{-3} in both 5-cm and 25-cm soil moisture in all members except PX, in which the bias was highly variable in time. However, the soil moisture tendency forecast was good at the Chandler site, where very little precipitation was forecast. This was the case at other sites where soil moisture observations were available as well. Biases in forecast soil moisture are likely the result of either poor initial soil conditions provided by the NAM model analysis or representativeness errors resulting from interpolating from a 12 km grid.

5.6 Case 2: 19-20 May 2013

5.6.1 Overview and observations

There was a regional outbreak of severe weather on both 19 and 20 May 2013, with over 1000 combined reports of severe weather, the majority of which were in the Plains and Midwest. The focus of these two events was in Kansas and Oklahoma, where

numerous tornadoes, several rated EF2 and higher, occurred. In particular, an EF4 tornado occurred near Shawnee, Oklahoma on the 19th and an EF5 tornado struck Moore, Oklahoma on the 20th.

The synoptic scale pattern was favorable for a large severe weather outbreak. A deep and slow moving trough sat over the Rocky Mountains with a strong jet for late May across the Plains. Lower atmospheric moisture was also high, with surface dewpoints ranging from 67-72°F (Fig. 5.20). Finally, a common triggering mechanism for the region, a dryline, was present both days, with a strong moisture gradient across the dryline. The resultant high instability, high shear, and strong triggering mechanism provided the necessary ingredients for severe weather.

The evolution of storms each day is summarized briefly as follows. The first storms developed along the dryline in south-central Kansas around 1900 UTC 19 May. Within an hour, additional storms developed just west of Oklahoma City, Oklahoma with additional storms partially filling in along the dryline between Oklahoma City and Wichita, Kansas, as well as southwest of Oklahoma City. All storms quickly became supercellular. After a series of storm interactions, three main supercells remained across central Oklahoma by 2200, producing a few of the strong tornadoes. Meanwhile, to the north, storms across the Kansas-Oklahoma border rapidly became linear by the late afternoon, but additional storms redeveloped along the tail end of the line in north-central Oklahoma, achieving supercellular characteristics briefly. These storms continued producing severe weather well into the evening, finally gradually weakening after 0600 UTC 20 May.

A stationary front had moved into northern Oklahoma and eastern Kansas by the morning of the 20th. The dryline remained in southern Oklahoma and Texas. Storms rapidly developed along both boundaries around 1900 UTC. Several discrete cells developed along the dryline in southern and central Oklahoma with little space between them. Therefore, storm interaction played a role in which storms survived to produce tornadoes, including the EF5 that impacted Moore, which occurred at around 2000. By late in the afternoon, a string of large and strong thunderstorms existed, stretching from near Wichita Falls, Texas northeast through northern Missouri. These storms slowly grew upscale to form a squall line after 0000 UTC 21 May, when the simulations for this case ended.

It should be noted that observations from the Oklahoma mesonet, as well as from National Weather Service observing locations in Kansas and Oklahoma were insufficiently dense to determine why storms formed in the specific locations that they did on 19 May, although Trier et al. (2015) provide some insights. Meso-gamma-scale thermodynamic variability along the dryline appears to have played a significant role.

5.6.2 Control member forecast

Simulations in this case were initialized at 1200 UTC 19 May 2013 and integrated for 36 hours. The initialization time represents a major difference from the first case, since the initiation of the target storms in this case was within the first 12 forecast hours, whereas in the first case, the target storms developed after forecast hour 18 and after an intervening episode of convection. Therefore, unlike the first case, initial condition errors had less time to impact the forecast and upscale growth of storm scale error was not a significant factor in the forecast skill of this case.

The control member predicted the storm mode and location with great accuracy on the 19th. However, some mesoscale aspects of the forecast were not as good. As in the first case, winds near the surface were veered in the model compared to the observations. Observed wind directions at 1900 UTC ranged from 170°-190° throughout Oklahoma east of the dryline. In the model, near-surface winds were uniformly from about 180°-190° (Fig. 5.21). Also, the dryline was rotated in the model compared to the observations. The observed dryline was oriented south-north, whereas in the model the dryline was oriented more SSW-NNE and was farther east by a few 10s of kilometers than in the observations. There was some evidence of moisture pooling along the dryline in north-central Oklahoma and south-central Kansas in the observations. Moisture pooling was also suggested in the model, but to a lesser extent. In general, the model appeared to be too dry near the surface by 1-2 g kg^{-1}. Unlike the first case, no obvious double-dryline structure was apparent, which is somewhat at odds with a study of the same case using a similar configuration of the model (Trier et al. 2015).

Storms initiated between 2000 and 2100 in Butler County, Kansas, about one county too far east compared to observed reflectivity (Fig. 5.22), and about 1-2 hours late. These storms progressed northeastward, but did not evolve into a linear configuration as in the observations, instead remaining cellular and discrete. Meanwhile, in Oklahoma, a storm initiated in Oklahoma and Lincoln Counties just after 2100, almost two hours late (Fig. 5.22). However, the storm was in approximately the same location as an observed storm at the time. Additional storms developed shortly after, one in particular, in Pottawattamie and Seminole Counties, very near where an

EF4 tornado occurred near Shawnee. These meso-gamma-scale aspects of the forecast are of remarkable accuracy for a model with 4 km grid spacing. However, the forecast evolution of convection afterward did not match the observed evolution of convection as closely. Whereas the observed discrete storms in Oklahoma dissipated from northeast-southwest in the late afternoon to early evening, the modeled storms maintained their intensity and expanded in coverage as they moved into northeast Oklahoma (Fig. 5.22). Also, there was no redevelopment in north-central Oklahoma in the model, but due to the spurious storms forecast by the model, the forecast reflectivity resembled observed reflectivity in southeast Kansas and northeast Oklahoma in the evening.

While the reflectivity forecast was fairly accurate despite a late initiation, other aspects of the control forecast differed from observations. Observed surface winds backed throughout the afternoon and into the early evening, reaching 150° in many areas, especially in south-central Oklahoma, and even to 135° within the cold pool of the central Oklahoma storms as the cold pool advected northwestward into north-central Oklahoma. Winds also backed in the model, but not enough to match the observations. Wind directions only reached about 180° in south-central Oklahoma, but did reach 160° in portions of eastern Oklahoma. Near-surface winds in north-central Oklahoma backed significantly within the cold pool of the forecast storms, matching the observed winds in the cold pool. However, the observations suggested a separation between the cold pool air mass and the dryline, which remained approximately stationary throughout the day. No such separation was evident in the forecast, as the dryline mixed eastward somewhat until the early evening when it halted and began to retrograde. Forecast storms appeared

to remain mostly anchored to the dryline until about 2300, an hour or so later than any hint of anchoring in the observations. Finally, the cold pool was too dry in the model compared to the observations. Observed dewpoint temperatures at stations impacted by the cold pool from the Oklahoma storms bottomed out in the 60-65 °F range before rebounding, which converts to mixing ratio values of approximately 11.5-13 g kg^{-1}. Forecast near-surface mixing ratio values dropped to less than 11 g kg^{-1} over a several-county area in north-central Oklahoma, with some locations dropping below 8 g kg^{-1}. Such errors with moisture content are likely related to microphysical processes. Errors with dryline position are probably the result of PBL scheme errors.

Around midday on 20 May, the forecast dryline and front were too far east compared to the observations, by over 100 km for the stationary front in northern Oklahoma, and by approximately 50 km for the dryline in southern Oklahoma. Also, the moisture gradient was much too relaxed in the model, where the rich moisture was displaced more than 100 km east of the forecast dryline, but where the observations suggested rich moisture remained immediately against the dryline. Also, as on the 19th, forecast near-surface winds were veered relative to the observations by 10-30°. Storms developed in the model around 2000, 1-2 hours too late, and as a result of the eastward bias of the surface boundaries, about 100 km east of observed storms (Fig. 5.23). However, forecast convection developed into a broken line of intense storms in eastern Oklahoma, roughly matching the observed convective morphology.

In summary, much like in the first case, errors from the PBL scheme likely caused errors with the evolution of the moisture content and winds within the PBL during the hours leading up to convection initiation. Also, the forecast dryline had an

eastward bias. Storms also developed a few hours late, not uncommon with 4 km forecasts of explicit convection, but acquired a similar morphology as in observed radar imagery.

5.6.3 Sensitivity in the FLSM ensemble

5.6.3.1 Reflectivity evolution

With essentially no convection impacting Oklahoma or far southern Kansas at any point in the simulation prior to initiation of the severe storms on the afternoon of 19 May, FLSM ensemble members behaved very similarly, both in the mesoscale environment and in terms of the location and morphology of the initial storms. There were differences in timing of initiation, however. Members MYJ-T, MYJ-MY, YSU-WD, MYNN-F, and to a lesser extent, YSU-MO initiated convection entirely in southern Kansas between 1900 and 2000 UTC. The convection initiated right along the dryline in a SSW-NNE oriented broken line from Sumner to Marion Counties, Kansas, a few tens of kilometers east of where observed convection developed. Initiation of convection occurred in the remaining members by 2100, also with a slight easterly bias (Fig. 5.24). A solid line of higher reflectivity was already present in some members by that time, matching the observed convective morphology. By 2100, the first members to develop storms in Oklahoma, MYJ-T and MYJ-MY, had done so in essentially identical locations in central and southern Oklahoma, followed within an hour by the remaining members. In general, as would be expected, development and early evolution of convection was very similar in members that used the same PBL scheme. Also, only in members MYJ-T and MYJ-MY did storms develop in southern Oklahoma. In the other members, storm coverage in Oklahoma was limited to the northern half of the state.

Similar to the control member, most other members of the FLSM ensemble did not properly handle the later evolution of convection. The ensemble members failed to continuously redevelop convection trailing behind the initial storms that formed in southern Kansas. However, at or after 0300 UTC 20 May, convection developed in northeast Oklahoma and southeast Kansas, behind the initial wave of storms, in all members except MYNN-MO, MYNN-F, and MYNN-N (not shown), similar to observed reflectivity. The forcing mechanism for this new development was a band of convergence resulting from accelerated southeasterly flow from the cold pool of the first wave of storms colliding with west and northwest winds behind the advancing dryline (combining with a weak cold front) in a region with 1000-2000 J kg^{-1} of MLCAPE and less than 100 J kg^{-1} of MLCIN. It should be noted that this evolution requires dryline-forced storms to have moved away from the boundary (section 5.6.2).

In the early afternoon of 20 May, storms first developed in members MYJ-T and MYJ-MY; they developed along and south of the Red River in Texas around 1900 UTC. Development of storms in Oklahoma occurred latest (around 2100) in members MYNN-MO, MYNN-F, and MYNN-N (Fig. 5.25). Storms developed at least one hour late and around 100 km too far east in each member. While a broken line of storms developed from central Texas northeastward into central Missouri in all members, the coverage of storms along the dryline and front varied among the members, with differences again appearing to be clustered by PBL scheme (not shown). Members MYJ-T and MYJ-MY contained three or four large clusters of storms with little else along the boundaries. Members MYNN-MO, YSU-WD, YSU-T, and YSU-MO, on the other hand, developed a configuration featuring generally uniform storm sizes nearly

equally spaced along the boundaries. Members ACM-WS and ACM-MY also had more uniform coverage of storms along the boundaries, but convection tended to fill in the gaps along the line as the simulation progressed. By 0000 UTC 21 May – the end of the forecast – these members forecast a nearly filled squall line extending northeastward from central Texas into western Illinois, whereas in the other members, there were large gaps in storm coverage, but the areal extent of storms was similar.

5.6.3.2 Mesoscale environment prior to convection initiation

The forecast in members using the same PBL scheme evolved nearly identically before initiation of storms on 19 May. There were only very small differences in dryline placement even among members using different PBL schemes, however. Point soundings east of the dryline also show the strong dependence on PBL scheme (Fig. 5.26). The vertical profile includes a very moist, yet capped environment was present in the warm sector with an elevated mixed layer atop the growing PBL, resulting in extreme instability, but also strong capping. The lifting and removal of this cap by large scale ascent can clearly be seen at later forecast hours (not shown). PBL depth ordering among members was similar to that in case #1, with the ACM2 scheme producing the deepest PBLs, the MYJ scheme producing the shallowest PBLs, and the other two schemes in between. Temperature in the PBL was very similar among ensemble members (Figs. 5.27c,d), leaving moisture content as the largest difference between members. Members using the MYJ scheme tended to be the most moist in both air masses (Figs. 5.27e,f) while members using the ACM2 scheme were the driest. Members using the MYNN and YSU schemes had very similar mixing ratios. These behaviors are consistent with those from case study #1. It may be that in this case study,

the stronger synoptic scale forcing was able to overwhelm any sensitivity to PBL scheme and force a relatively consistent dryline position among the members.

There was more variability in the heat and moisture fields east of the dryline/front in the afternoon of 20 May, owing to differences in convective evolution the previous evening. However, the same tendencies were still present.

Thermodynamic fields also clustered by PBL scheme. Members MYJ-T and MYJ-MY had the highest MLCAPE near the dryline, forecasting a narrow band of over 3500 J kg^{-1} in northern Oklahoma by 1900 UTC 19 May, whereas members ACM-WS and ACM-MY forecast less than 3000 J kg^{-1} in the same area. The remaining members had between 3000 and 3500 J kg^{-1} MLCAPE (Fig. 5.28). All members had less than 10 J kg^{-1} MLCIN in the area by 1900 UTC. It should be noted that, according to a RUC analysis, the observed MLCAPE was over 4000 J kg^{-1} in northern Oklahoma at the time, so all members underforecast instability, some by more than 1000 J kg^{-1}. It seems reasonable, then, that convection initiated first in MYJ-T and MYJ-MY since they were the most moist and had the highest CAPE and lowest LFCs, assuming forced ascent was similar among the members (of which there is little reason to suspect otherwise), and vice versa for ACM-WS and ACM-MY.

At 1800 UTC 20 May, all FLSM members depicted a weakly active cold front extending northeastward from its intersection with the dryline around Oklahoma City, Oklahoma, with the dryline extending SSW from there. The location of the cold front differed by 50-100 km among members and there were differences in the structure of the moisture gradient along it and the near-surface winds behind it. While most members had light W or SW winds immediately behind the front, in member MYJ-MY

there were light NE winds behind the front in Oklahoma. Also in member MYJ-MY, the front bulged northwestward coincident with a divergent near-surface wind pattern associated with very light precipitation falling along the front around that time. Members MYJ-T and MYJ-MY also had the highest CAPE among the FLSM ensemble members across Oklahoma at 1800 UTC.

The location of the dryline and the moisture gradient was very similar among the FLSM ensemble members. The gradient was diffuse, with mixing ratio gradually decreasing from around 14-15 g kg^{-1} in the warm sector to 3-4 g kg^{-1} over a 100-km distance. Near-surface winds also gradually veered from S or SSW to SW or WSW over that same distance. In member MYJ-MY the winds were backed around 10° compared to the other members, which was the most extreme difference in wind direction in the FLSM ensemble.

The most noticeable diversity at 1800 was with the location of the portion of the cold front that extended into the dry air behind the dryline in Oklahoma. A west-east oriented thermodynamic boundary with an associated wind shift was present, but in a much different location and shape in each member (not shown). In member YSU-T, the boundary was oriented perfectly west-east and extended from Dewey to Logan County. In member MYNN-N, on the other hand, the boundary had surged southward, arcing from Beckham County southeastward towards Kiowa and Comanche Counties and then northeastward across Caddo and Grady Counties before joining the dryline. Member MYJ-MY had a similar placement as member MYNN-N, but with the boundary displaced northward a few tens of kilometers. Other members were within the two

extremes presented above. While these differences were substantial, they did not appear to play a role in the location or timing of development of convection a few hours later.

5.6.4 Sensitivity in the LSMO ensemble

5.6.4.1 Reflectivity evolution

Initiation of storms in south central Kansas occurred between 2000 and 2100 UTC 19 May in each member (Fig. 5.29). In members Z01, Z25, Z50, MP1, and MP5 the morphology was a small cluster of cells with one dominant cell located in Butler County in each of those members. In members MP2, MP3, and MP4 only one large cell was present (but with a small secondary cell developing), also in Butler County. Only a small storm or two was present in members RUC and PX, located in nearly identical locations across Butler and Cowley Counties. A storm also developed in central Oklahoma in members MP1 and MP5 by 2100. Storms developed in central Oklahoma in the remaining members over the next two hours, with initiation latest in members RUC and PX.

The convective morphology over the next few hours was similar in each member – that of a broken line of cells extending from far southeastern Nebraska and southwestern Iowa through central and eastern Oklahoma – with varying degrees of filling along the line (not shown). Only member PX exhibited any tendency for immediate linear growth with the storms in Kansas. Members MP2, MP3, and PX had a mostly filled line by 0100 UTC 20 May, whereas larger gaps remained in the other members. In the extreme case – member RUC – only two storms (both supercells according to HMUH) were forecast in Oklahoma before 0100. Additionally, convection also tended to develop sequentially down the dryline in most members, although with

different frequencies. In members MP1, MP2, MP5, and PX, three to four additional cells developed along the dryline after the initial cell and south of where it developed. In member Z50, no additional cells developed in central Oklahoma after the first one.

Subsequent reflectivity evolution differed from that in the FLSM ensemble. While LSMO ensemble members also generally failed to correctly evolve the convection into the late evening, in only member RUC did secondary development occur across southeastern Kansas. This was due to the lack of storm coverage in this member across northeastern Oklahoma and southeastern Kansas. In member RUC, the air mass recovered behind the initial storms that developed in Kansas and a substantial amount of instability remained into the evening. Forcing was provided by a combination of the dryline and an eastward advancing OFB from post-dryline storms that developed farther to the west, in central Kansas, in the late afternoon. This evolution was different from that in the other members, where convergence was present in a similar location, but there was insufficient instability to support new convection. As a result, convection was absent from all but the eastern most column of counties in Oklahoma by 0500, except in member MP5, which contained a weak and dying cell across east-central Oklahoma (not shown).

In the early afternoon of 20 May, storms first developed in members MP1, MP2, and RUC. In member MP1, the first cell developed in central Texas, whereas in member MP2, two storms developed in Oklahoma, one in the northeastern corner of the state and one in the south-central portion. In member RUC, storms developed in both Texas and Oklahoma (Fig. 5.30). In each member, initiation occurred along the combined dryline/front. Storms formed in the remaining members over the next two

hours, with initiation latest in member PX. The location of the first storm varied along the combined dryline/front. A small storm developed in far northeastern Oklahoma or far southeastern Kansas in each member, but development was preferred across central Texas in members Z01, Z25, Z50, and MP1, and straddling the Red River in members MP5 and RUC. Initial convection expanded more quickly in member MP2, but by 2200 UTC each member had similar storm coverage (not shown). Convective morphology was a broken line of cells with varying degrees of coverage along the line. By the end of the forecast, there were large breaks in the line in members MP1 and RUC, whereas the line remained almost completely filled in members MP2 and PX. Overall, however, storm coverage and morphology among LSMO members was less diverse than among FLSM ensemble members. Each member contained rotating storms in various locations along the line (according to HMUH, not shown), but HMUH did not reach larger values until 2200 or later, indicating a late bias in the forecast of strongly rotating storms. If the eastward bias of the forcing boundaries was offset, then some members forecast rotating storms near Moore, Oklahoma.

5.6.4.2 Mesoscale environment

Dryline location on the afternoon of 19 May in the LSMO ensemble was approximately as variable as that in the FLSM ensemble, with differences of 20 km or less among members (not shown). However, there was more diversity in the moisture content both throughout the warm sector and immediately west of the dryline in Oklahoma. West of the dryline, members RUC, MP4, and PX were generally the coolest and most moist, and members MP4 and PX had the shallowest PBLs (Figs. 5.31a,c,e). Member PX was overall the most moist owing to very high moisture flux

(Fig. 5.32c) and rapid drying in member RUC during the afternoon. Members MP1 and MP5, and to some extent MP3, were overall warmer and drier with deeper PBLs, owing to larger sensible heat flux and reduced moisture flux, especially in members MP1 and MP5 (Figs. 5.32a,c). This behavior is the same as in case #1. The reason for the drying in member RUC is related to dramatic PBL growth in that member (Fig. 5.31a). There is a frequency difference or phase offset in the diurnal cycle of sensible heat and moisture flux in the RUC LSM compared to the other LSMs. As can be seen in Fig. 5.32, moisture flux in member RUC reached approximately the same maximum value as in member PX, but two hours earlier (moisture flux peaked at about the same time in all other members), and decreased sharply afterward, becoming the member with the lowest moisture flux west of the dryline by about 2200 UTC. Sensible heat flux followed a complementary behavior by reaching a maximum comparable to that of members MP1 and MP5, but one hour later (and later than all other LSMO ensemble members) with a sharp increase beginning shortly after initialization (at which point it had the second smallest sensible heat flux). Between 2100 and 0000 UTC, member RUC had the highest sensible heat flux, exceeding some members by as much as 200 W m^{-2}. This behavior favors slow growth and a tendency to keep PBL averaged moisture at a higher level than other members early in the day, followed by rapid growth, heating, and drying towards early evening, which is consistent with the time tendency seen in point soundings (not shown). Strangely enough, this behavior was generally not present east of the dryline except on the second day of the forecast. It was seen in case #1, however, but also only west of the dryline.

East of the dryline, member RUC was much more moist throughout the depth of the PBL with mixing ratios 1.5-2.0 g kg^{-1} higher than the other members (Fig. 5.31f), the result of large moisture flux (Fig. 5.32d). The PBL was also cooler and shallower (Figs 31b,d). Members Z01 and PX were also more moist than average, although moisture dipped in member PX during the late afternoon, perhaps due to an increase in sensible heat flux (Fig. 5.32b). On the other hand, members MP1 and MP5 had much warmer PBLs, which were also noticeably drier and deeper as well, similar to the first case. Members MP2 and MP3 were also somewhat more dry and warm, but to a lesser extent.

Consistent with being more moist east of the dryline, MLCAPE was also highest in the vicinity of the dryline in member RUC with maximum values exceeding 4000 J kg^{-1} in north-central Oklahoma by 2200 UTC (Fig. 5.33). In members MP1 and MP5, maximum MLCAPE barely exceeded 3000 J kg^{-1} in the same area. In most of the other members, maximum MLCAPE did not exceed 3500 J kg^{-1}; if it did, it was only at a small number of grid points. In general, this level of variability is similar to or slightly less than that seen in the FLSM ensemble (not shown). |MLCIN| decreased to less than 10 J kg^{-1} in south-central Kansas in all members by 1800-1900 UTC, and in Oklahoma by 2000 (not shown), suggesting nearly complete removal of any resistance to initiation by early afternoon across the entire region. Despite being more moist, CIN was not appreciably lower in member RUC than any other member, although the region of near zero CIN extended further east of the dryline in that member than in other members. This behavior in the thermodynamic fields is consistent with differences in PBL heights among the members. Members MP1 and MP5 – having deeper, drier, and warmer PBLs

– still mixed enough to remove CIN at the expense of reduced theta-e and hence, CAPE. The PBL was shallower and cooler, but more moist in member RUC, so theta-e and hence CAPE was higher, but the structure of the capping layer above the PBL was similar to that in the former members, thus resulting in similar levels of CIN.

Differences in the timing of initiation are likely related to meso-gamma-scale variations that are hard to pin down on a 4 km grid, as discussed with case #1. However, in each member, initiation in southern Kansas was preceded by a local maximum in forcing, whether integrated convergence or vertical motion near the top of the PBL. It should be noted that this forcing was not present in members RUC and PX at the same hour it was present in the other members, as initiation was somewhat delayed in members RUC and PX compared to the other members. But it was present the next hour, just as initiation was occurring. Initiation in central Oklahoma was not preceded by any particularly strong forcing maximum. Rather, there was a broad area of convergence along the dryline in central Oklahoma in each member.

There was considerably more variability in the mesoscale environment by early afternoon on 20 May, consistent with the longer lead time and for more time for differences in the previous day's convection to cause divergence of ensemble members. For instance, MLCAPE was more variable along the fronts in Oklahoma and Texas than it was along the dryline the day before (not shown). While there were similar differences in the magnitude of maximum MLCAPE, the member with the largest value and the location of the maximum were in much different locations among ensemble members, ranging from central Texas in members Z01 and PX, for example, to areas of eastern Oklahoma in members MP2 and MP4 (Fig. 5.34). However, |MLCIN| decreased

to near-zero values just before initiation in each member and over a similar area, similar to the day before. Additionally, the variability in thermodynamics along and just east of the boundaries was larger than that in the FLSM ensemble (not shown). This indicates that a combination of direct perturbations to the LSM component and the interaction between LSM perturbations and convective evolution were sufficient to cause more diversity in 30+-hour forecasts than perturbations to microphysics and PBL parameterizations. This further suggests that it may take longer for LSM perturbations to generate large ensemble diversity, but once a sufficient amount of time has passed, the generated diversity can be substantial.

5.7 Implications for storm-scale ensemble design

5.7.1 Ensemble diversity

Overall ensemble forecast quality can only be approximated here due to the small sample size – only a few cases were examined. Notwithstanding, the effectiveness of the LSM perturbations compared to the other physics combinations can still be evaluated. Since many limited-area, convection-allowing ensembles not incorporating lateral boundary perturbations tend to be underdispersive (Nutter et al. 2004 provides some insights), increased spread is generally preferable, as it indicates a greater tendency for the ensemble dispersion to reflect the underlying forecast uncertainty and capture the full range of variability in the verifying atmospheric state.

Ensemble spread in the two cases followed a diurnal cycle, with some departures from that behavior. For example, mixing ratio spread at 925 and 850 hPa in case #1 increased approximately linearly through almost the entire forecast before leveling off somewhat (Fig. 5.35). This behavior was not seen at other levels or in other

fields. I speculate that it could be related to the nocturnal boundary layer wind maximum. There were also differences in the rate of spread growth between the cases. In case #1, spread increased rapidly during the first six forecast hours before leveling off, then decreasing slightly (following the diurnal cycle), whereas in case #2, spread growth within the first six forecast hours was slower, but then accelerated during the subsequent six hours before following the diurnal cycle. This may be due to different initialization times and where initialization lay within the diurnal cycle, and also due to the meteorological scenario. The development of an MCS during the spin up period in case #1 may be responsible for immediate rapid spread growth. In case #2, the sensible weather was comparatively quiescent during the early portion of the forecast, but became very active during the severe weather outbreak which began around forecast hour 9. Maximum spread during any portion of the forecast was slightly larger in case #1 than in case #2, likely reflecting overall larger uncertainty in the evolution of the atmosphere in case #1, which was more mesoscale driven, compared to case #2, which was strongly forced on the large scale.

As seen for both cases in Fig. 5.35, the MLSM ensemble had larger spread than the FLSM ensemble at nearly all forecast hours and in all fields, but especially in fields closer to the surface. This includes not only horizontal winds, temperature, and mixing ratio, but also stability indices such as surface-based and mixed-layer CAPE and CIN (not shown), although the difference in CIN spread between the FLSM and MLSM ensembles was very small. It was hypothesized early in the research that members that used the RUC and PX LSMs may be outliers since they are formulated more differently than the Noah and Noah-MP LSMs, which are formulated more similarly. Therefore,

spread from nine-member sub-ensembles with each member of each ensemble removed was also examined for robustness in the signal. Results suggest that members RUC and PX were indeed outliers since sub-ensemble spread with those members removed was much lower than with other members removed (Fig. 5.36). However, even with these outlier members accounted for, the MLSM ensemble spread was still generally larger than that of the FLSM ensemble especially for temperature and mixing ratio, indicating the LSM perturbations added diversity to the ensemble by perturbing the model in a dimension not perturbed by microphysics and PBL variations. It is encouraging to see spread of the horizontal wind components in the MLSM ensemble was also somewhat larger than in the FLSM ensemble, likely the result of feedbacks between the thermodynamic and kinematic aspects of the model. It is also encouraging to note that temperature and mixing ratio spreads from the LSMO and FLSM ensembles were comparable at a subset of forecast hours, indicating that LSM perturbations alone were almost as effective at generating ensemble diversity as a combination of microphysics and PBL physics perturbations.

5.7.2 Precipitation verification

Increased ensemble spread is not desirable if it is accompanied by a degraded forecast. Probabilistic forecasts of one-hour accumulated precipitation were verified using Stage IV precipitation analyses as verifying data. The data were regridded to match the model grid using bilinear interpolation. PQPFs were constructed using a 50 km circular neighborhood (similar to Chapters 3 and 4). The results were fairly similar among a variety of metrics, so only the fractions Brier score (Schwartz et al. 2010) is discussed (Fig. 5.37). There was very little difference in forecast performance between

the FLSM and MLSM ensembles in each case. At a small number of forecast hours the MLSM ensemble was slightly better, but the opposite was true at other forecast hours. The LSMO ensemble performed more poorly than either the FLSM or MLSM ensembles, although only slightly so during the periods in each case where convection was active. This is broadly consistent with the spread comparison between the FLSM and LSMO ensembles in that microphysics and PBL perturbations were slightly more effective in making skillful PQPFs than just LSM perturbations. However, some member forecasts from the LSMO ensemble were better than those from either of the FLSM or MLSM ensembles (not shown). In general, individual member verification was most tightly clustered in the LSMO ensemble and least so in the MLSM ensemble. While certain members had biases and scores that stood out from the rest of the ensemble for specific metrics, thresholds, or forecast hours, no member was consistently an outlier as either a more or less skillful member.

5.7.3 Perturbation methodology

Three strategies for generating LSM perturbations were used in this study: multiple separate LSMs (hereafter multi-LSM), altering a set of similar parameters or processes within the Noah-MP LSM (hereafter perturbed parameter), and varying one specific parameter with documented uncertainty (the CZIL parameter in the Noah members). Since only three members contained CZIL diversity, that particular method was not compared to the other two. However, the results suggest there is value in perturbing this parameter, although better results may be achieved in future work by using a more disparate set of values. Four separate LSMs were used in this study, and they comprise a multi-LSM sub-ensemble. Five members used the Noah-MP LSM with

changes only to the formulation of a set of physical processes and comprise a perturbed parameter sub-ensemble. It is natural to consider whether either approach individually is better than the combination of approaches. Ensemble spread and precipitation verification were used to answer the question of which perturbation method is better.

The results were consistent between the two cases: the multi-LSM approach resulted in larger spread for temperature and moisture, and in some locations for the horizontal wind components, too (Fig. 5.38). The signal was more apparent at lower levels. This is particularly interesting since the multi-LSM sub-ensemble only included four members, a smaller size than the perturbed parameter sub-ensemble, and ensemble size in the range of a few to 10 members was shown to significantly impact ensemble spread in Chapter 3. Despite having fewer members, the multi-LSM sub-ensemble, using separate LSMs with more diversity in soil structure and formulations of physical processes, contained greater diversity than changing the formulation of a few processes within a single LSM. However, there was notable uncertainty when checking the statistical robustness of the results by examining spread from sub-ensembles with various members removed. In particular, removing either the RUC or PX LSM from the multi-LSM sub-ensemble caused a large decrease in spread (not shown), commonly to the point where it became smaller than that of the perturbed parameter sub-ensembles. Since these LSMs were only used in one member each, and since 8 of the 10 ensemble members used Noah LSMs, the experimental design may have forced this result. At the same time, however, there is support for the utility of separate LSMs to generate ensemble diversity rather than perturbing parameters within a single LSM, a result also found in Chapter 3. However, the caveat to this is the same as that in Chapter 3 in that it

is typically more difficult to develop and maintain a suite of separate LSMs than it is to perturb parameters within a single LSM that can be updated with less effort.

Verification of one-hour accumulated precipitation was also consistent between the cases. The scores were very similar between the sub-ensembles. When they differed, the result is as follows. Fractions Brier scores were better in the multi-LSM sub-ensemble than in the perturbed parameter sub-ensemble in the second half of the forecast, whereas they were better for the perturbed parameter sub-ensemble at the 2.54- and 12.7-mm thresholds in the first half of the forecast (Fig. 5.39). The area under the ROC curve also suggests the multi-LSM sub-ensemble was more skillful during the second half of the second case, but not of the first case, where the sub-ensemble with better scores alternated in time (not shown).

5.8 Summary and conclusions

The sensitivity of convective-scale forecasts to model physics perturbations of the WRF model, with an emphasis on perturbations to the LSM component, were evaluated for two cases involving intense deep moist convection that produced severe weather in the Plains region of the U.S. One case was strongly forced by synoptic scale processes and included a major severe weather outbreak including multiple tornadic supercells. The other case was weakly forced on the synoptic scale and was more dependent on mesoscale features (i.e., an outflow boundary left by an MCS from the previous night). Three ensembles were configured, each with 10 members, perturbing different combinations of WRF model physics components. The case studies focused on differences in sensitivity of the forecasts of convection and precipitation between two of the ensembles, the FLSM ensemble in which the same LSM was used with varying

microphysics and PBL parameterizations, and the LSMO ensemble which included LSM perturbations but used the same microphysics and PBL scheme. There were multiple purposes of conducting this experiment. One purpose was to determine if adding LSM perturbations to other physics perturbations already used in experimental storm-scale ensembles improved ensemble spread and probabilistic forecasts. The LSM component contains uncertainties that have not been sampled in prior experimental storm-scale ensemble forecast systems. Another purpose was to document the variability that LSM perturbations alone can generate in a convective-scale NWP forecast. A third purpose was to document the performance and behavior of a new LSM, the multiparameterization-Noah (Noah-MP), which has not yet been used in experimental warm- season convective-scale forecasts.

The case studies suggest LSM perturbations may take longer to grow before generating sufficient ensemble diversity to meaningfully alter the mesoscale details of the ensemble forecast compared to microphysics and PBL perturbations. However, given sufficient integration time, the LSM perturbations can grow and cause ensemble diversity to exceed that generated by only microphysics and PBL parameterizations. Generally, LSM perturbations can cause storm scale differences to emerge even after 6-9 hours of forecast integration assuming the forecast period includes the daytime during which surface-atmosphere exchange processes are active. However, an ensemble in which only the LSM component is perturbed is not recommended.

Adding LSM perturbations to the FLSM ensemble to construct the MLSM ensemble had a positive impact on ensemble diversity and PQPFs. Ensemble spread in the MLSM ensemble was larger than that of the FLSM ensemble for many fields in the

lower atmosphere, with differences generally diminishing with ascent. This was the case even after accounting for possible outlier members in the MLSM ensemble, which included the members that used the RUC and PX LSMs. Verification of PQPFs also suggested a positive benefit of adding LSM perturbations, although the difference in scores between the FLSM and MLSM ensemble was small, impeding the ability to draw robust conclusions regarding the effectiveness of the added perturbations on precipitation forecasts.

Three methods were used to generate perturbations to the LSM component. One method was to use separate LSMs (multi-LSM). Four LSMs were used in this study: Noah, Noah-MP, RUC, and PX. The second method was to alter the namelist options controlling the formulation of various physical properties within the Noah-MP LSM (perturbed parameter). Five members were generated using this approach (one overlapping with the multi-LSM approach). The third method was to perturb the CZIL parameter in the Noah LSM. Three values were used (one overlaps with a member used in the multi-LSM approach). A separate test regarding which particular perturbation methodology – multi-LSM or perturbed parameter – was more effective at increasing ensemble diversity revealed that using separate LSMs may be preferable. However, there may be difficulties in developing and maintaining separate LSMs, a difficulty that would not be encountered using a perturbed parameter approach. No additional testing was performed comparing the perturbation of CZIL to the other two methods.

The Noah-MP LSM offered a convenient way to generate ensemble members by changing namelist options for parameters thought to be important for warm season forecasts. It should be noted that one particular option for calculating the surface

exchange coefficient in the Noah-MP LSM seemed flawed, however. The calculated exchange coefficients were too low, causing skin temperature, and subsequently, ground heat flux and soil temperature to be too high. This caused large error in soil temperatures and 2-m temperatures, which is not desirable in ensemble design, which seeks to address uncertainties rather than perturb a value or process to the point of being unreasonable. Interestingly, soil moisture and precipitation forecasts were not degraded in members using this particular option, somewhat neutralizing my opinion regarding this parameter.

It is important to stress that this was an exploratory study on LSM perturbations. The perturbation methodology, although based on literature review and convenience of available code in the WRF model, may have faults. It is not guaranteed that the particular set of perturbations applied in this study is the best such set of perturbations to use. It is certainly not the only possible set of perturbations. Future work should consider using other combinations of namelist options to generate ensemble members using the Noah-MP LSM (18 possible combinations are possible), as well as determine how sensitive each option actually is. Both prior research and the results herein suggest that convective-scale forecasts are not as sensitive to handling heat flux at the bottom of the soil (opt_tbot) as they are to calculating the surface exchange coefficient (opt_sfc). It may turn out that not all 18 combinations of Noah-MP options would contribute meaningfully to ensemble diversity. Finally, it may be necessary to perturb CZIL over a larger range, as literature review suggests a wider range of values may be necessary to capture all variability associated with that parameter. It may also be useful to include other methods of calculating the thermal roughness length, some of which do not

include CZIL. The important result from this work, however, is that adding a reasonable set of LSM perturbations appears to improve convective-scale ensemble forecast performance.

Figures

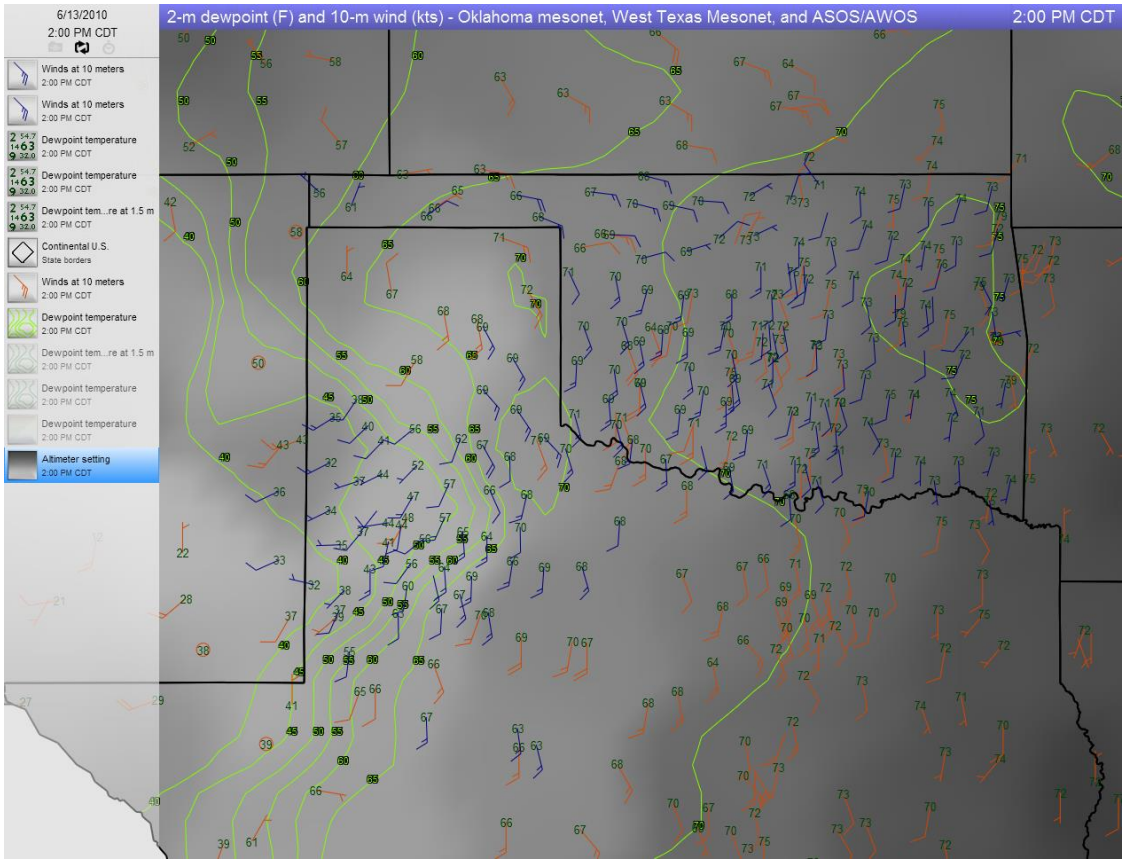

Fig. 5.1. Surface dewpoint (°F) and 10-m wind barbs (kt) valid 1900 UTC 13 June 2010. Green contours represent analyzed dewpoint contours, while gray shading represents mean sea-level pressure, with lighter shades representing lower pressure.

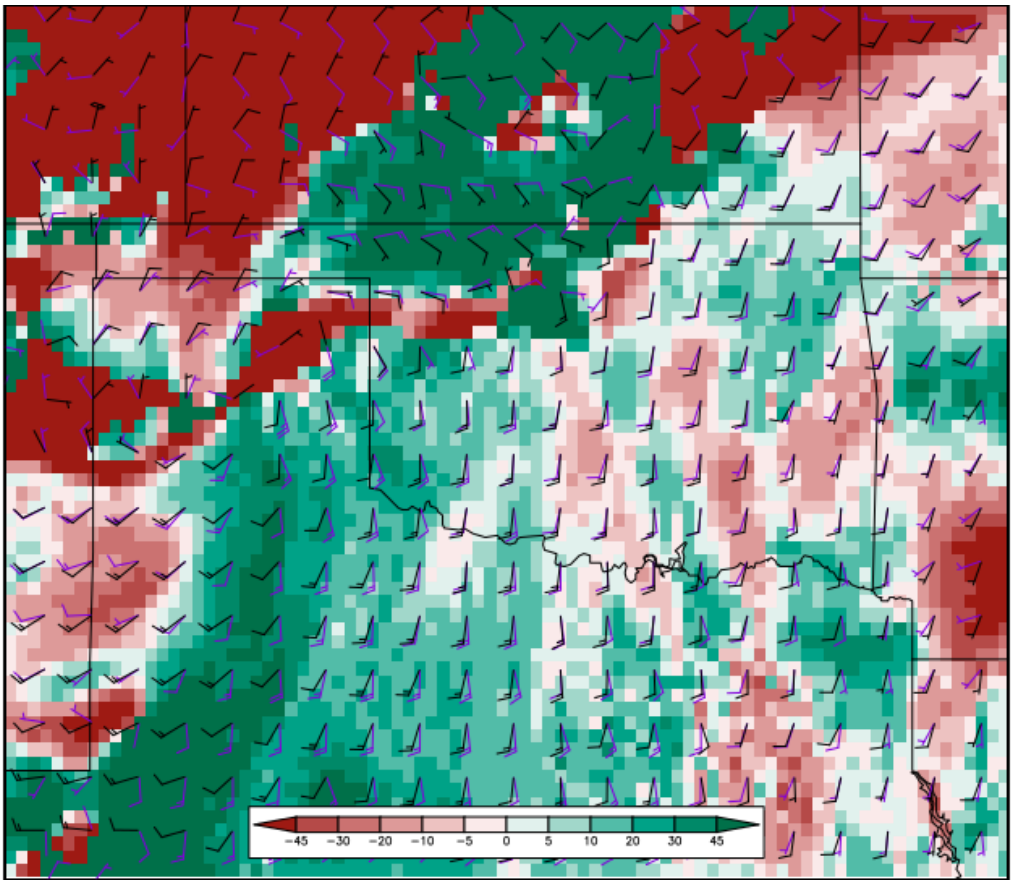

Fig. 5.2. Wind direction error (degrees, color shaded) between a RUC analysis and the control member forecast at 1800 UTC 13 June 2010. Green colors denote areas where the forecast wind directions are veered relative to the RUC analyzed winds. Forecast wind barbs (kts) are shown in black, whereas RUC analyzed wind barbs are in purple.

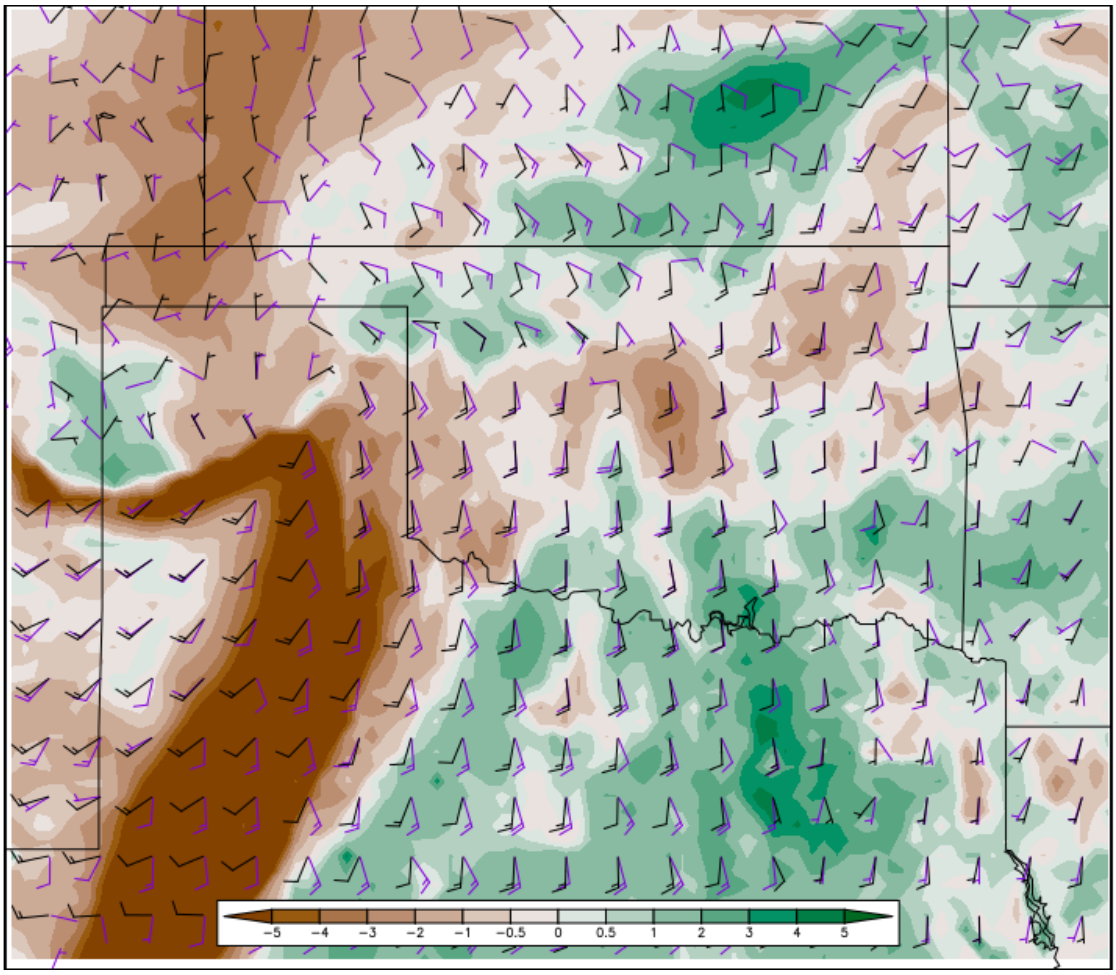

Fig. 5.3. 2-m mixing ratio differences (g kg^{-1}, color shaded) between the control forecast and a RUC analysis, valid at 2000 UTC 13 June 2010. Brown shades denote where the control forecast was drier than the RUC analysis. Forecast wind barbs (kts) are in black, whereas RUC analysis wind barbs are in purple.

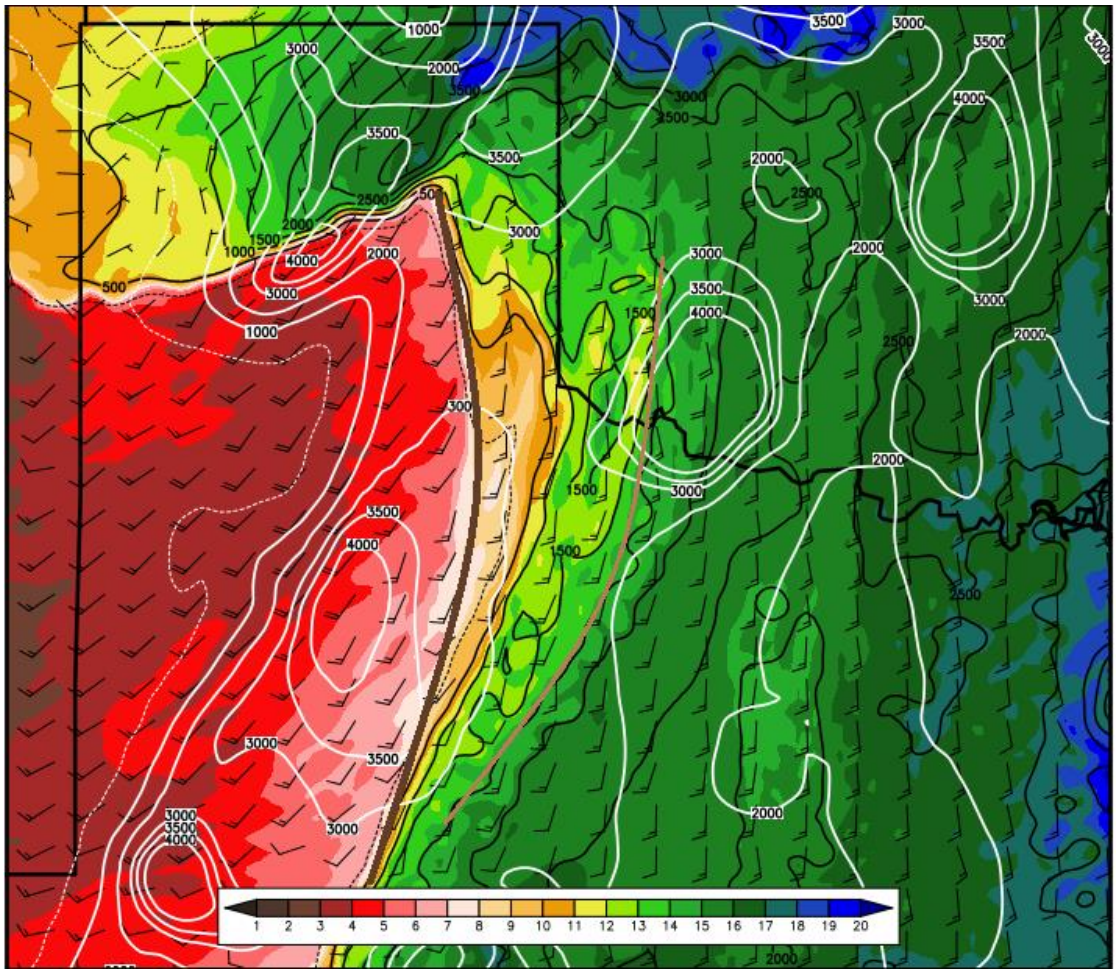

Fig. 5.4. Control member forecast mixing ratio (g kg^{-1}, color shaded) and wind barbs (kts) at the first model level with forecast SBCAPE (black) and RUC analyzed SBCAPE (white) contoured at 2100 UTC 13 June 2010. The 100-J kg^{-1} contour is dashed for both sources. The two moisture gradients composing the double dryline structure are indicated in brown arcs.

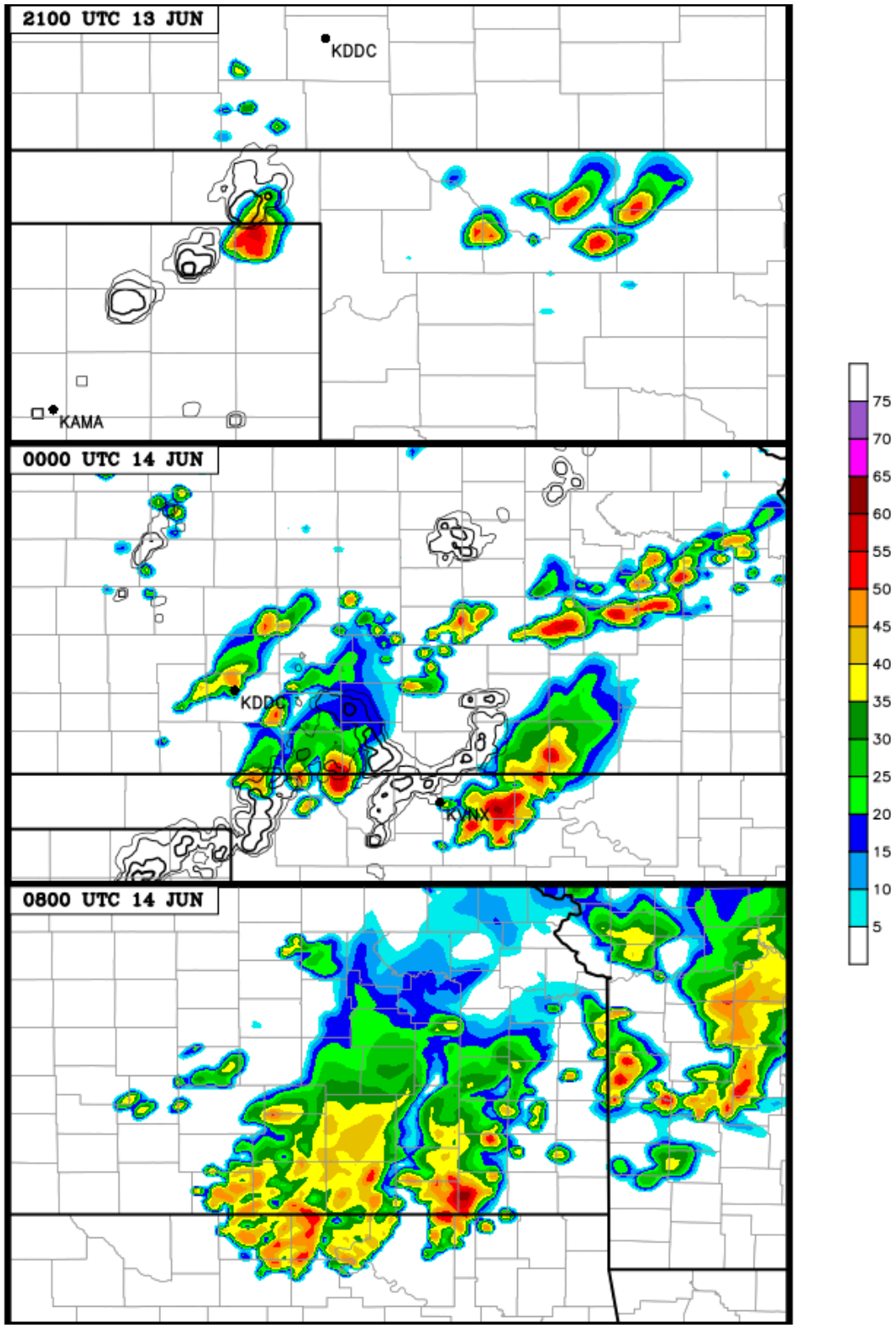

Fig. 5.5. Control member forecast composite reflectivity (dBZ) with observed reflectivity (contoured every 10 dBZ starting at 30 dBZ in progressively thicker black contours) from indicated radar sites, except in the bottom panel.

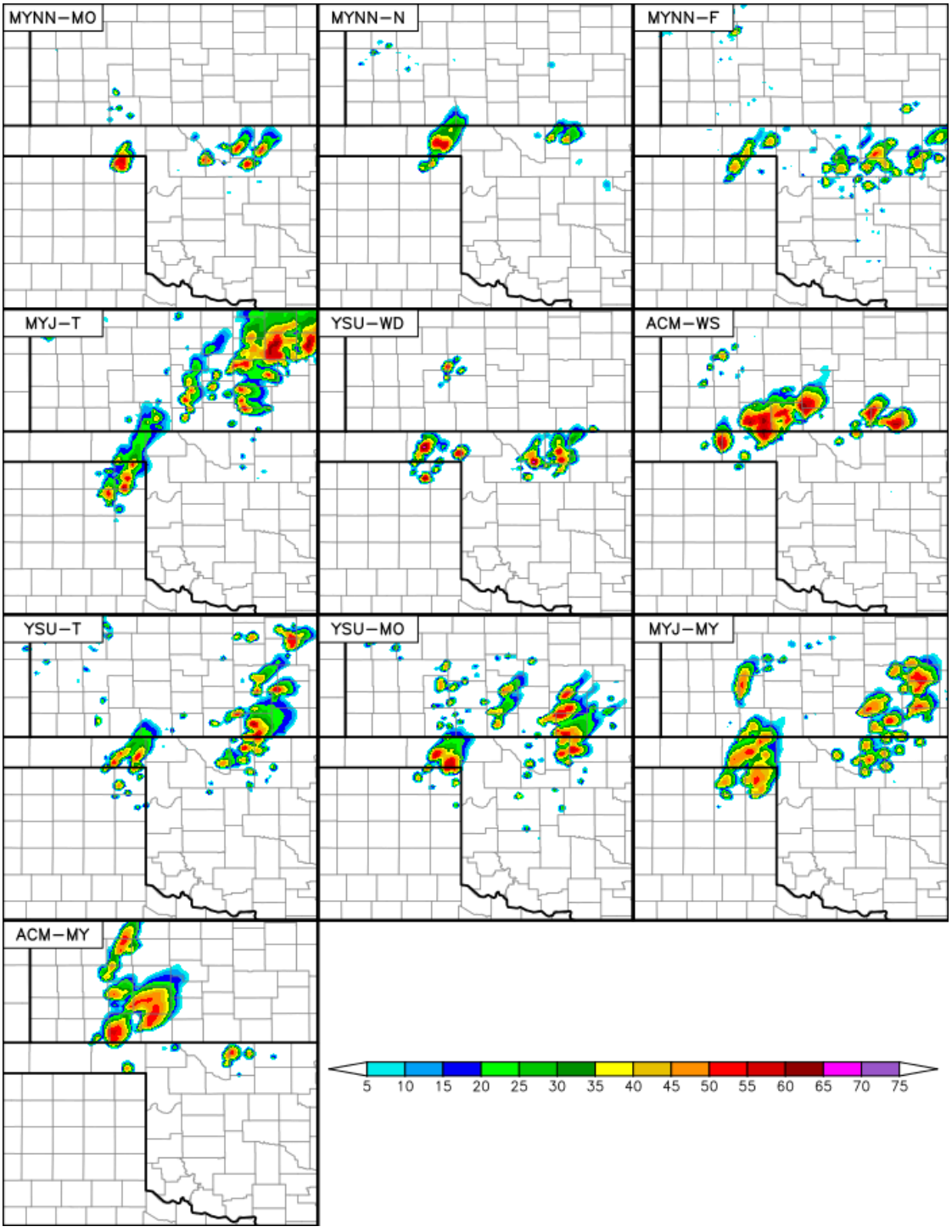

Fig. 5.6. Simulated composite reflectivity (dBZ) from FLSM ensemble members; a 21-hour forecast valid 2100 UTC 13 June 2010.

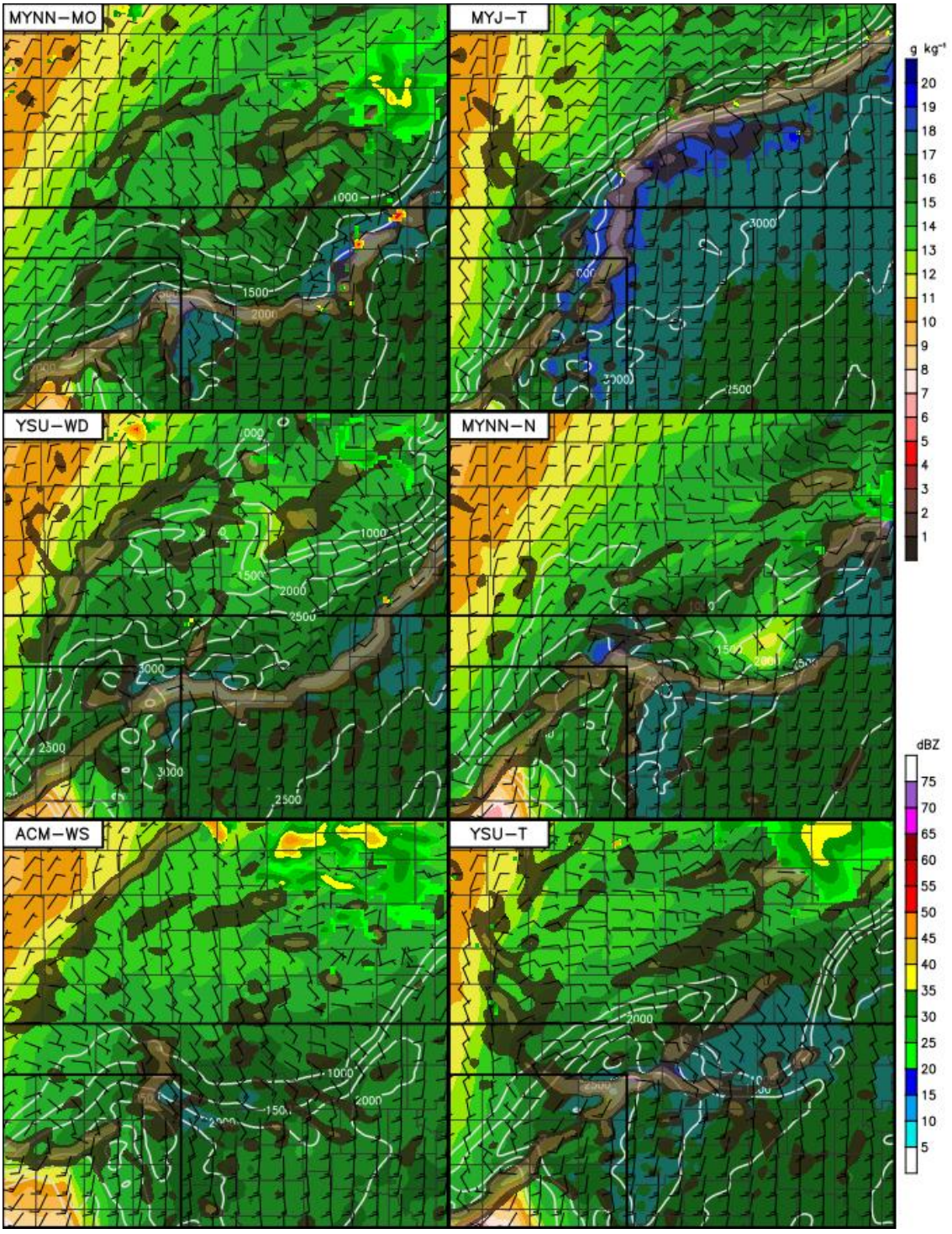

Fig. 5.7. Forecast mixing ratio (g kg^{-1}, color shaded), wind barbs (kts), and moisture convergence (transparent brown shades) at the first model level, MLCAPE (J kg^{-1}, white contours), and forecast reflectivity (dBZ, color shades above 20 dBZ) from select FLSM ensemble members at 1800 UTC 13 June 2010.

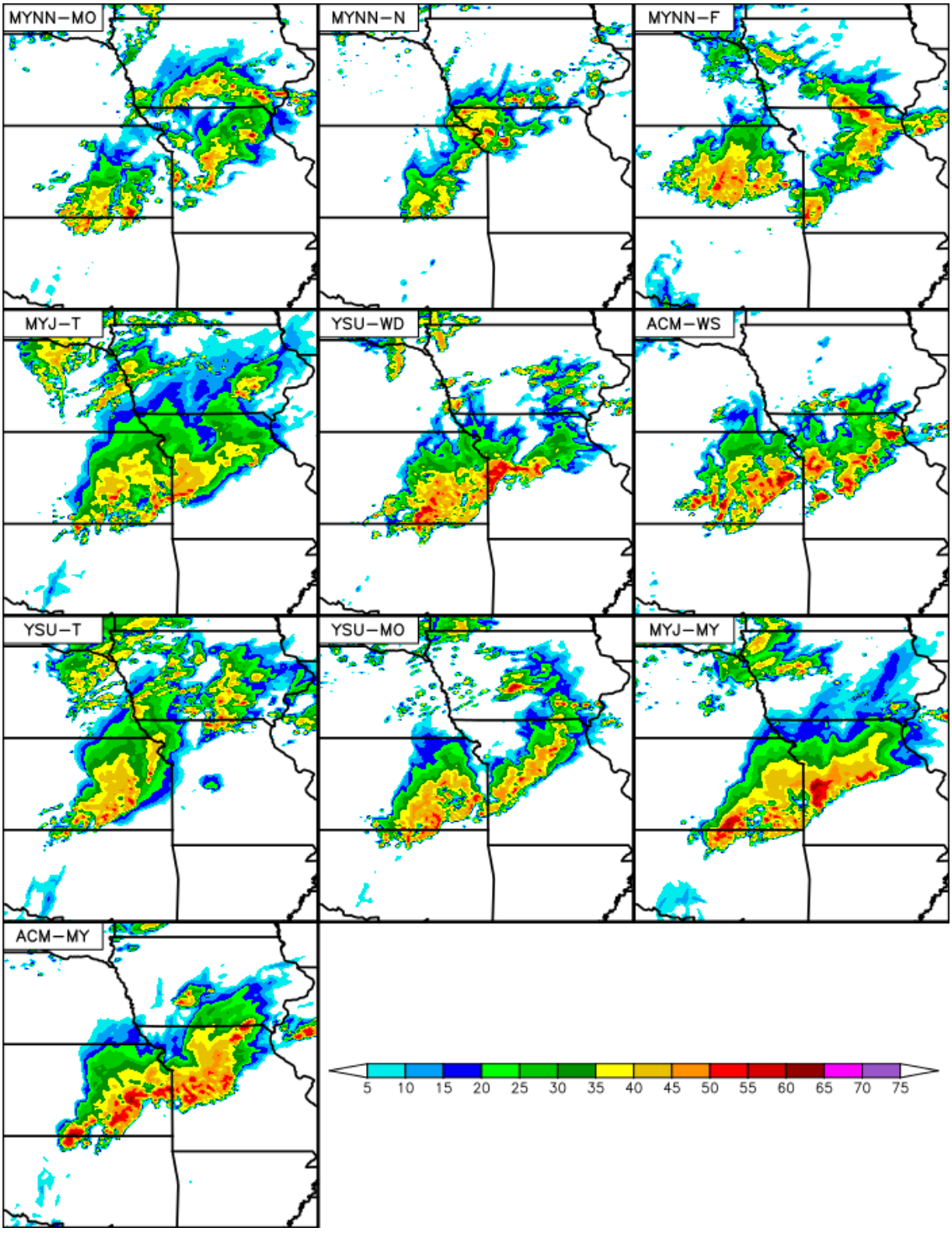

Fig. 5.8. As in Fig. 5.6 except for a 32-hour forecast valid 0800 UTC 14 June 2010.

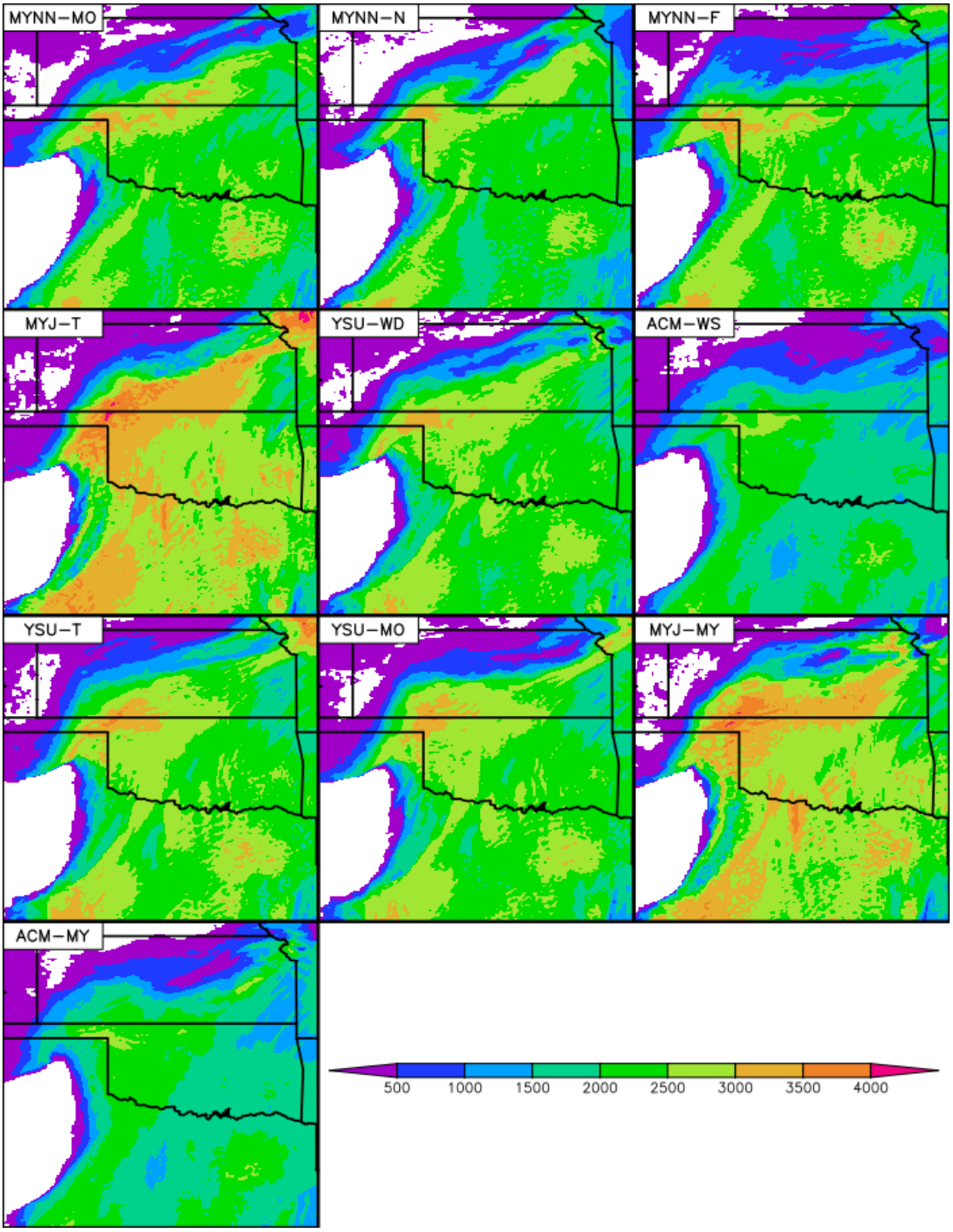

Fig. 5.9. Maximum MLCAPE (J kg^{-1}) between 1900 and 2100 UTC 13 June 2010 from FLSM ensemble members.

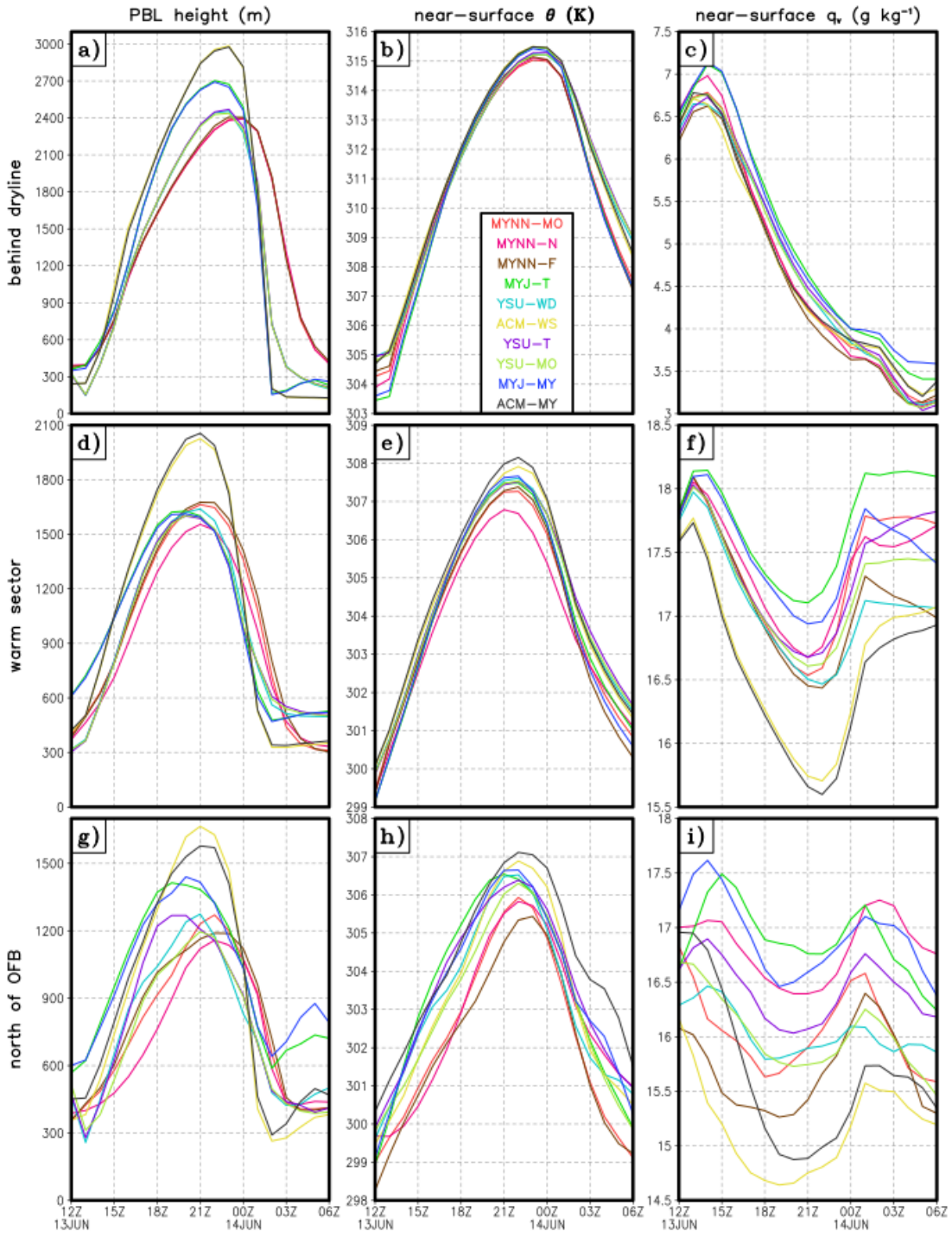

Fig. 5.10. Air mass averaged (a), (d), (g) PBL height, (b), (e), (h) potential temperature, and (c), (f), (i) mixing ratio (a)-(c) behind the dryline, (d)-(f) east of the dryline, and (g)-(i) north of the OFB for FLSM ensemble members.

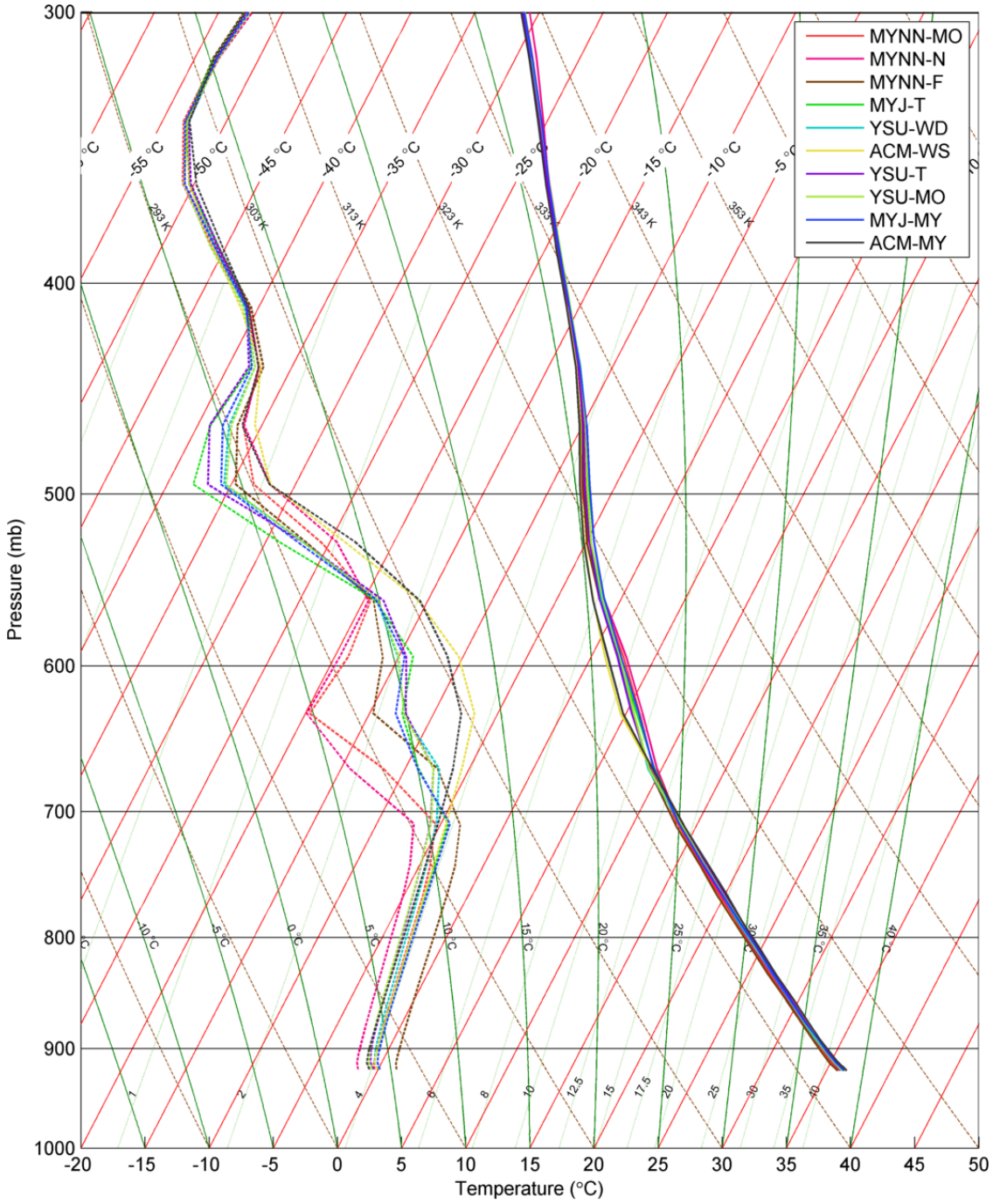

Fig. 5.11. Point forecast skew-T/log-P diagram for FLSM ensemble members at 2200 UTC 13 June 2010. The location is at (lat,lon) = (34.3°N,101.0°W).

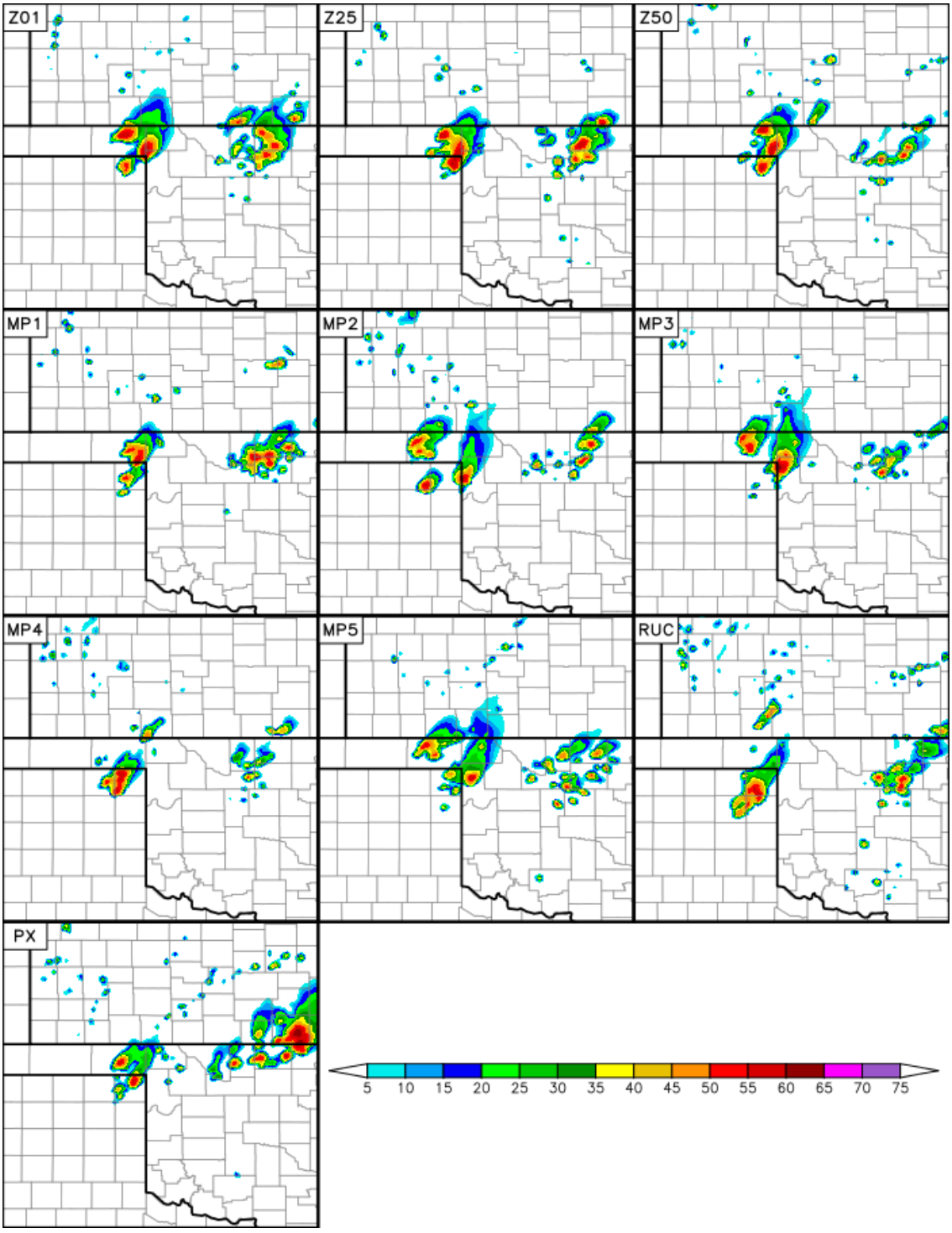

Fig. 5.12. Same as Fig. 5.6 except from the LSMO ensemble and valid at 2200 UTC 13 June 2010.

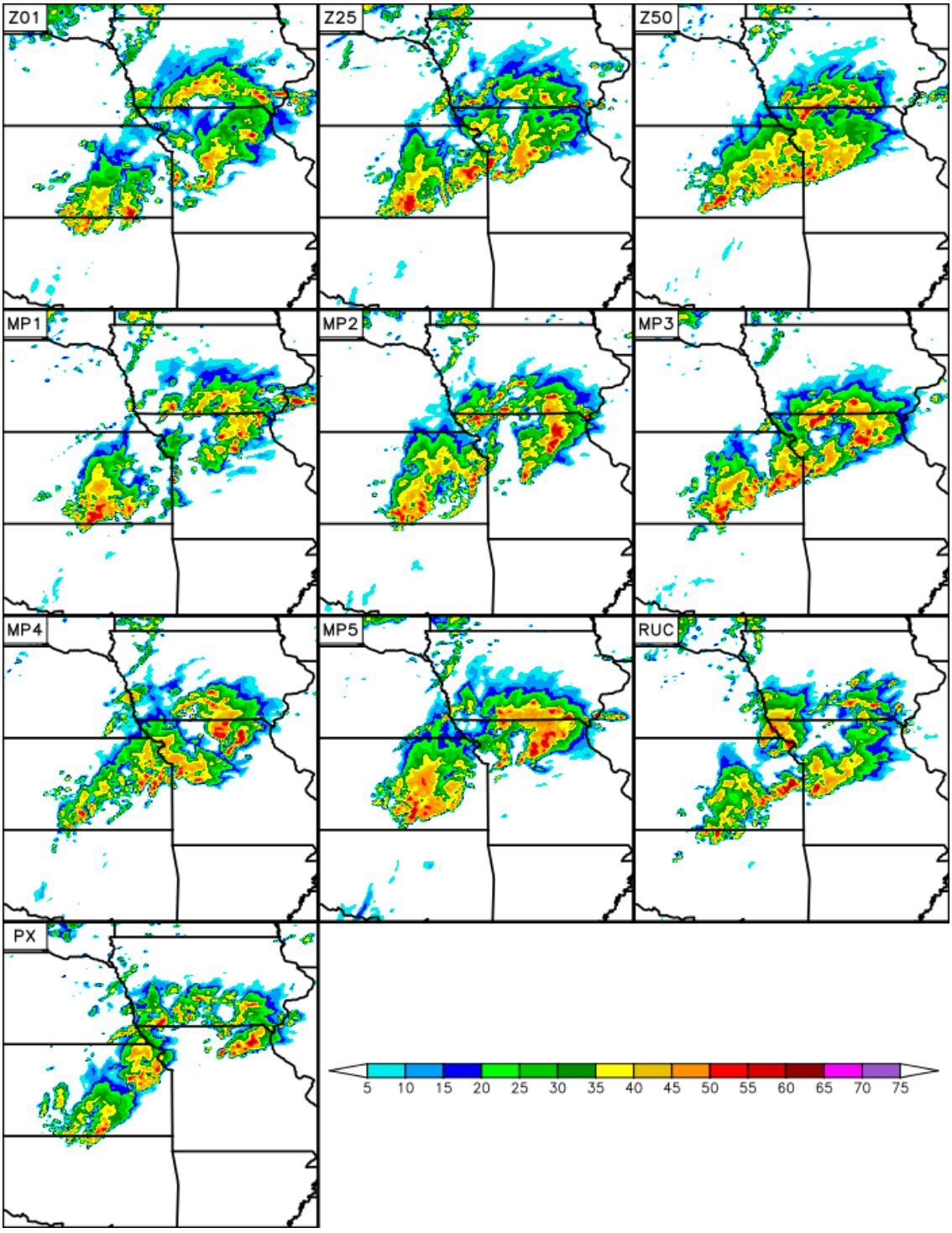

Fig. 5.13. Same as Fig. 5.8 except from the LSMO ensemble.

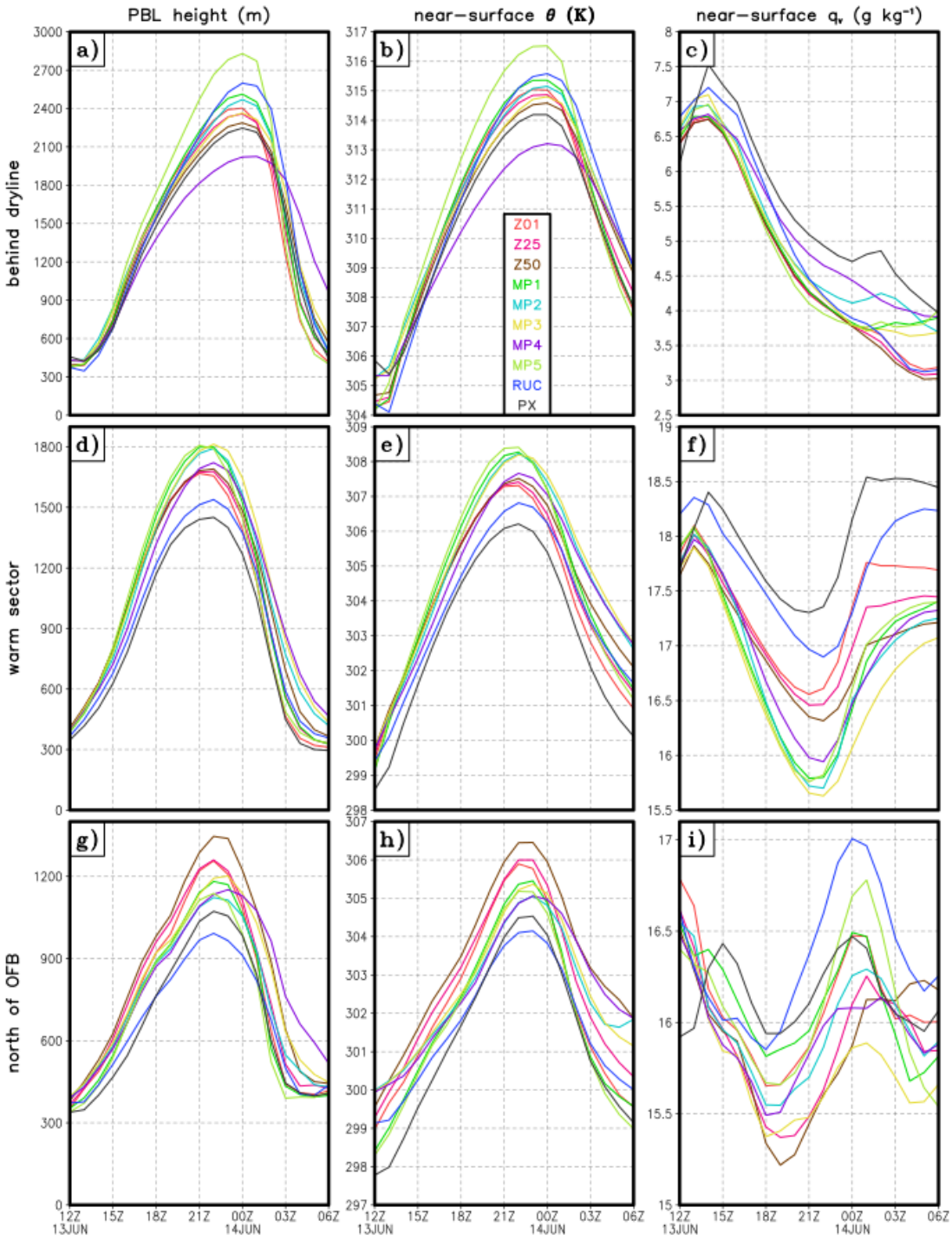

Fig. 5.14. Same as Fig. 5.10 except for the LSMO ensemble.

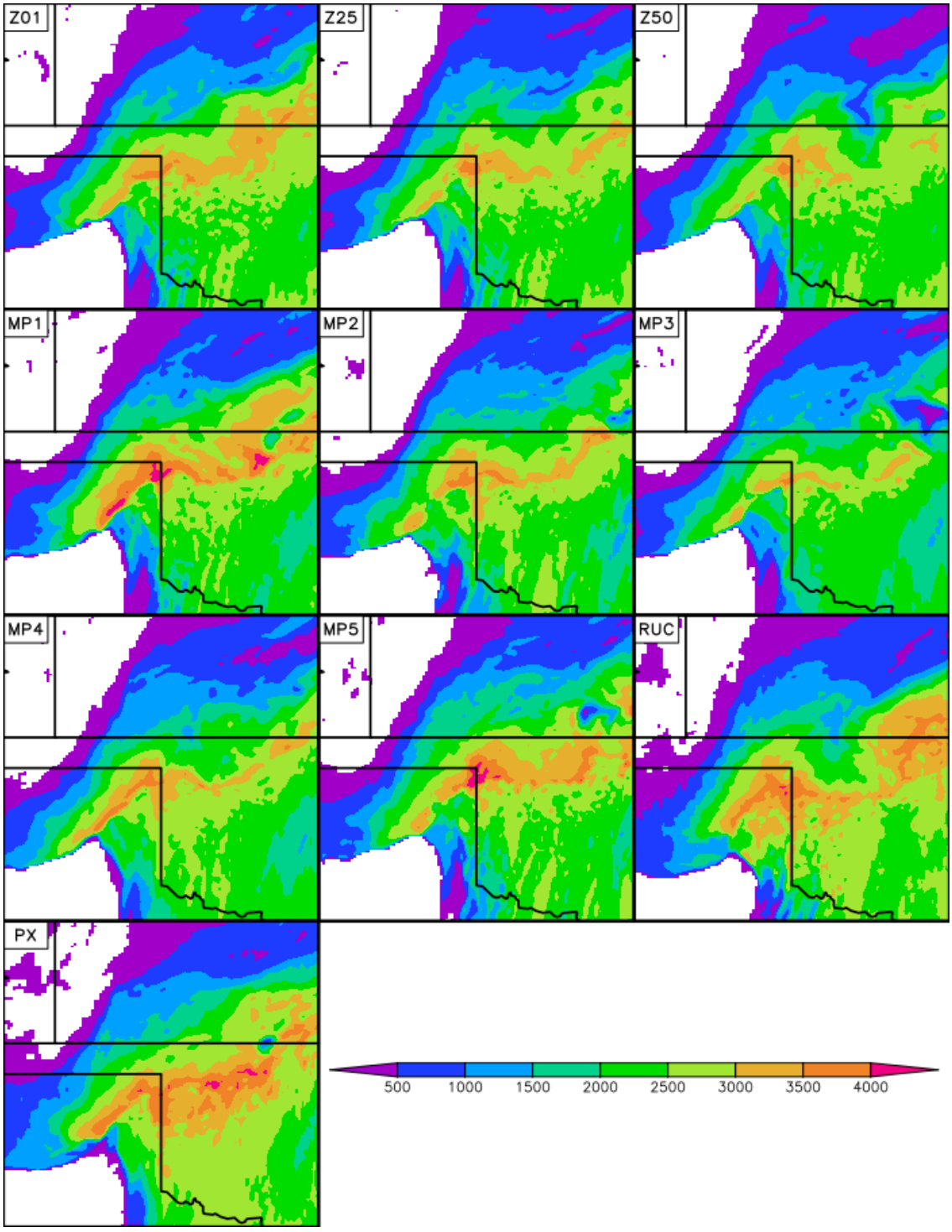

Fig. 5.15. Surface-based CAPE (J kg^{-1}) at 2000 UTC 13 June 2010 for LSMO ensemble members.

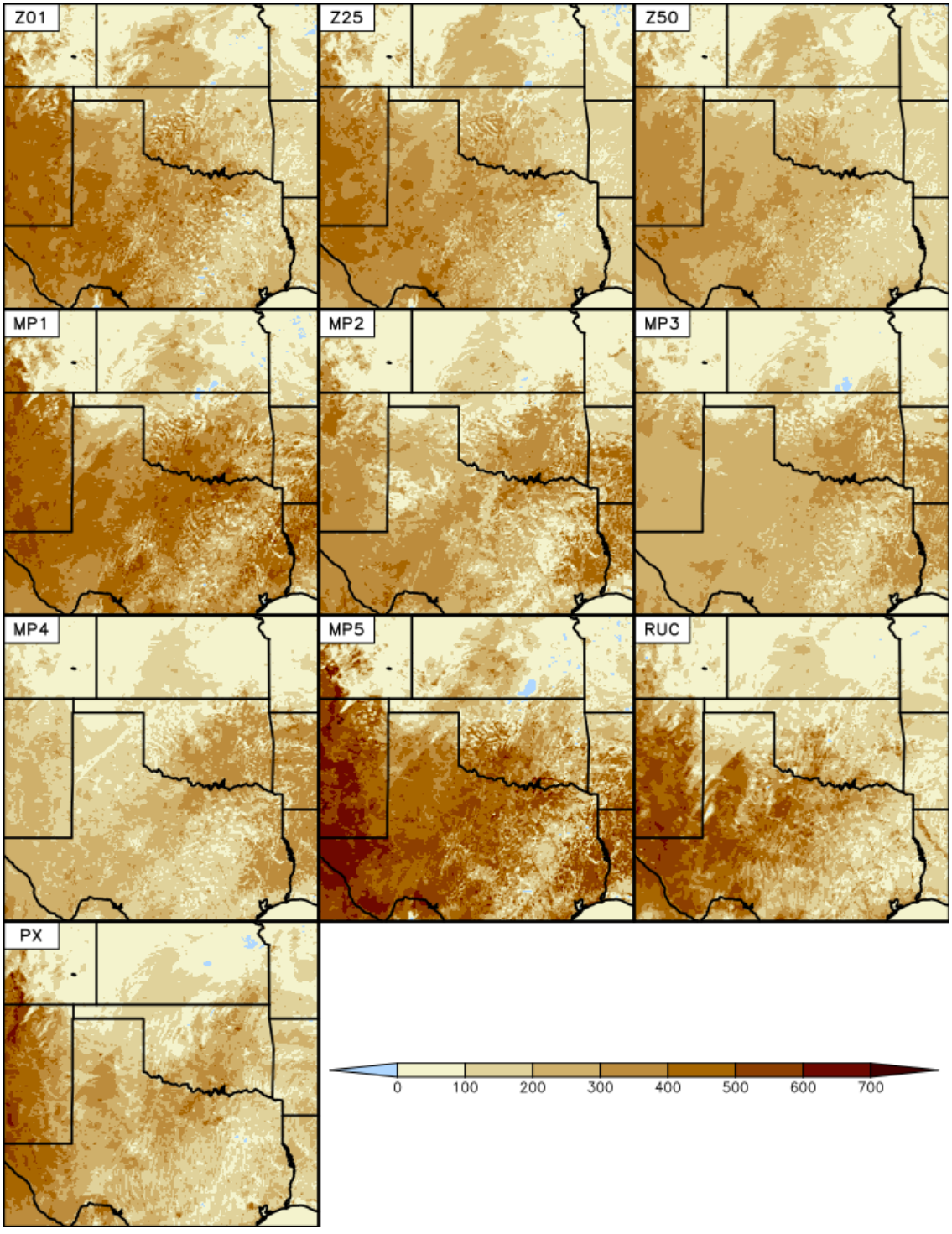

Fig. 5.16. Sensible heat flux (W m^{-2}) at 1900 UTC 13 June 2010 in the LSMO ensemble.

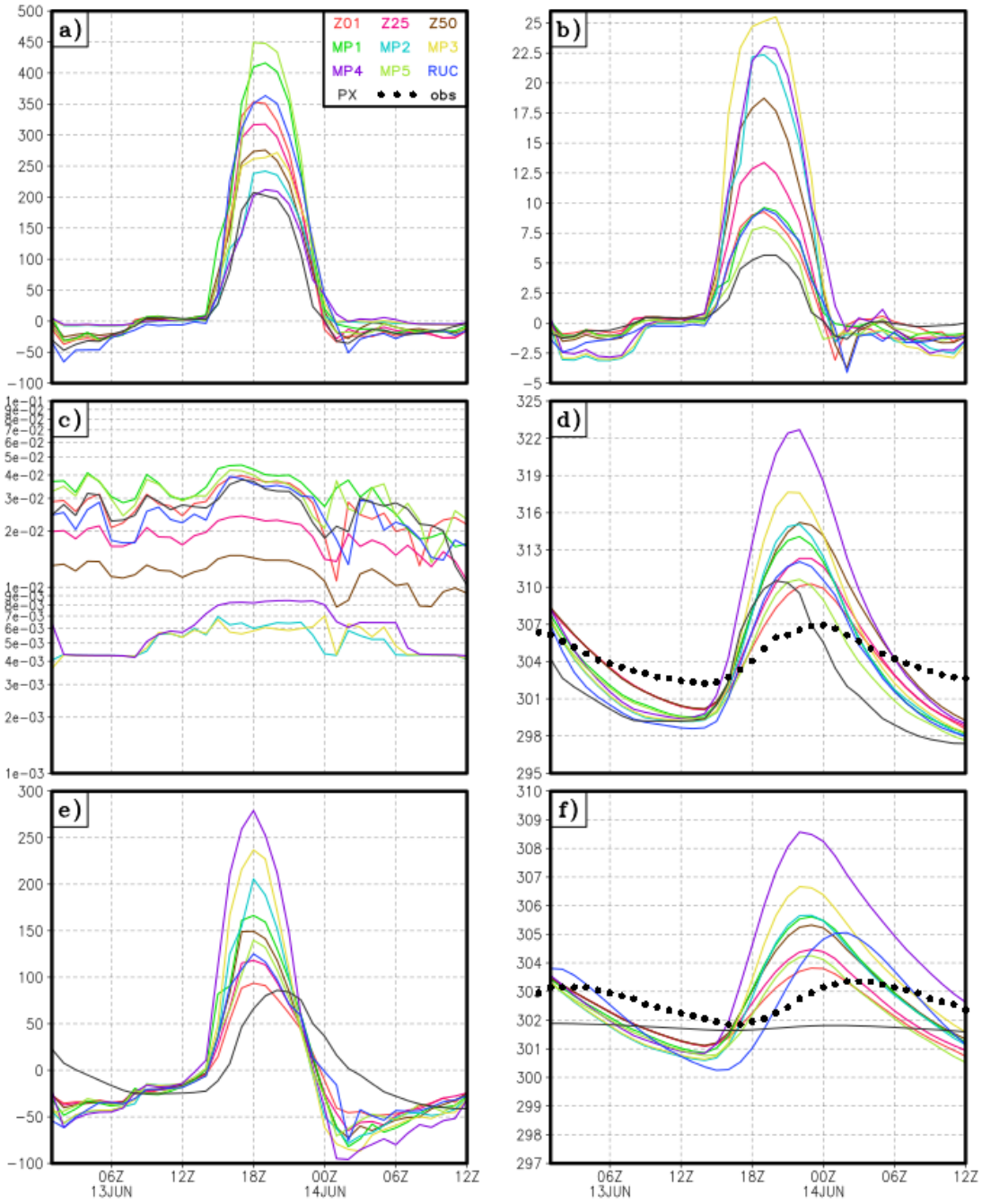

Fig. 5.17. Terms in the surface heat budget from LSMO ensemble members: (a) sensible heat flux (W m^{-2}), (b) difference between skin temperature and temperature at the first model level (K), (c) exchange coefficient for heat (unitless), (d) and (f), soil temperature at 5 and 20 cm depth, and (e) ground heat flux (W m^{-2}). The location is at (lat, lon) = (33.168° N, 100.568° W), corresponding to the West Texas Mesonet site near Jayton, Texas (KJTS). Soil temperature observations are included in (d) and (f).

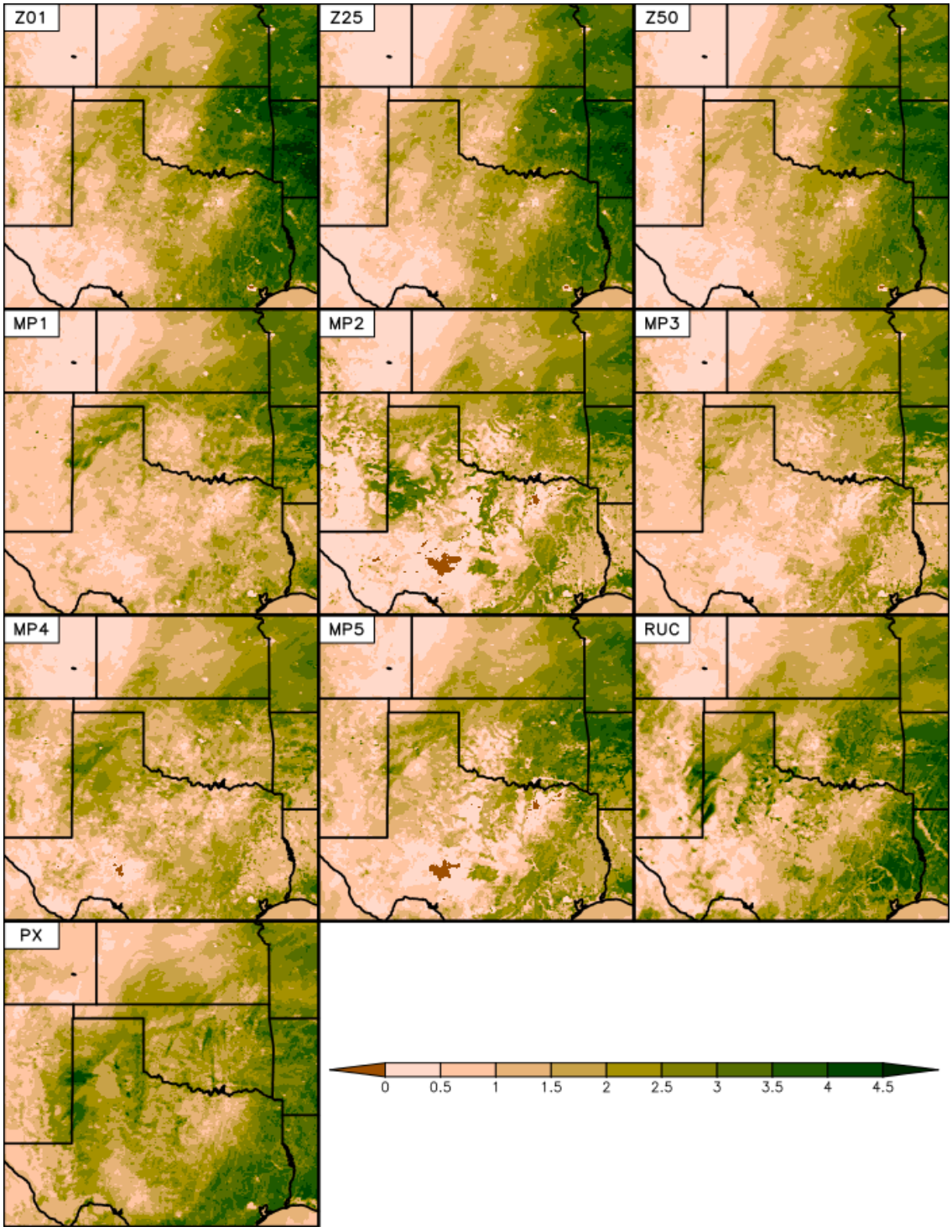

Fig. 5.18. Accumulated moisture flux (kg m^{-2} s^{-1}) between 1200 and 2000 UTC 13 June 2010 from LSMO ensemble members.

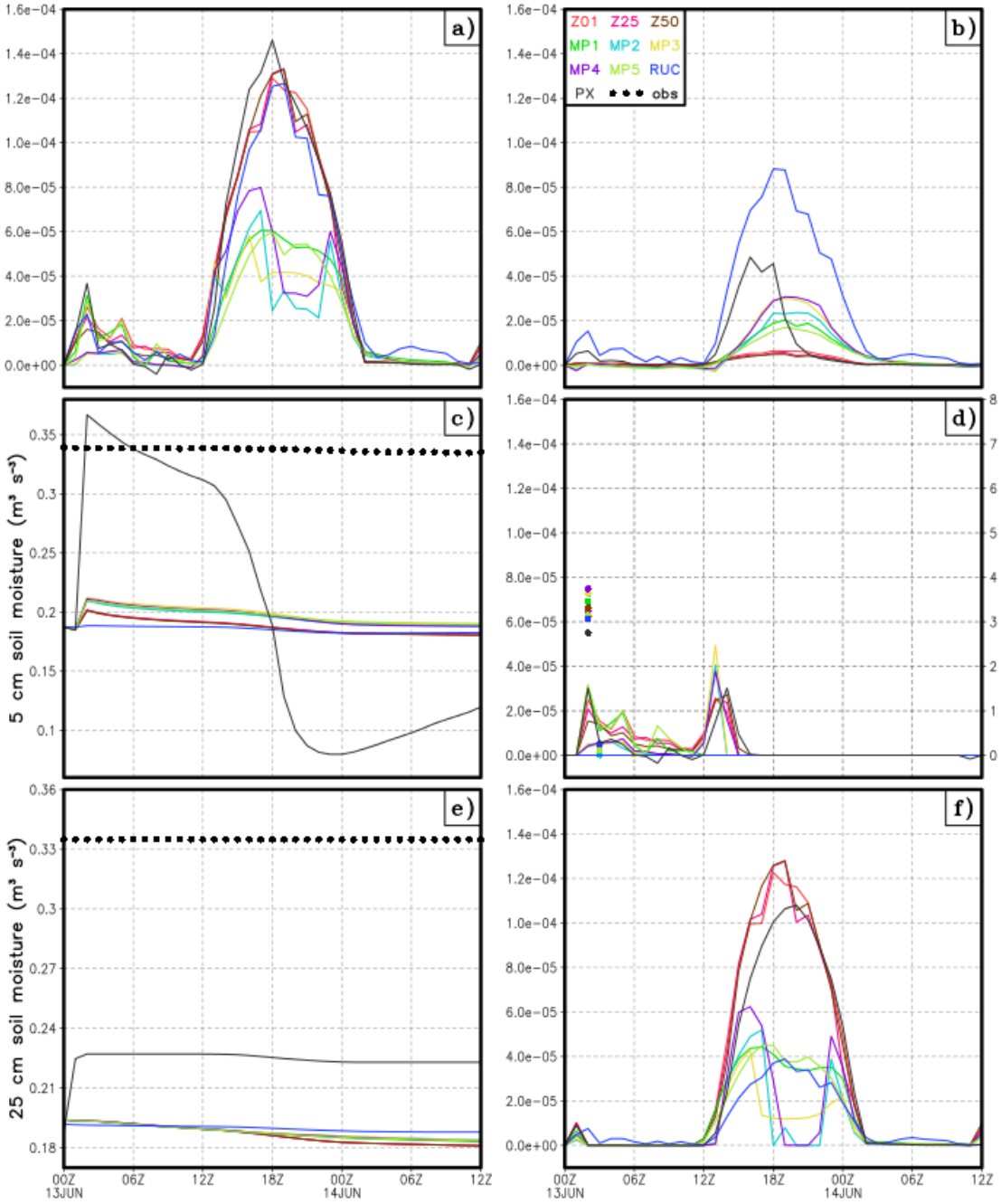

Fig. 5.19. (a) Surface moisture flux, (b) bare soil evaporation, (d) canopy water evaporation, and (f) plant transpiration (kg m^{-2} s^{1}). Soil moisture at (c) 5 cm and (e) 25 cm depths from the LSMO ensemble at (lat, lon) = (36.65°N, 96.80°W), with observed soil moisture provided by the Oklahoma Mesonet site located near Chandler, Oklahoma. One-hour accumulated precipitation (mm) is also dotted and scaled to the right-hand vertical axis in (d).

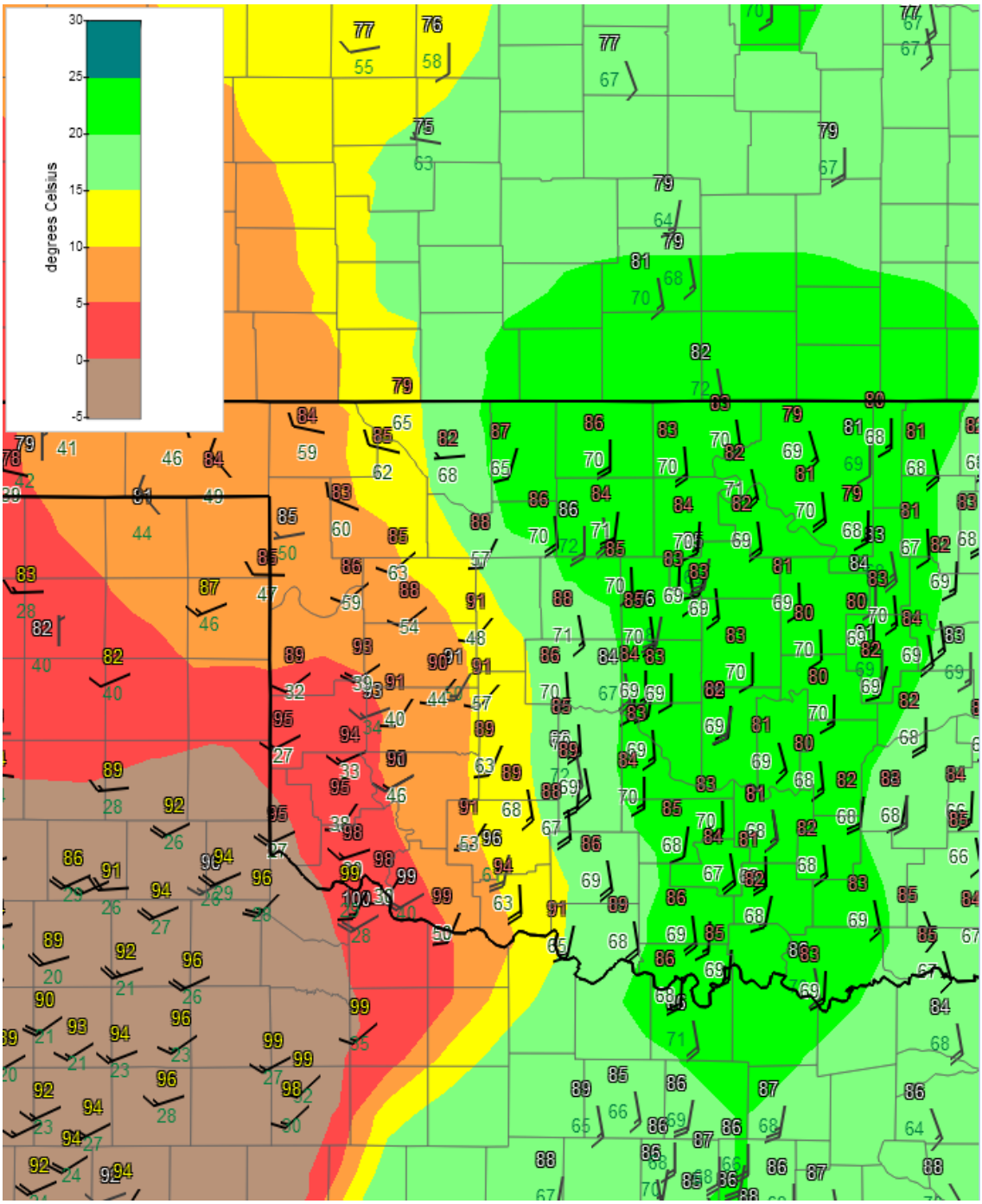

Fig. 5.20. Observed 2-m temperature and dewpoint (°F) and 10-m wind barbs (kts) from various sources indicated by color of temperature observations, including the Oklahoma Mesonet (light red), West Texas Mesonet (yellow), and National Weather Service (white) at 1900 UTC 19 May 2013. An objective analysis of dewpoint (°C) is color shaded.

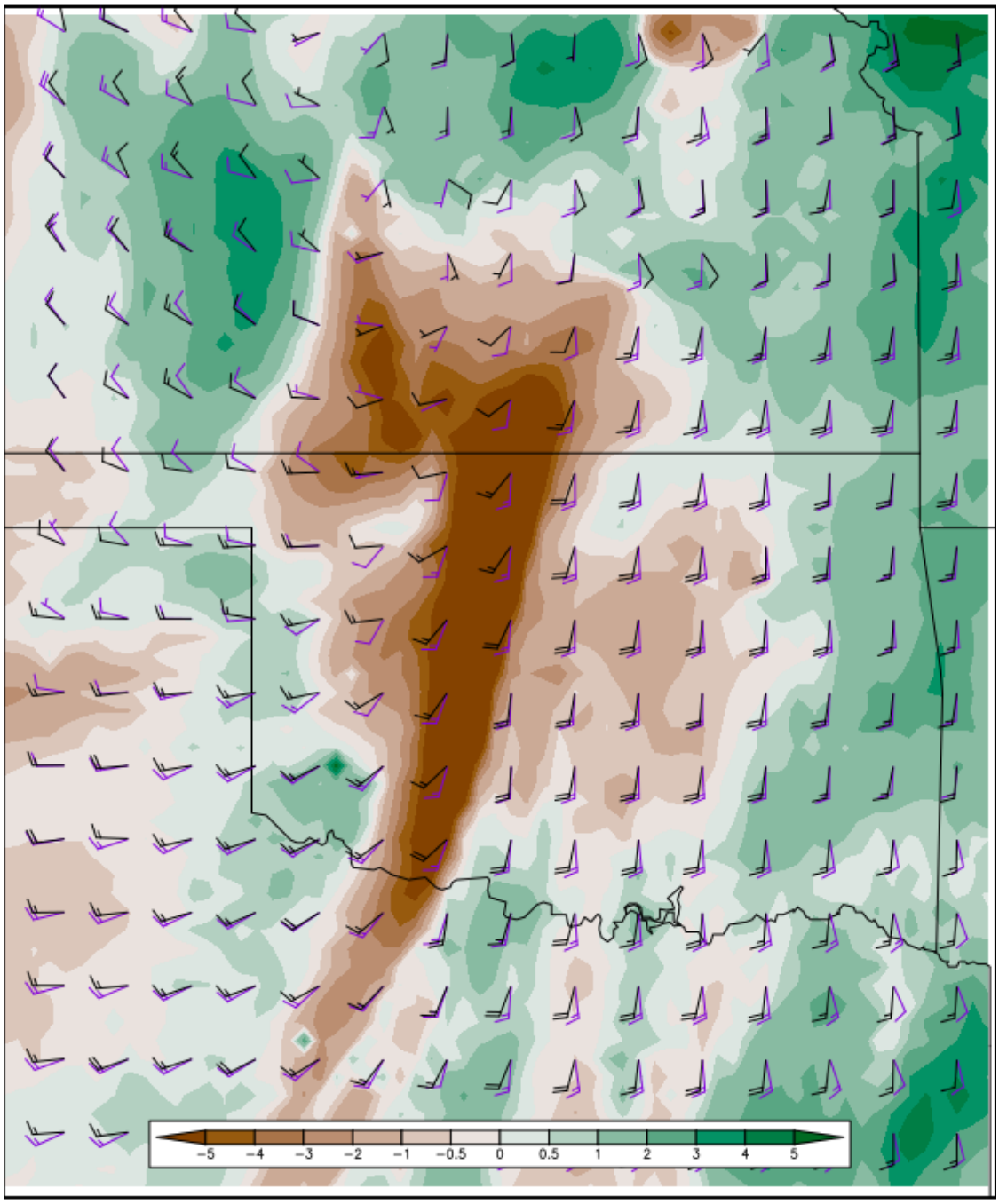

Fig. 5.21. Same as Fig. 5.3 except valid at 1900 UTC 19 May 2013.

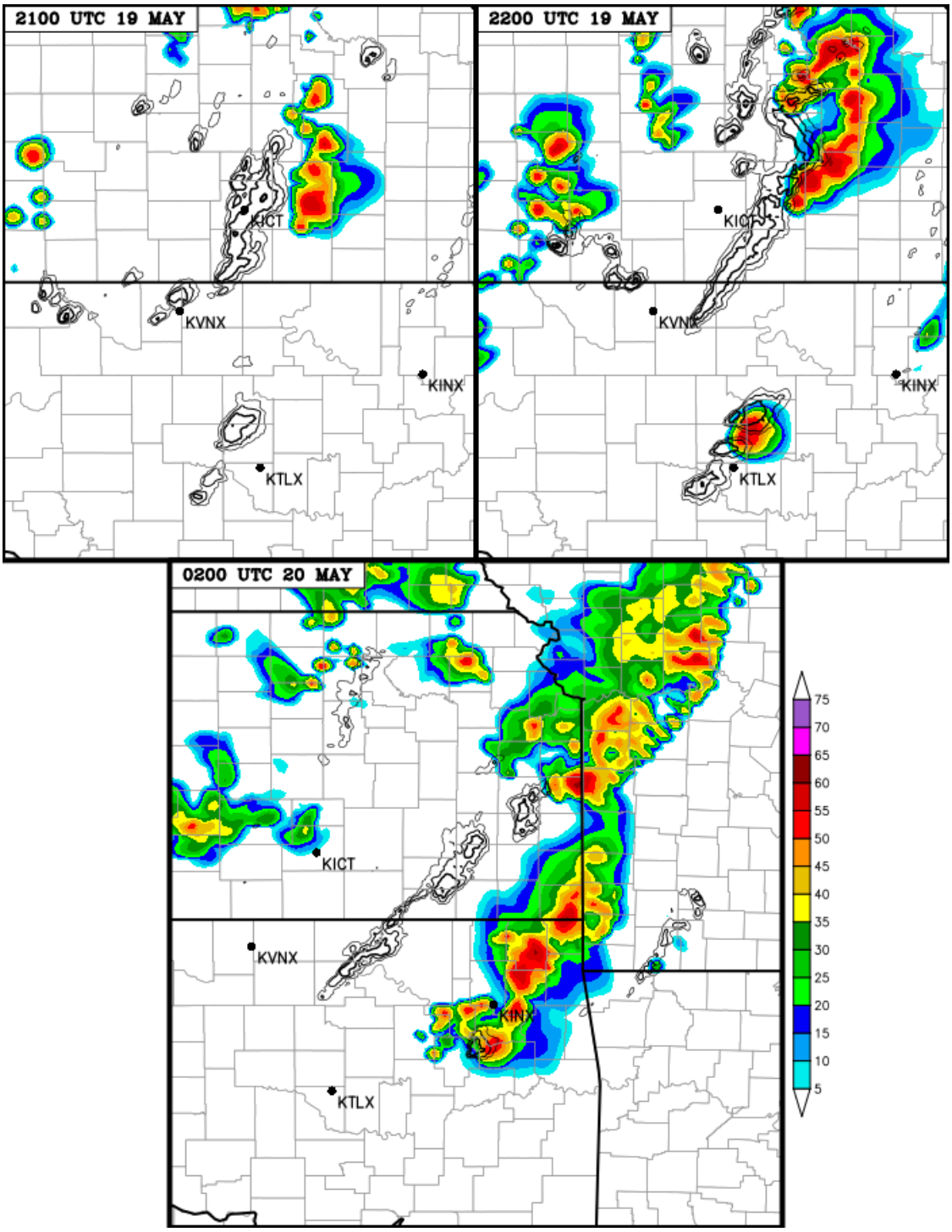

Fig. 5.22. Same as Fig. 5.5 except for the times indicated. Only observed reflectivity from indicated sites is plotted.

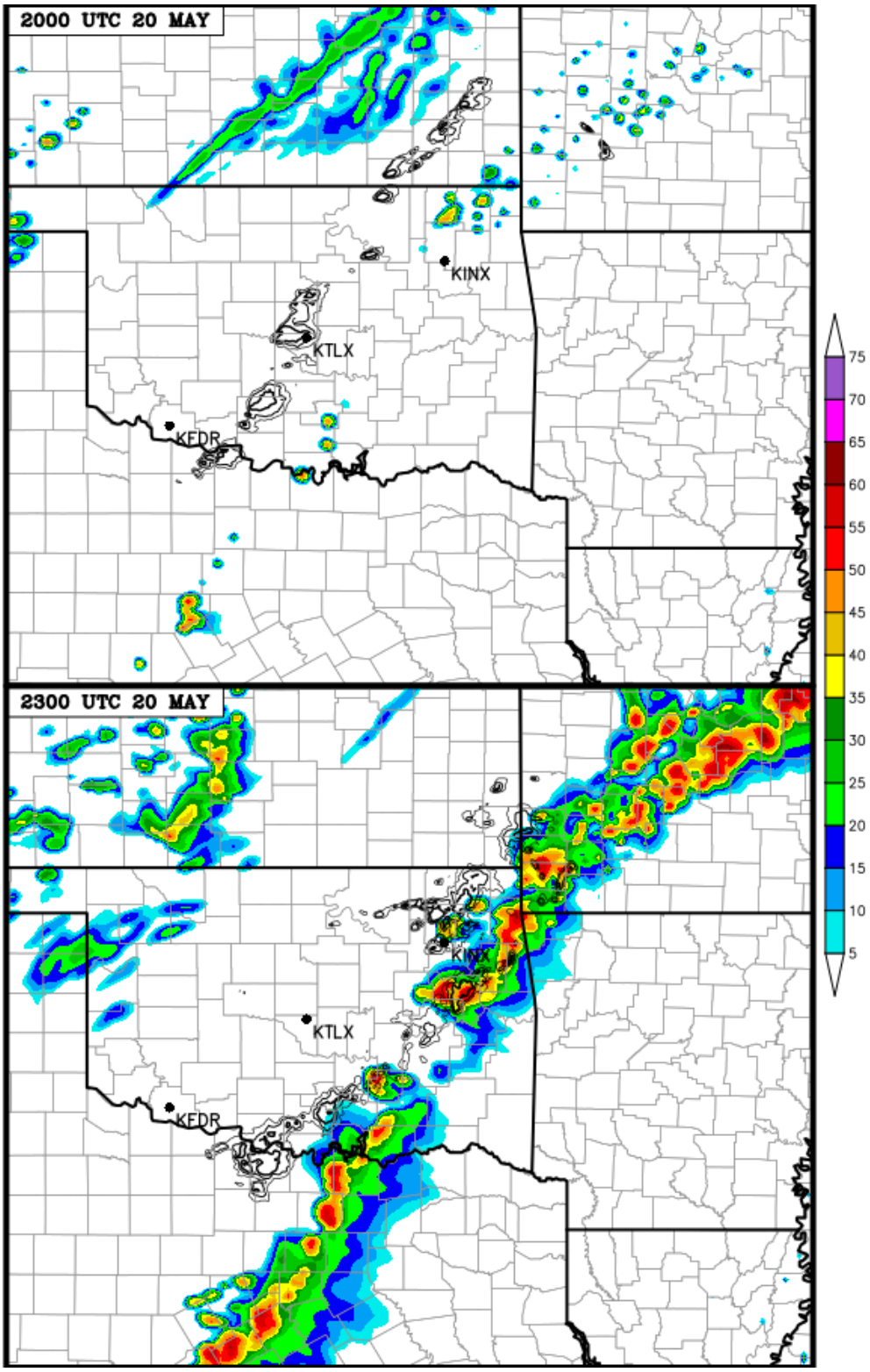

Fig. 5.23. Same as Fig. 5.22 except for the times indicated.

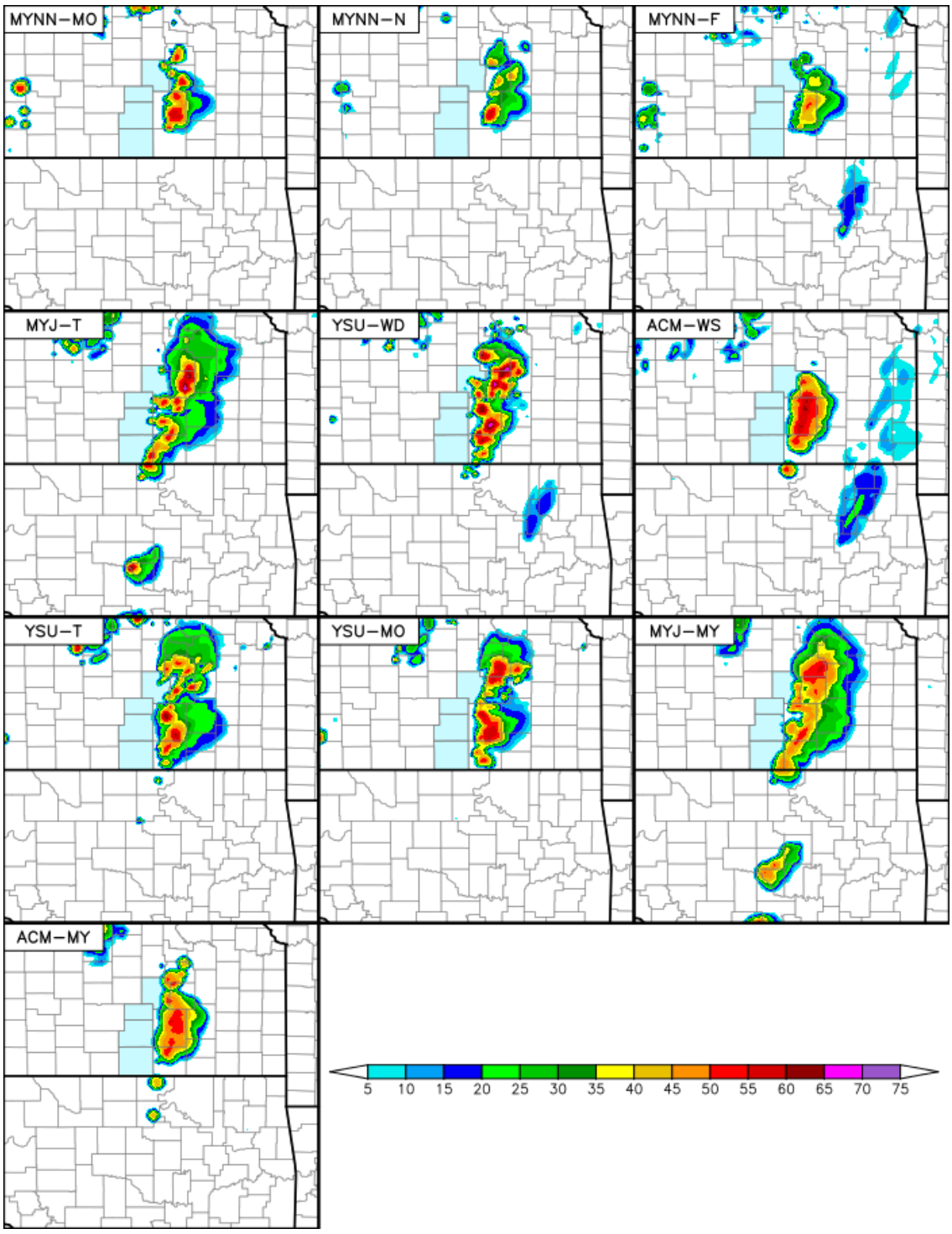

Fig. 5.24. Simulated composite reflectivity (dBZ) from FLSM ensemble members; a 9-hour forecast valid 2100 UTC 19 May 2013. The following counties in Kansas are shaded in light blue from southwest to northeast: Sumner, Sedgwick, Harvey, and Marion.

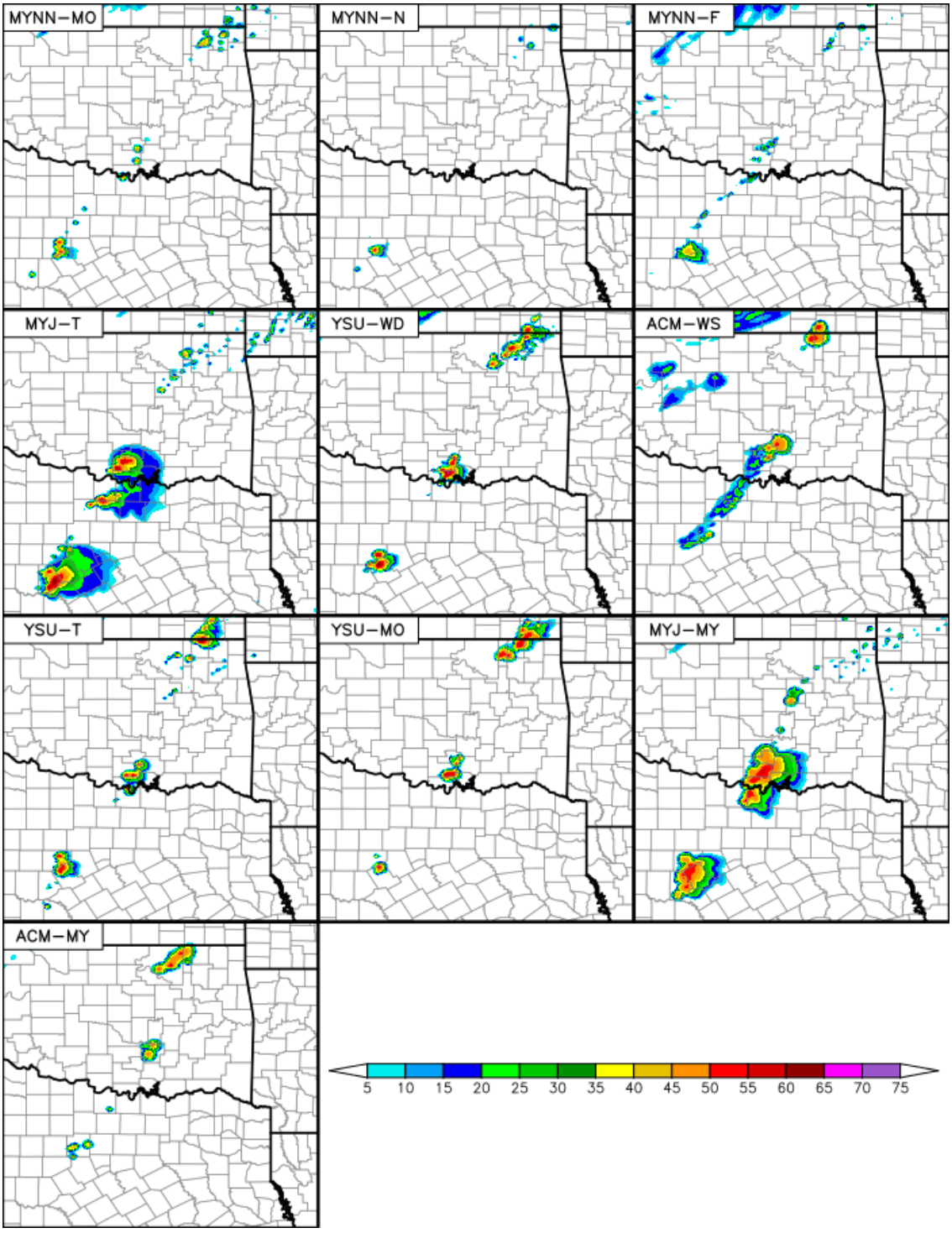

Fig. 5.25. Same as Fig. 5.24 except valid at 2000 UTC 20 May 2013 (a 32-hour forecast).

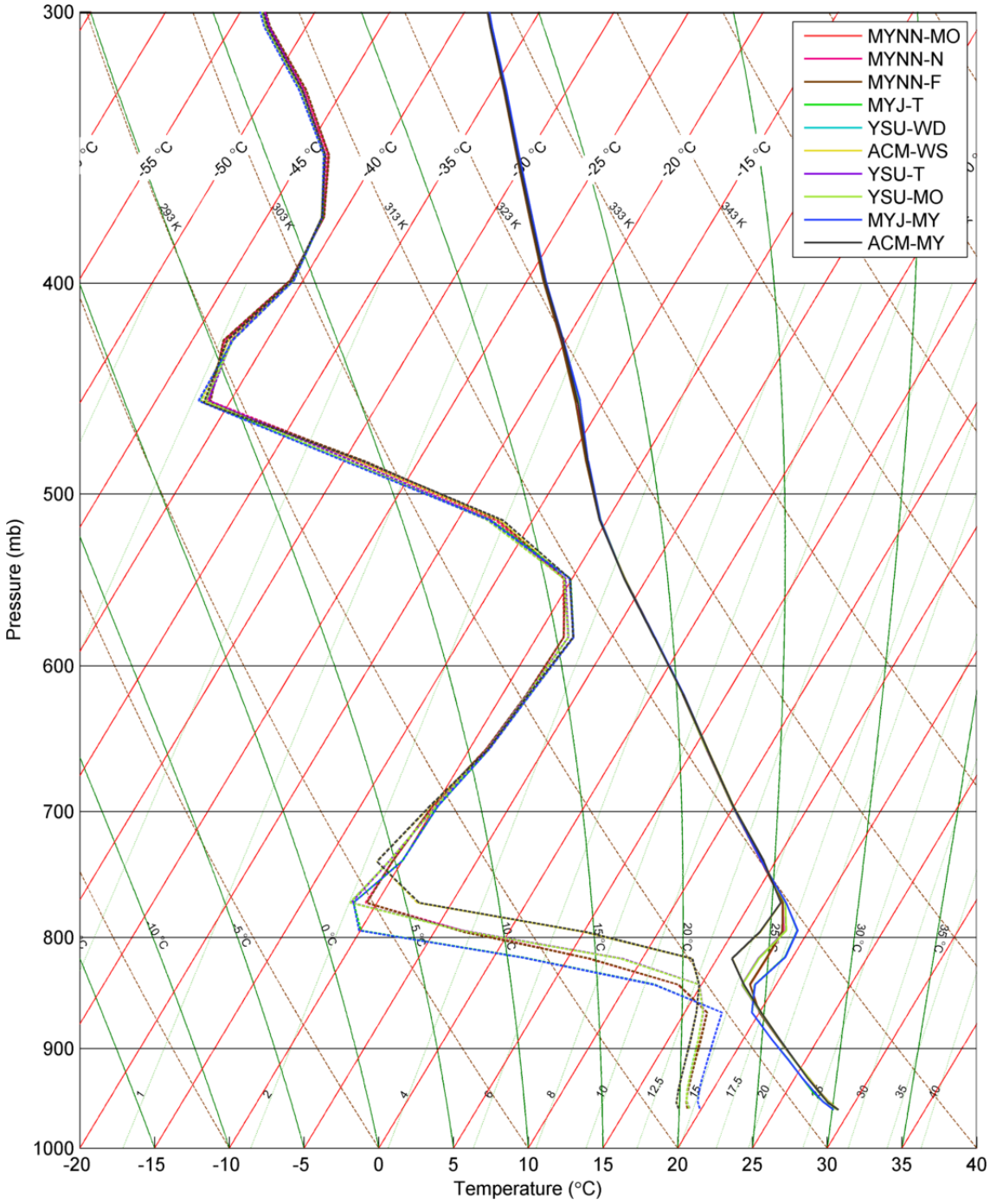

Fig. 5.26. Same as Fig. 5.11 except a 7-hour forecast valid at 1900 UTC 13 May 2013 and taken from (lat, lon) = (35.375°N, 97.14°W), between Oklahoma City and Shawnee, OK.

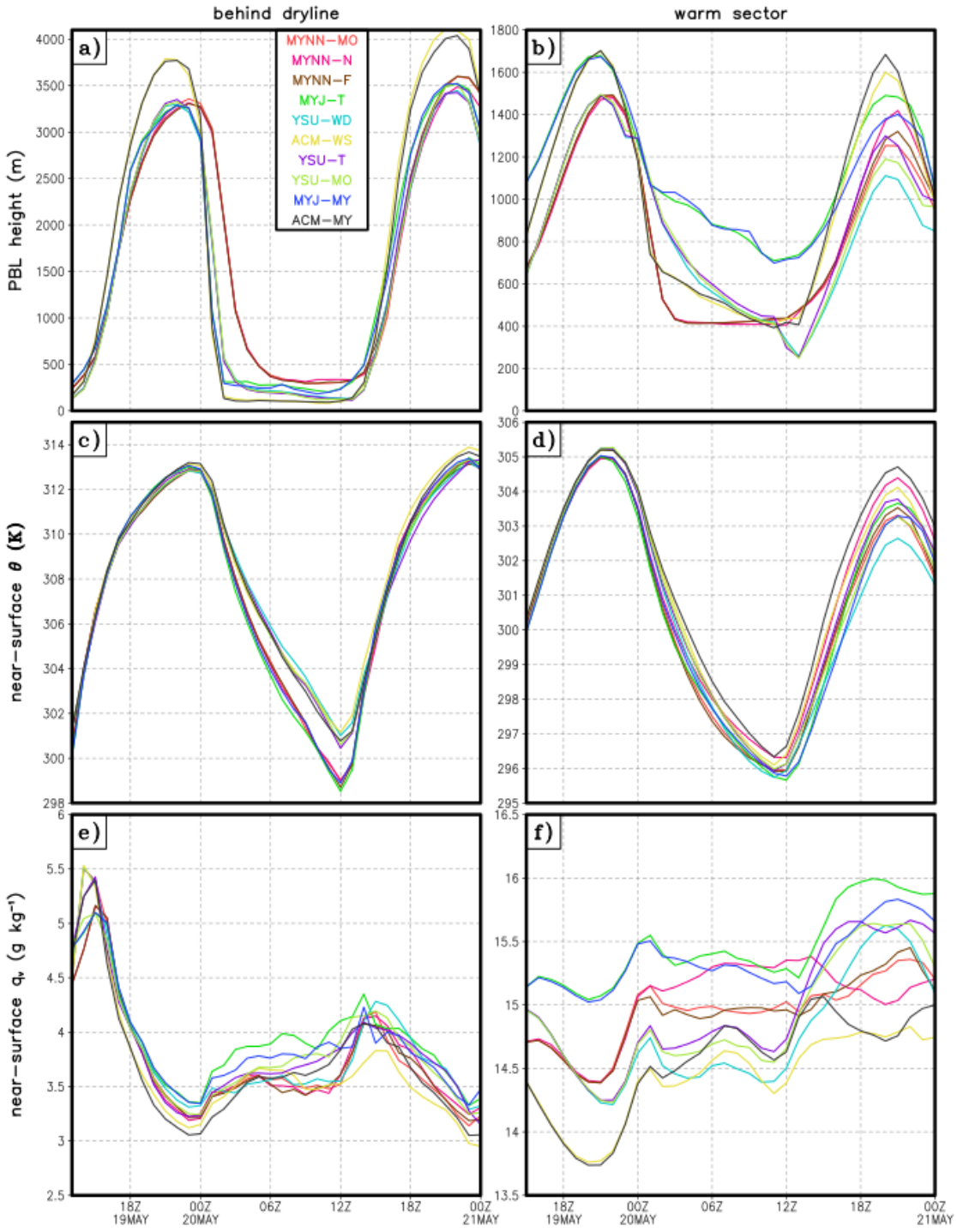

Fig. 5.27. Air mass averaged (a),(b) PBL height (m); (c),(d) near-surface potential temperature (K); and (e),(f) near-surface mixing ratio (g kg^{-1}) (a),(c),(e) behind the dryline, and (b),(d),(f) in the warm sector for FLSM ensemble members.

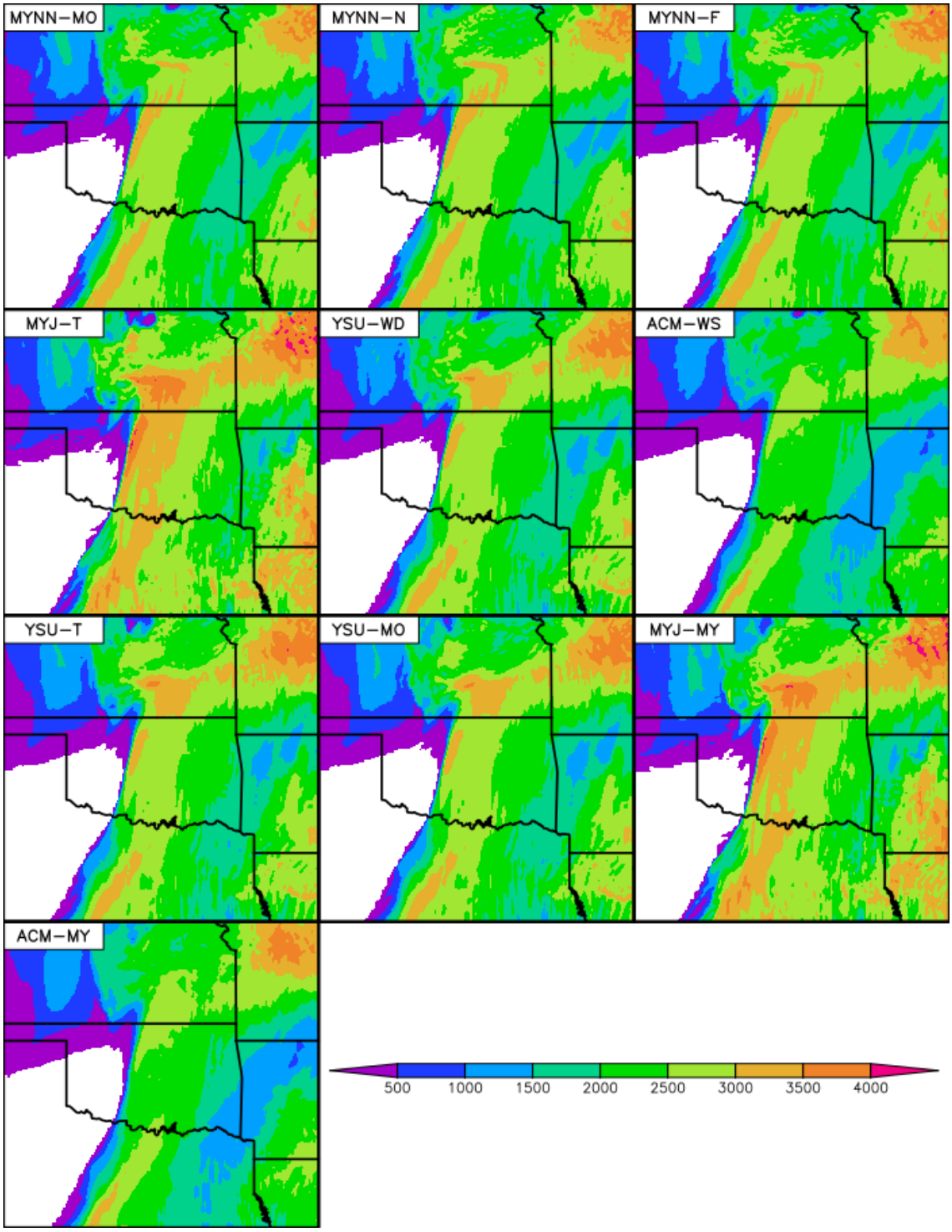

Fig. 5.28. MLCAPE (J kg^{-1}) at 1900 UTC 19 May 2013 from FLSM ensemble members.

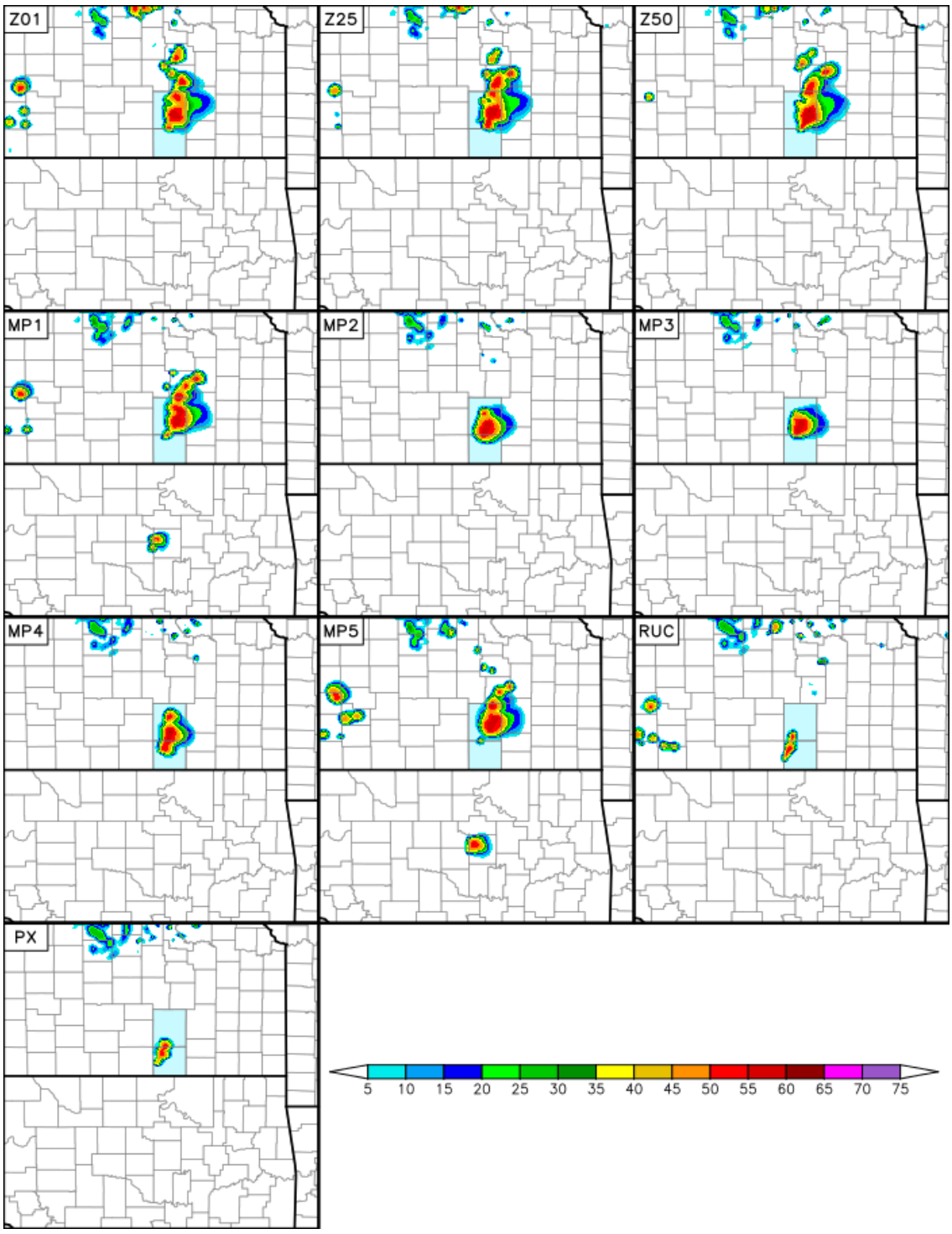

Fig. 5.29. Same as Fig. 5.24 except from the LSMO ensemble. Butler and Cowley Counties, Kansas are shaded light blue.

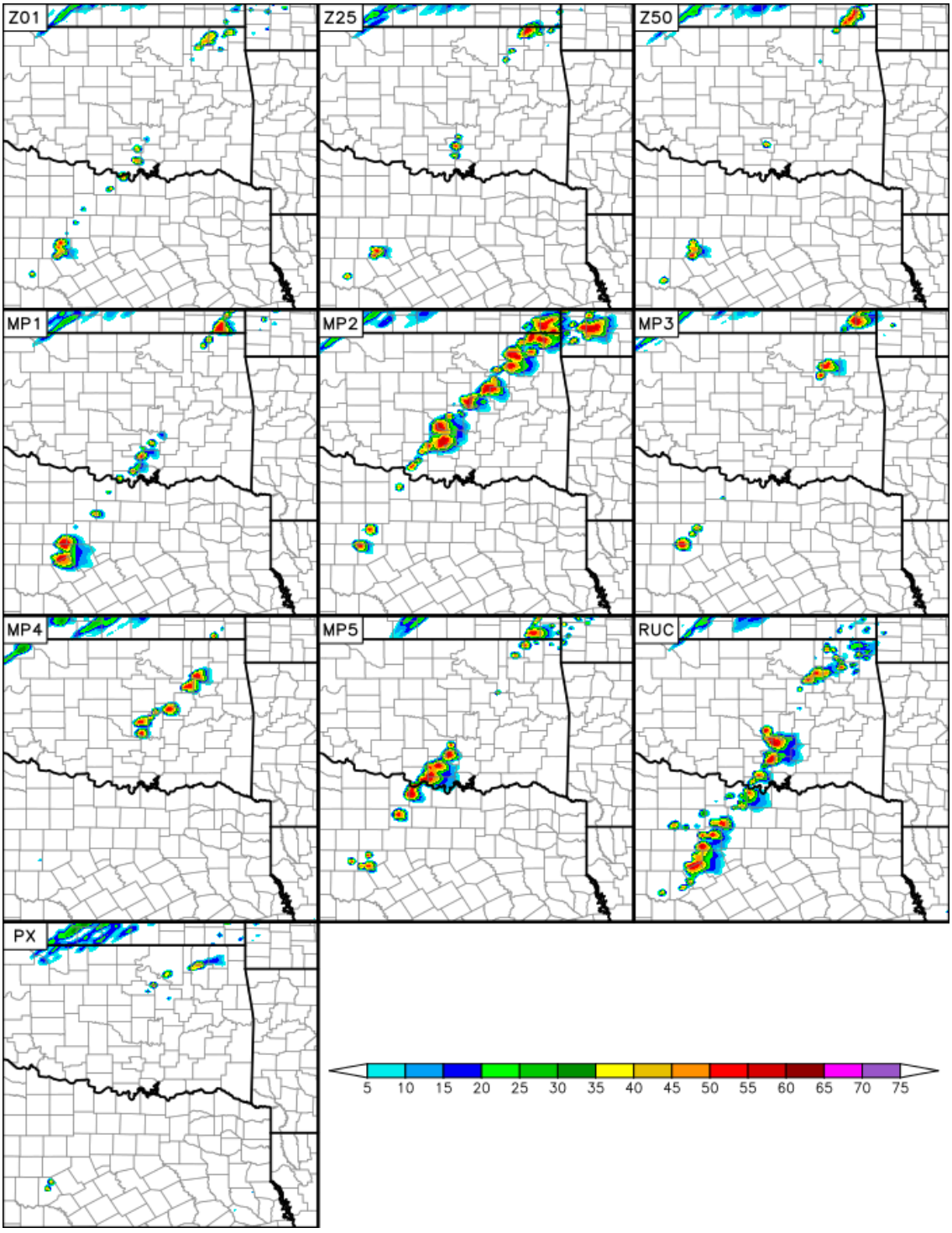

Fig. 5.30. Same as Fig. 5.25 except for members of the LSMO ensemble.

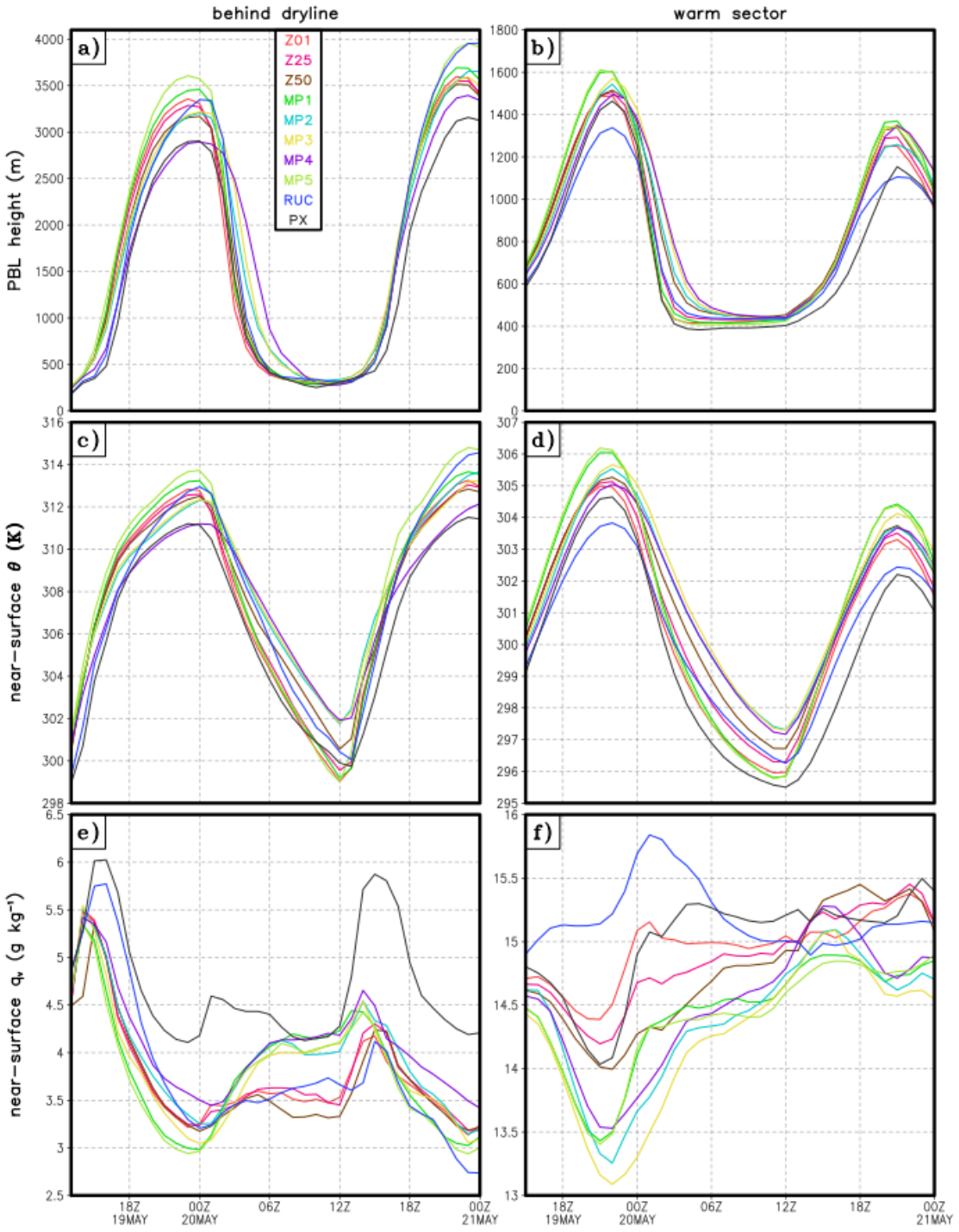

Fig. 5.31. Same as Fig. 5.27 except for the LSMO ensemble members.

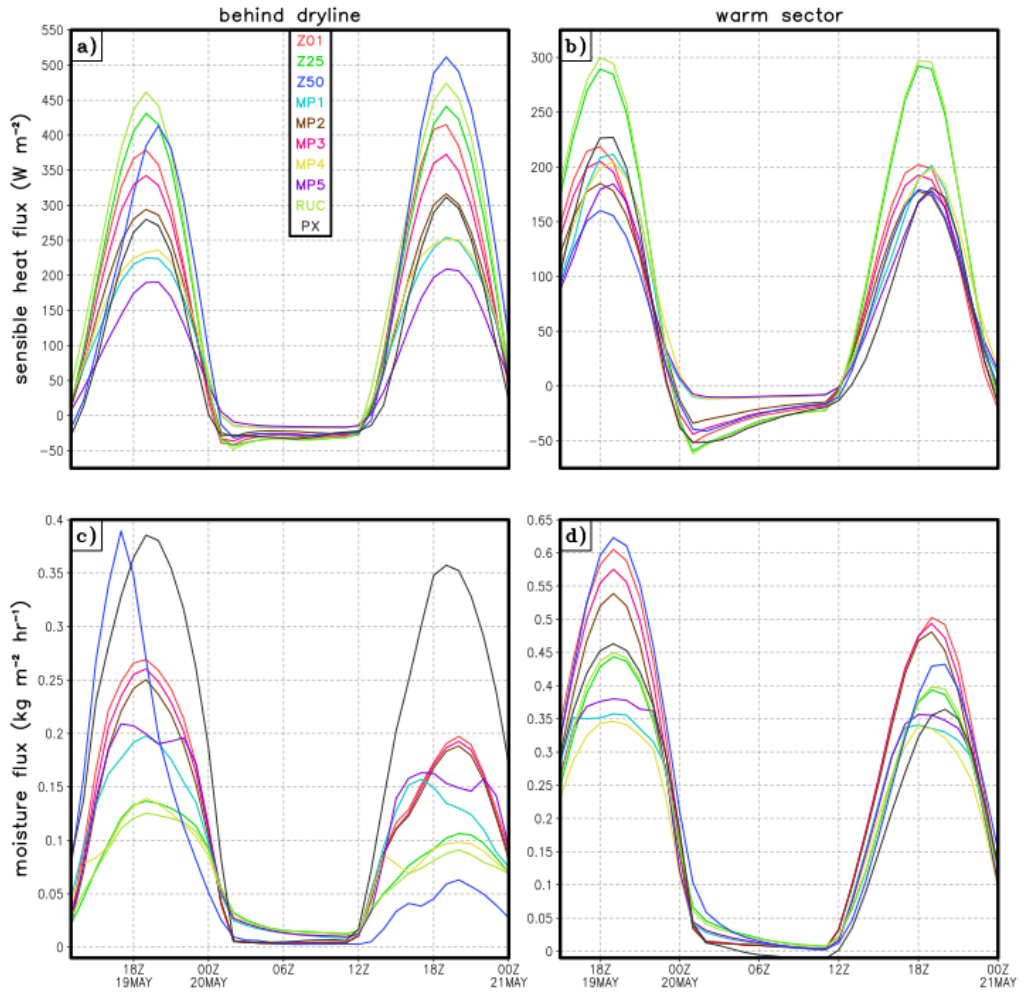

Fig. 5.32. Air mass averaged (a),(b) sensible heat flux and (c),(d) moisture flux (a),(c) behind the dryline and (b),(d) in the warm sector for case #2 from LSMO ensemble members. All units are W m^{-2}.

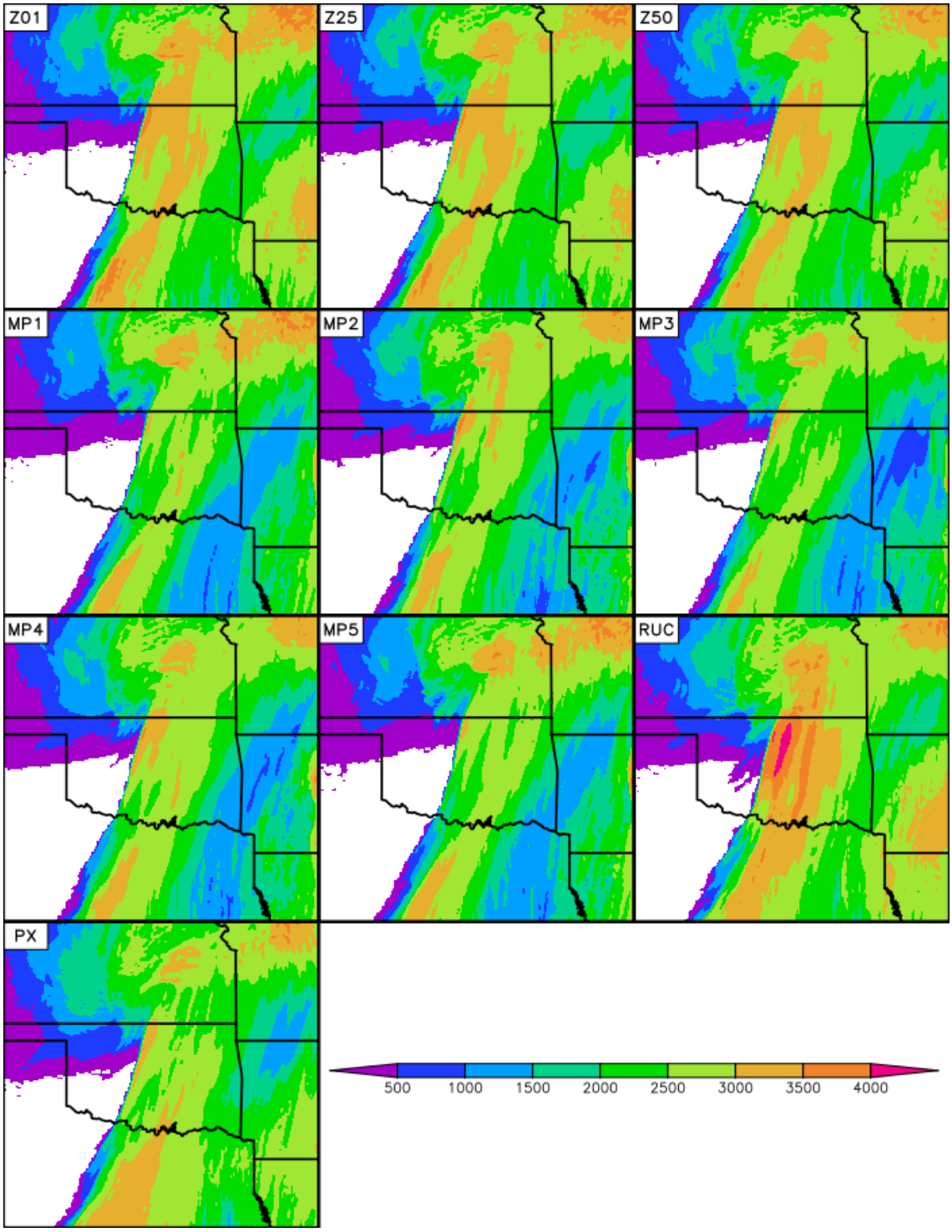

Fig. 5.33. Maximum MLCAPE (J kg^{-1}) between 2000 and 2200 UTC 19 May 2013 from LSMO ensemble members.

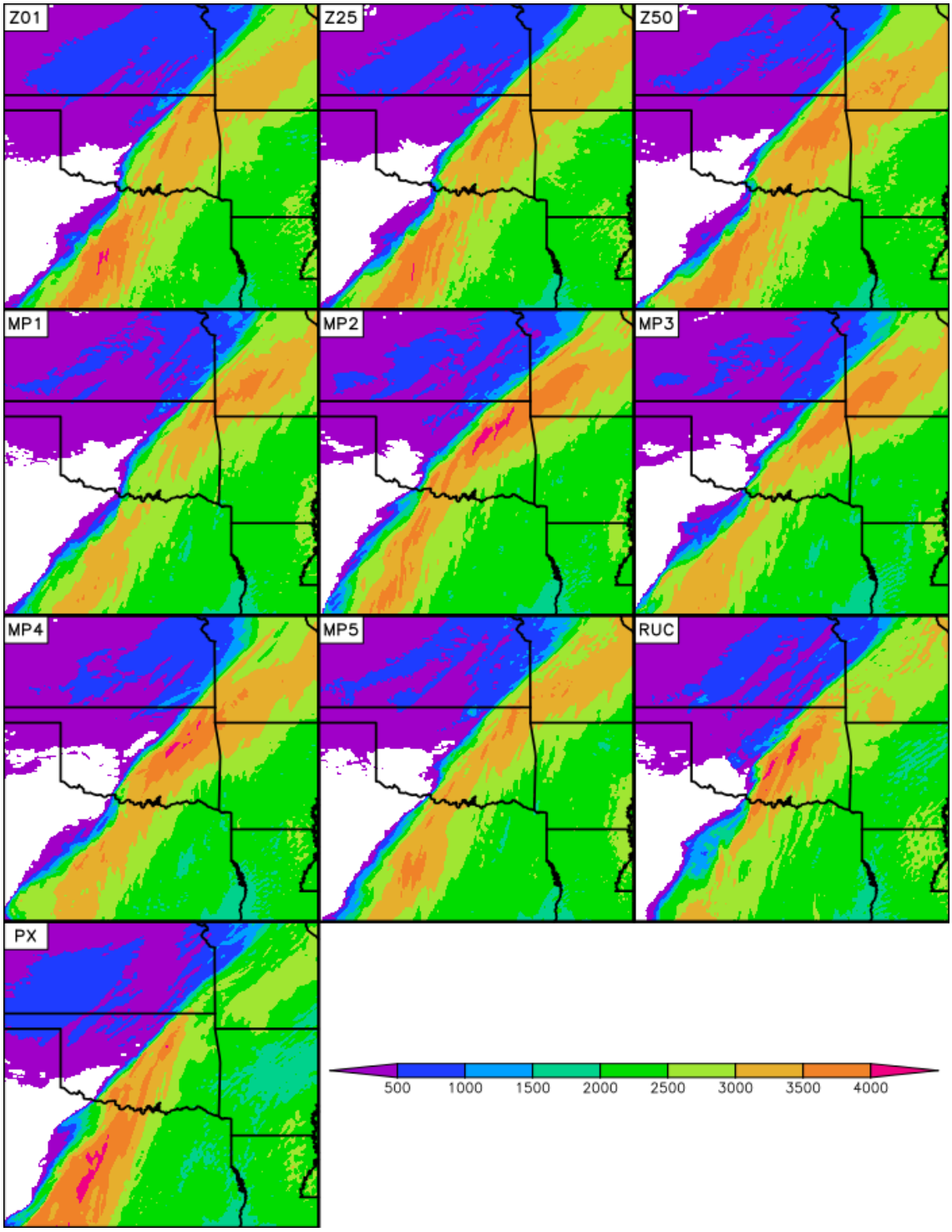

Fig. 5.34. Same as Fig. 5.33 except for between 1800 and 2000 UTC 20 May 2013.

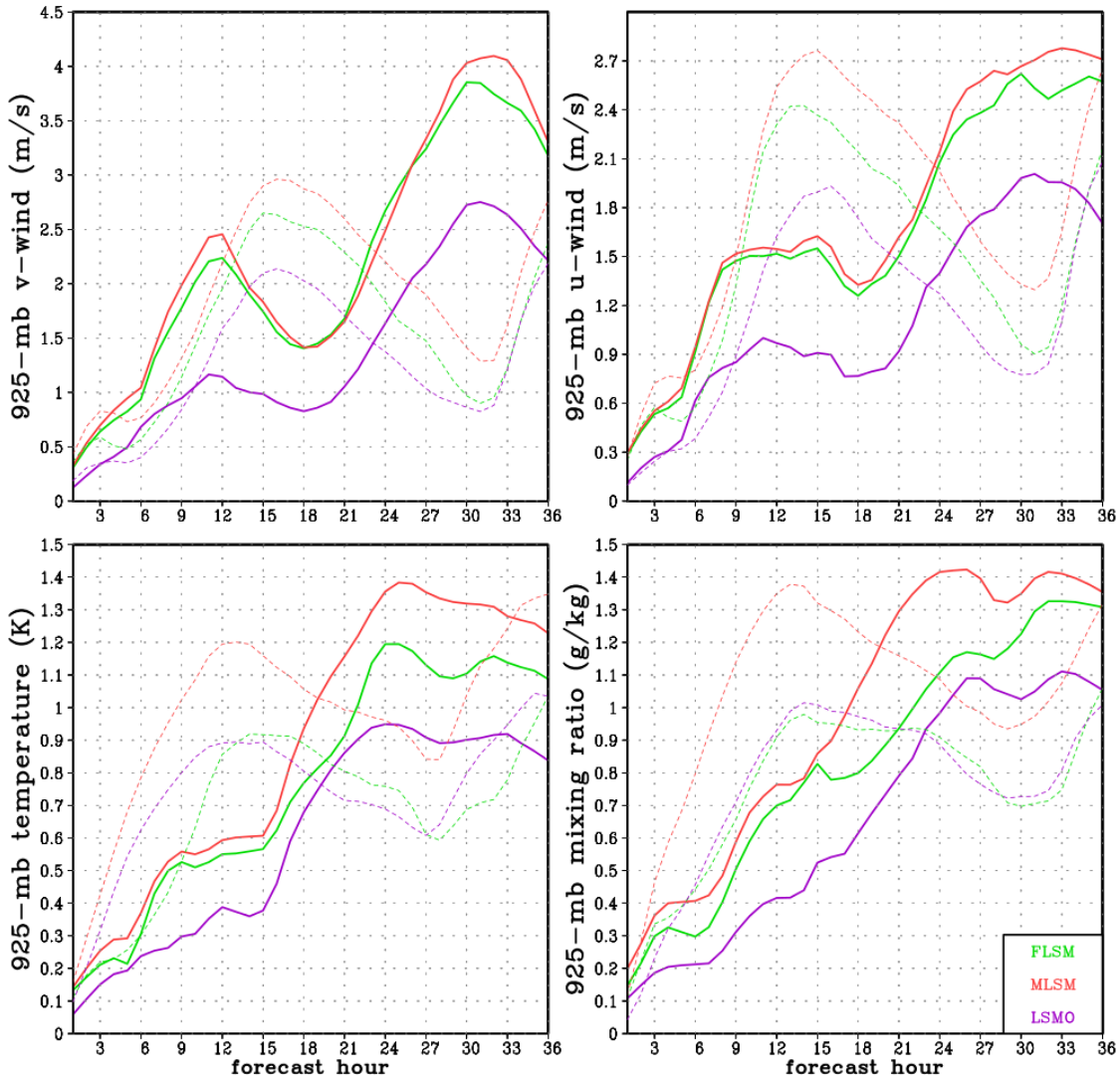

Fig. 5.35. Ensemble spread for the indicated fields over a subset of the model domain covering most of the southern Plains. The bold solid (thin dashed) lines are for case 1 (2).

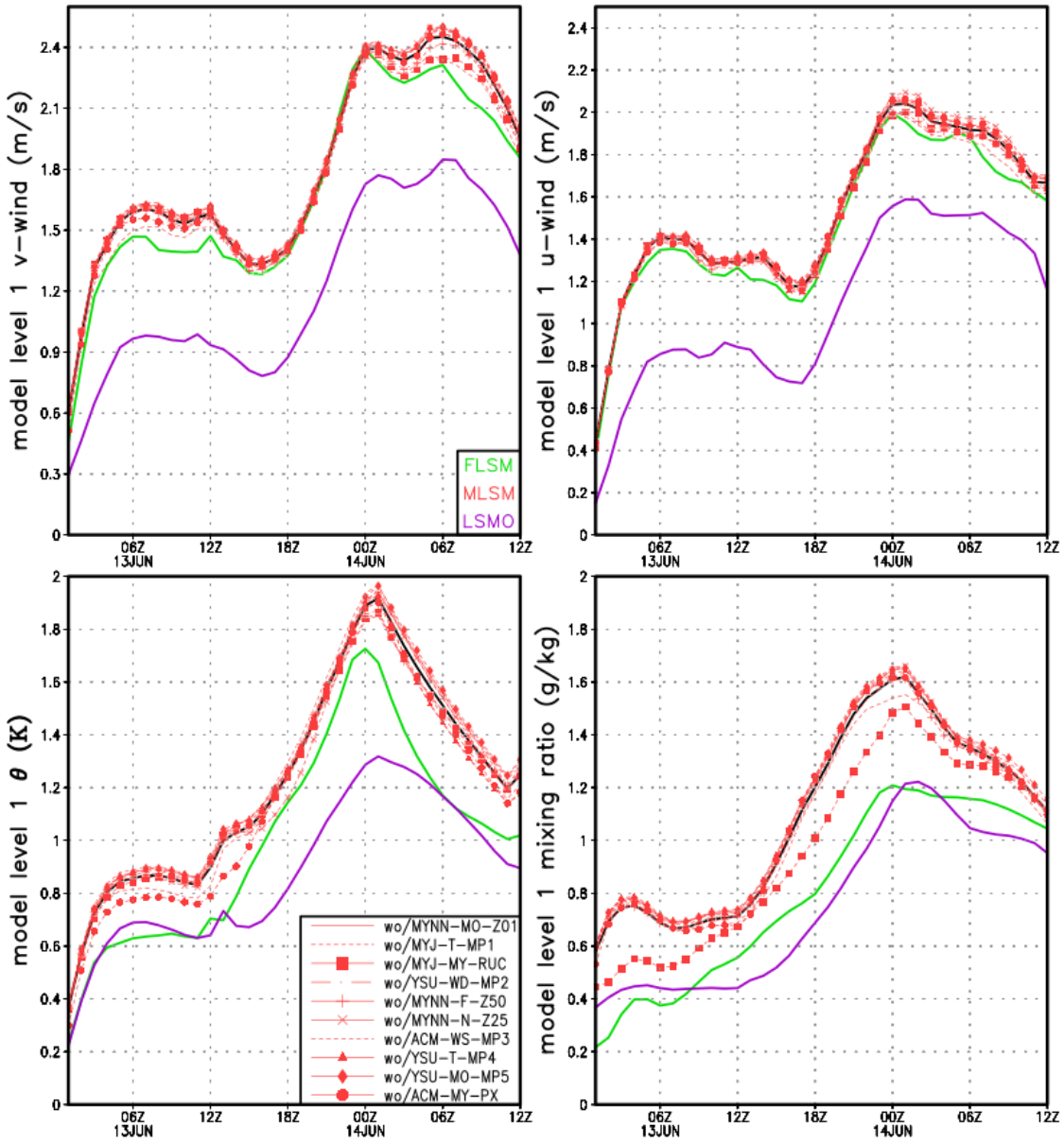

Fig. 5.36. Ensemble spread as in Fig. 5.31 except for the case initialized at 0000 UTC 13 June 2010. 10-member ensemble spread is shown in the bold lines, whereas the thin lines with symbols delineate spread of 9-member ensembles with the indicated member removed.

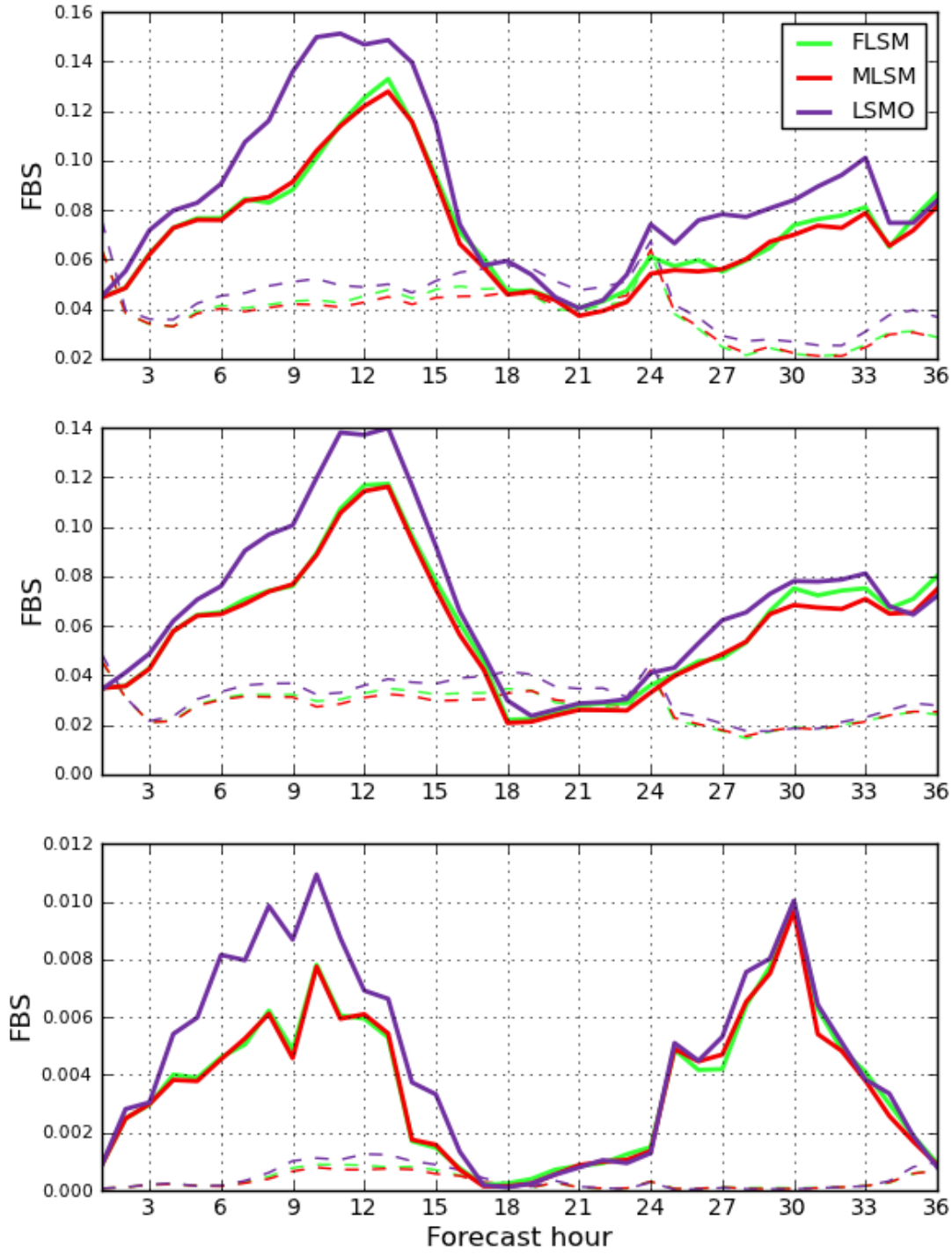

Fig. 5.37. Fractions Brier scores for one-hour accumulated precipitation at the following thresholds: (top) 0.254 mm, (middle) 2.54 mm, (bottom) 12.7 mm. Solid lines indicate case #1 scores with case #2 scores in dashed lines.

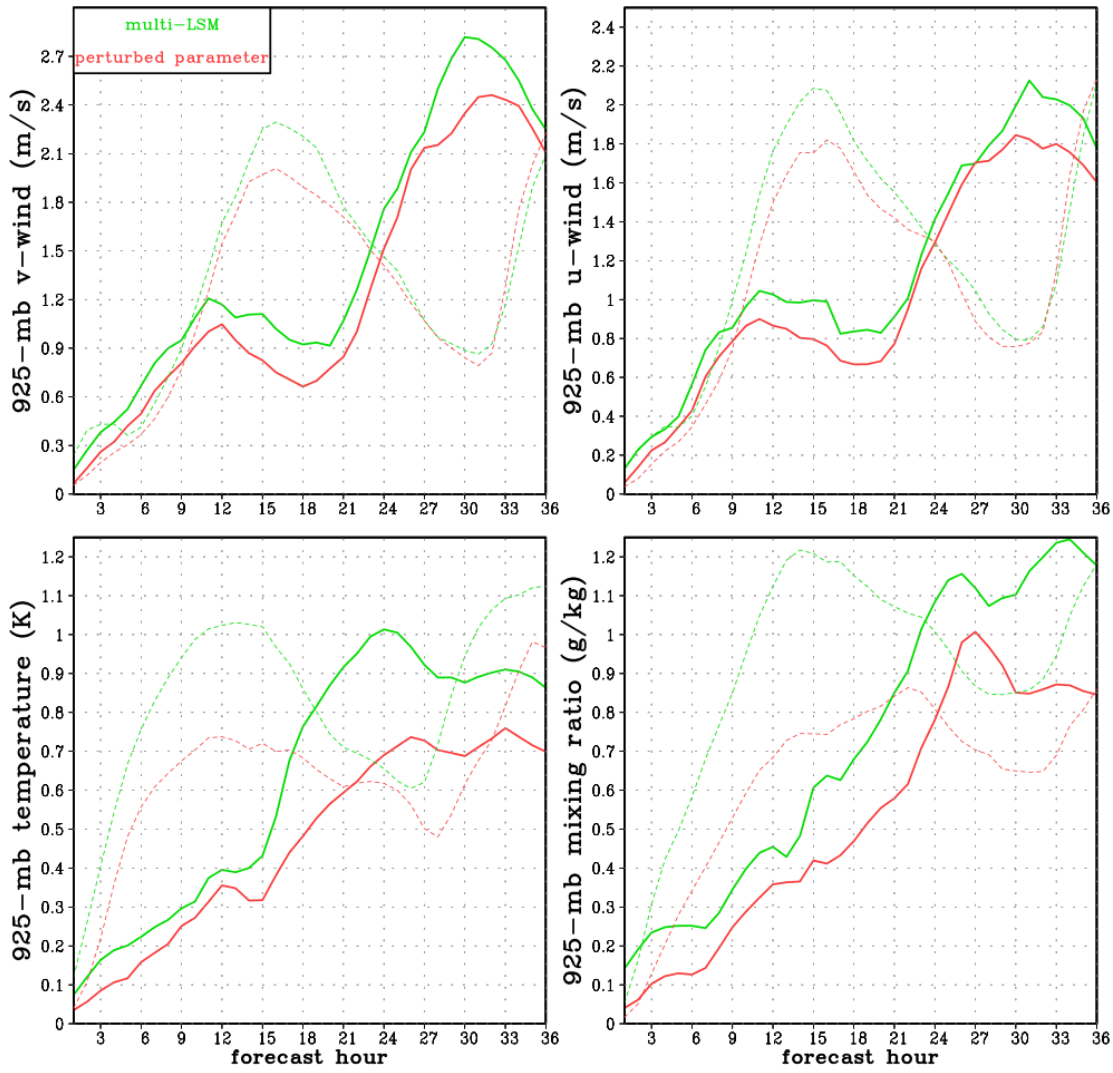

Fig. 5.38. Same as Fig. 5.31 except for a 4-member multi-LSM sub-ensemble and a 5-member perturbed parameter sub-ensemble.

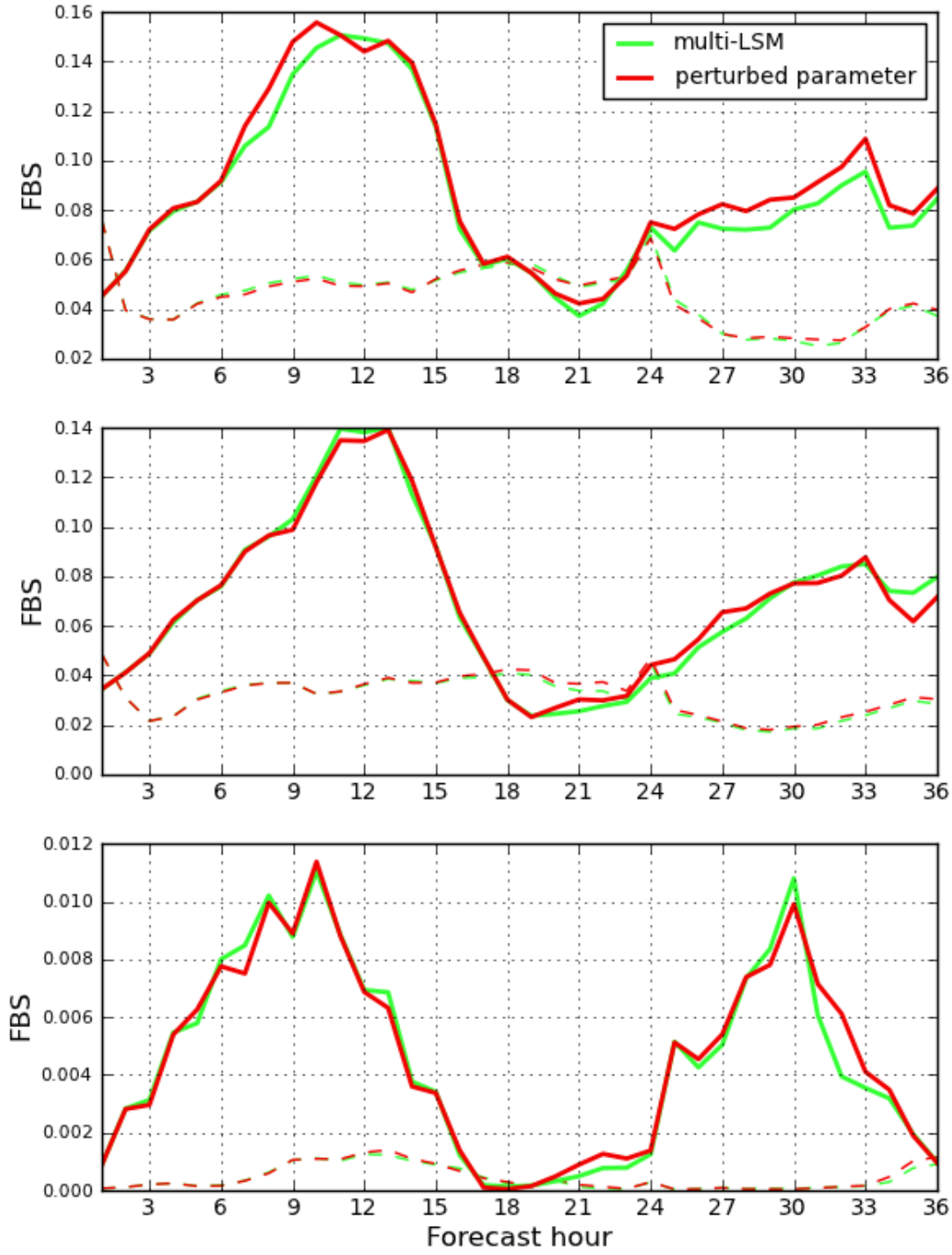

Fig. 5.39. As in Fig. 5.33 except comparing ensemble perturbation methodologies.

Chapter 6. Summary and Future Work

6.1 Summary

The purpose of this dissertation was to investigate convection-allowing ensemble design. Since such a topic is so broad, focus was directed towards model error sources. Initial and boundary condition error are equally critical components as model error in making accurate and reliable probabilistic forecasts, but detailed work on those error sources belongs to dissertations and papers by other students and researchers (e.g., Johnson and Wang 2016).

The WRF model was used as the primary tool for this investigation due to its community design and high utility. A large number of physical parameterizations have been designed by members of the atmospheric research and NWP community, and by allowing their codes to be implemented into the WRF model, a vast array of physical parameterizations is available which was ideal for this research.

The specific topics covered in this dissertation included perturbation methods for microphysics and surface-atmosphere exchange (LSM) physics, as well as addition of a physical parameterization previously missing from WRF model physics suites that is also applicable to ensemble forecasting. By "perturbation methods," I mean methods of addressing uncertainty by changing some aspect of the physics component. While a perturbation is traditionally considered in the atmospheric sciences to be a displacement from some established state, when separate microphysics schemes or LSMs are used, it can be hard to argue that one particular scheme is "established" while the other is a perturbation from it. However, the spirit of the term "perturbation" is that of a change or

difference between two schemes meant to perform the same task, which reconciles the discrepancy between the traditional definition and its use herein.

Two methods of microphysics perturbations were tested. One was called mixed microphysics and used a set of separate microphysics schemes. The schemes were among the more sophisticated available in the WRF. The other method was called perturbed parameter microphysics in which members were created by varying fixed parameters (rain and graupel size distribution intercept parameters and graupel density) within a single microphysics scheme. Results suggested that while the mixed-microphysics method tended to be somewhat better, neither method was clearly superior to the other in generating reliable and properly dispersive probabilistic forecasts.

Exploratory research on methods of perturbing the LSM component was described in this dissertation. The purpose of this work was to fill in a gap in existing research on how to best perturb the LSM component for ensemble forecasting in general. Three methods of perturbation were tested, including mixed LSM, where separate LSMs were used, a perturbed parameter approach using varied values of a fixed constant (CZIL in the thermal roughness length calculation) within a single LSM, and a second perturbed parameter approach in which formulations for physical processes were varied rather than constant parameters. A new LSM, the Noah-MP, was used for this approach. Results showed that adding LSM perturbations added diversity to forecasts already incorporating other physics perturbations. Moreover, it was shown that forecasts of deep convection can be almost as sensitive to LSM perturbations as to a combination of microphysics and PBL perturbations.

A missing physical parameterization was added to the WRF. Prior research suggested that excessive kinetic energy dissipation from three sources (numerical interpolation and diffusion, deep convection, and mountain/gravity wave drag) was occurring in the ECMWF model. The SKEB scheme was designed to randomly inject kinetic energy into the model to replace the lost energy. The inclusion of a random number process allowed testing of the effectiveness of the SKEB scheme in improving probabilistic forecasts at the convective scale. The results showed an overwhelming improvement in probabilistic forecasts that used the SKEB scheme on top of other physics perturbations, specifically through increased spread. Additionally, perturbations from the SKEB scheme could also improve forecasts above the PBL even when no physics diversity was incorporated into the ensemble design.

6.2 Overall thoughts and future work

Some additional observations and thoughts on convective-scale ensemble design are now discussed. The specific method of accounting for uncertainty may not be important. The results from Chapter 3 suggest that neither method of perturbing the microphysics scheme was dramatically better than the other as the pooled ensemble using both methods did not outperform either method. While the mixed microphysics method showed marginally better results, the difficulty associated with maintaining a large enough or diverse enough set of complex microphysics schemes may make this method difficult to implement operationally.

On a related note, other aspects of the ensemble design may be more important than the method of perturbing model physics. Using the results from the microphysics uncertainty estimation in Chapter 3, it seemed that ensemble size made a larger

difference in verification statistics than method of perturbing microphysics. In general, choosing an appropriate ensemble size for testing scientific questions about ensemble design was somewhat arbitrary. Different ensemble sizes were used in the different studies in this dissertation, but the same general principles were used to determine ensemble size. A balance was sought between available computer resources and sufficient sample size to avoid invalid results due to poor statistical robustness. The variety of available physics parameterizations was also an important limitation in studies of physics uncertainty. There are a large number of available microphysics and PBL schemes in the WRF, although not all are appropriate for convective-scale simulations. There are far fewer LSMs and radiation schemes, and only one or two of each may be appropriate, which imposes limitations on the number of ensemble members that can be constructed. It was discovered over the course of the projects comprising this dissertation that either some sets of physics combinations did not work well together or some schemes would not run successfully for certain cases. For instance, there were cases for which the ACM2 and QNSE PBL schemes caused the WRF to terminate either immediately or early in a simulation, and thus the case had to be thrown out. Also, while somewhat unrelated, Warner et al. (1997) and Nutter et al. (2004) point out that a judicious selection of model domain can alleviate problems associated with not perturbing lateral boundary conditions. Specifically, constraints on error growth are reduced on larger domains. This single fact justifies the use of as large of a domain size as is practically possible. Strictly speaking from a standpoint of model error representation, however, this fact is not relevant. But from an overall ensemble design standpoint, it is important.

Finally, including a random number process has a surprisingly positive impact on ensemble forecasts. The SKEB scheme in particular seems like a very good scheme to use to improve probabilistic forecasts at the convective scale. The one drawback is that it requires tuning, and different sets of parameters may be required for different sets of model physics. While no other stochastic model error schemes were tested in this dissertation, other research suggests the stochastically perturbed parameterization tendencies (SPPT; Palmer et al. 2009) and even the cellular automaton stochastic backscatter scheme (CASBS; Shutts 2005) may also be viable options. Perhaps more important than the specific stochastic model error scheme is simply insertion of a random number process. Also as opposed to perturbed parameter approaches with other physics components (e.g., microphysics and LSM), large ensemble membership can be generated with a fraction of the effort by using different random number sequences.

This work was part of a broader research project investigating optimal ensemble design at the convective scale. To keep the project focused, initial and lateral boundary condition error was not accounted for in any of the studies covered in this dissertation. While this was necessary to isolate the impact of methods for accounting for model error, it also means none of the results explicitly incorporated a full set of methods to account for all forecast error sources. Therefore, the results from this research underestimate the benefits of applying model perturbations, and the verifications, especially for QPFs, are likely poorer than they would be if initial and lateral boundary condition perturbation methods were also included. The results from this work were not applied to official SSEF runs for recent NOAA HWT SFEs (i.e., 2014 and 2015 program documents available at forecast.caps.ou.edu and by manipulating URLs

within), so there is currently no published research on combining the impacts from the model error perturbation methods researched herein with those from initial and lateral boundary condition perturbation methods being researched contemporaneously by others at the University of Oklahoma. Such changes should be implemented in the SSEF for upcoming SFEs so a more complete understanding of the utility of all of these methods can be realized.

The results presented in this dissertation are promising, however, as they indicate that the variety of options already present in the WRF model can be used effectively to account for model error in convection-allowing probabilistic forecasts.

References

Ahijevych, D., E. Gilleland, B. G. Brown, and E. E. Ebert, 2009: Application of spatial verification methods to idealized and NWP-gridded precipitation forecasts. *Wea. Forecasting*, **24**, 1485–1497.

Aligo, E. A., W. A. Gallus, and M. Segal, 2007: Summer rainfall forecast spread in an ensemble initialized with different soil moisture analyses. *Wea. Forecasting*, **22**, 299–314.

Anthes, R. A., 1984: Enhancement of convective precipitation by mesoscale variations in vegetative covering in semiarid regions. *J. Clim. Appl. Meteor.*, **23**, 541–554.

Baldwin, M. E., S. Lakshmivarahan, and J. S. Kain, 2001: Verification of mesoscale features in NWP models. Preprints, *Ninth Conf. on Mesoscale Processes*, Ft. Lauderdale, FL, Amer. Meteor. Soc., 255–258.

Basara, J. B., 2001: The value of point-scale measurements of soil moisture in planetary boundary layer simulations. Ph. D. dissertation, University of Oklahoma, 225 pp.

Berner, J., G. Shutts, M. Leutbecher, and T. Palmer, 2009: A spectral stochastic kinetic energy backscatter scheme and its impact on flow dependent predictability in the ECMWF Ensemble Prediction System. *J. Atmos. Sci.*, **66**, 603–626.

Berner, J., S.-Y. Ha, J. P. Hacker, A. Fournier, and C. Snyder, 2011: Model uncertainty in a mesoscale ensemble prediction system: stochastic versus multi-physics representations. *Mon. Wea. Rev.*, **139**, 1972–1995.

Berner, J., K. R. Fossell, S.-Y. Ha, J. P. Hacker, and C. Snyder, 2015: Increasing the skill of probabilistic forecasts: Understanding performance improvements from model-error representations. *Mon. Wea. Rev.*, **143**, 1295–1320.

Bouttier, F., B. Vie, O. Nuissier, and L. Raynaud, 2012: Impact of stochastic physics in a convection-permitting ensemble. *Mon. Wea. Rev.*, **140**, 3706–3721. doi: 10.1175/MWR-D-12-00031.1.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3.

Brown, J. M., and Coauthors, 2011: Improvement and testing of WRF physics options for application to Rapid Refresh and High Resolution Rapid Refresh. *Preprints, 14th Conf. on Mesoscale Processes/15th Conf. on Aviation, Range, and Aerospace Meteorology*, Los Angeles, CA, Amer. Meteor. Soc., 5.5. [Available online at https://ams.confex.com/ams/14Meso15ARAM/webprogram/Paper191234.html.]

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

Bryan, G. H., and H. Morrison, 2012: Sensitivity of a simulated squall line to horizontal resolution and parameterization of microphysics. *Mon. Wea. Rev.*, **140**, 202–225.

Buizza, R., M. Miller, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **125**, 2887–2908.

Candille, G., 2009: The multiensemble approach: The NAEFS example. *Mon. Wea. Rev.*, **137**, 1655–1665.

Charron, M., G. Pellerin, L. Spacek, P. L. Houtekamer, N. Gagnon, H. L. Mitchell, and L. Michelin, 2010: Toward random sampling of model error in the Canadian ensemble prediction system. *Mon. Wea. Rev.*, **138**, 1877–1901.

Chen, F., Z. Janjić, and K. Mitchell, 1997: Impact of atmospheric surface-layer parameterizations in the new land-surface scheme of the NCEP mesoscale Eta model. *Boundary-Layer Meteor.*, **85**, 391–421.

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

Chen, F., and Y. Zhang, 2009: On the coupling strength between the land surface and the atmosphere: From viewpoint of surface exchange coefficients. *Geophys. Res. Lett.*, **36**, L10404, doi: 10.1029/2009GL037980.

Chen, Y., K. Yang, D. Zhou, J. Qin, and X. Guo, 2010: Improving the Noah land surface model in arid regions with an appropriate parameterization of the thermal roughness length. *J. Hydrometeor.*, **11**, 995–1006, doi: http://dx.doi.org/10.1175/2010JHM1185.1.

Clark, C. A., and R. W. Arritt, 1995: Numerical simulations of the effect of soil moisture and vegetation cover on the development of deep convection. *J. Appl. Meteor.*, **34**, 2029–2045.

Clark, A.J., W.A. Gallus, and T. C. Chen, 2008: Contributions of mixed physics versus perturbed initial/lateral boundary conditions to ensemble-based precipitation forecast skill. *Mon. Wea. Rev.*, **136**, 2140–2156.

Clark, A. J., W. A. Gallus, M. Xue, and F. Kong, 2009: A comparison of precipitation forecast skill between small convection-allowing and large convection parameterizing

ensembles. *Wea. Forecasting*, **24**, 1121–1140. doi: http://dx.doi.org/10.1175/2009WAF2222222.1

Clark, A. J., W. A. Gallus, M. Xue, and F. Kong, 2010a: Growth of spread in convection-allowing and convection-parameterizing ensembles. *Wea. Forecasting*, **25**, 594–612.

Clark, A. J., W. A. Gallus, M. L. Weisman, 2010b: Neighborhood-based verification of precipitation forecasts from convection-allowing NCAR WRF model simulations and the operational NAM. *Wea. Forecasting*, **25**, 1495–1509.

Clark, A. J., and Coauthors, 2011: Probabilistic precipitation forecast skill as a function of ensemble size and spatial scale in a convection-allowing ensemble. *Mon. Wea. Rev.*, **139**, 1410–1418.

Clark, A. J., and Coauthors, 2012: An Overview of the 2010 Hazardous Weather Testbed Experimental Forecast Program Spring Experiment. *Bull. Amer. Meteor. Soc.*, **93**, 55–74. doi: http://dx.doi.org/10.1175/BAMS-D-11-00040.1

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

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

Doblas-Reyes, F. J., M. Déqué, and J.-P. Piedelievre, 2000: Multi-model spread and probabilistic seasonal forecasts in PROVOST. *Quart. J. Roy. Meteor. Soc.*, **126**, 1069–2807.

Du, J., S. L. Mullen, and F. Sanders, 1997: Short-range ensemble forecasting of quantitative precipitation. *Mon. Wea. Rev.*, **125**, 2427–2459.

Du, J., G. DiMego, B. Zhou, D. Jovic, B. Ferrier, B. Yang, and S. Benjamin, 2014: NCEP Regional Ensembles: Evolving toward hourly-updated convection-allowing scale and storm-scale predictions within a unified regional modeling system. *22nd Conf. on Numerical Weather Prediction and 26th Conf. on Weather Analysis and Forecasting*, Atlanta, GA, Amer. Meteor. Soc., Feb. 1-6, 2014, paper J1.4.

Duda, J. D., and W. A. Gallus, 2013: The impact of large-scale forcing on skill of simulated convective initiation and upscale evolution with convection-allowing grid spacings in the WRF. *Wea. Forecasting*, **28**, 994–1018.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480.

Ebert, E. E., 2009: Neighborhood verification: A strategy for rewarding close forecasts. *Wea. Forecasting*, **24**, 1498–1510.

Eckel, F. A., and C.F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting*, **20**, 328–350.

ECMWF, 2012: IFS documentation – cy37r2. Operational implementation 18 May 2011. Part V: Ensemble prediction system. [Available online at https://software.ecmwf.int/wiki/display/IFS/CY37R2+Official+IFS+Documentation]

Elmore, K. L., S. J. Weiss, and P. C. Banacos, 2003: Operational ensemble cloud model forecasts: Some preliminary results. *Wea. Forecasting*, **18**, 953–964.

Epstein, E. S., 1969: Stochastic dynamic prediction. *Tellus*, **21**, 739–759.

Ek, M. B., K. E. Mitchell, Y. Lin, E. Rogers, P. Grunmann, V. Koren, G. Gayno, and J. D. Tarplay, 2003: Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *J. Geophys. Res.*, **108,** 8851, doi:10.1029/2002JD003296.

Ferrier, B. S., Y. Jin, Y. Lin, T. Black, E. Rogers, and G. DiMego, 2002: Implementation of a new grid-scale cloud and precipitation scheme in an NCEP eta model. *Extended abstract, 19th Conf. on weather Analysis and Forecasting/15th Conf. on Numerical Weather Prediction,* San Antonio, TX, 10.1.

Ferrier, B. S., and Coauthors, 2003: Changes to the NCEP meso eta analysis and forecast system: Modified cloud microphysics, assimilation of GOES cloud-top pressure, assimilation of NEXRAD 88D radial wind velocity data. NCEP technical procedures bulletin. [Available online at http://www.emc.ncep.noaa.gov/mmb/tpb.spring03/tpb.htm.]

Ferrier. B. S., 2005: The cloud and precipitation scheme in the operational NCEP (regional) models: Description and system integration issues. Microsoft PowerPoint presentation, presented at SUNY-Stony Brook. 46 slides. [Available online at http://www.emc.ncep.noaa.gov/mmb/bf/presentations/Stony_Brook_3-1-05.ppt]

Gao, J.-D., M. Xue, K. Brewster, and K. K. Droegemeier, 2004: A three-dimensional variational data analysis method with recursive filter for Doppler radars. *J. Atmos. Oceanic Technol.*, **21**, 457–469.

Gilliam, R. C. and J. E. Pleim, 2010: Performance assessment of new land surface and planetary boundary layer physics in the WRF-ARW. *J. Appl. Meteor. Clim.*, **49**, 760–774, doi: http://dx.doi.org/10.1175/2009JAMC2126.1.

Gilmore, M. S., J. M. Straka, and E. N. Rasmussen, 2004: Precipitation uncertainty due to variations in precipitation particle parameters within a simple microphysics scheme. *Mon. Wea. Rev.*, **132**, 2610–2627.

Godfrey, C. M., D. J. Stensrud, and L. M. Leslie, 2005: The influence of improved land surface and soil data on mesoscale model predictions. *19th Conf. on Hydro.*, San Diego, CA, Amer. Meteor. Soc., paper 4.7, [Available online at https://ams.confex.com/ams/Annual2005/webprogram/Paper86050.html.]

Godfrey, C. M., and D. J. Stensrud, 2010: An empirical latent heat flux parameterization for the Noah land surface model. *J. Appl. Meteor. Clim.*, **49**, 1696–1713, doi: http://dx.doi.org/10.1175/2010JAMC2180.1.

Grimit, E. P., and C. F. Mass, 2007: Measuring the ensemble spread-error relationship with a probabilistic approach: Stochastic ensemble results. *Mon. Wea. Rev.*, **135**, 203–221.

Ha, S.-Y., J. Berner, C. Snyder, 2015: A comparison of model error representations in mesoscale ensemble data assimilation. *Mon. Wea. Rev.*, **143**, 3893–3911.

Hacker, J. P., and Coauthors, 2011: The U.S. Air Force Weather Agency's mesoscale ensemble: Scientific description and performance results. *Tellus,* **63A**, 1–17.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept. *Tellus*, **57A**, 219–233.

Hamill, T. M., and S. J. Colucci, 1997: Verification of Eta-RSM short range ensemble forecasts. *Mon. Wea. Rev.*, **125**, 1312–1327.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560.

Henderson-Sellers, A., K. McGuffie, and A. J. Pitman, 1996: The project for intercomparison of land-surface parametrization schemes (PILPS): 1992 to 1995. *Clim. Dynamics*, **12**, 849–859.

Hohenegger, C., and C. Schär, 2007: Atmospheric predictability at synoptic versus cloud-resolving scales. *Bull. Amer. Meteor. Soc.*, **88**, 1783–1793.

Hong, S.-Y. and J.-O. J. Lim, 2006a: the WRF single-moment 6-class microphysics scheme (WSM6). *J. Korean Meteor. Soc.*, **42**, 129–151.

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

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

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.*, **124**, 1225–1242.

Hu, M., M. Xue, and K. Brewster, 2006: 3DVARand cloud analysis with WSR-88D level-II data for the prediction of Fort Worth tornadic thunderstorms. Part I: Cloud analysis and its impact. *Mon. Wea. Rev.*, **134**, 675–698.

Jackson, C., Y. Xia, M. K. Sen, and P. L. Stoffa, 2003: Optimal parameter and uncertainty estimation of a land surface model: A case study using data from Cabauw, Netherlands. *J. Geophys. Res.*, **108**, 4583, doi:10.1029/2002JD002991.

Jacquemin, B., and J. Noilhan, 1990: Sensitivity study and validation of a land surface parameterization using the hapex-mobilhy data set. *Boundary-Layer Meteor.*, **52**, 93–134.

Janjić, Z., 2002: Nonsingular Implementation of the Mellor-Yamada Level 2.5 Scheme in the NCEP Meso model. NCEP Office Note #437, 61 pp. [Available online at http://www.emc.ncep.noaa.gov/officenotes/newernotes/on437.pdf.]

Johnson, A., X. Wang, F. Kong, and M. Xue, 2011a: Hierarchical cluster analysis of a convection- allowing ensemble during the Hazardous Weather Testbed 2009 Spring Experiment. Part I: Development of the object-oriented cluster analysis method for precipitation fields. *Mon. Wea. Rev.*, **139**, 3673–3693.

Johnson, A., X. Wang, F. Kong, and M. Xue, 2011b: Hierarchical cluster analysis of a convection- allowing ensemble during the Spring Experiment of the Hazardous Weather Testbed in 2009. Part II: Ensemble clustering over the whole experiment period. *Mon. Wea. Rev.*, **139**, 3694–3710.

Johnson, A. and X. Wang, 2012: Verification and Calibration of Neighborhood and Object-Based Probabilistic Precipitation Forecasts from a Multimodel Convection-Allowing Ensemble. *Mon. Wea. Rev.*, **140**, 3054–3077. doi: http://dx.doi.org/10.1175/MWR-D-11-00356.1.

Johnson, A. and X. Wang, 2013a: Object-Based Evaluation of a Storm-Scale Ensemble during the 2009 NOAA Hazardous Weather Testbed Spring Experiment. *Mon. Wea. Rev.*, **141**, 1079–1098. doi: http://dx.doi.org/10.1175/MWR-D-12-00140.1.

Johnson, A., X. Wang, F. Kong, and M. Xue, 2013b: Object-Based Evaluation of the Impact of Horizontal Grid Spacing on Convection-Allowing Forecasts. *Mon. Wea. Rev.*, **141**, 3413–3425.

Johnson, A., and Coauthors, 2014: Multiscale characteristics and evolution of perturbations for warm season convection-allowing precipitation forecasts: Dependence on background flow and method of perturbation. *Mon. Wea. Rev.,* **142**, 1053–1073. doi: http://dx.doi.org/10.1175/MWR-D-13-00204.1.

Johnson, A., X. Wang, J. Carley, L. Wicker, and C. Karstens, 2015: A comparison of multi-scale GSI-based EnKF and 3DVar data assimilation using radar and conventional observations for mid-latitude convective-scale precipitation forecasts. *Mon. Wea. Rev.*, **143**, 3087–3108.

Johnson and Wang, 2016: A study of multi-scale initial condition perturbation methods for convection-permitting ensemble forecasts. *Mon. Wea. Rev.*, in press.

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

Kong, F., K. K. Droegemeier, and N. L. Hickmon, 2006: Multi-resolution ensemble forecasts of an observed tornadic thunderstorm system, Part I: Comparison of coarse and fine-grid experiments. *Mon. Wea. Rev.*, **134,** 807–833.

Kong, F., K. K. Droegemeier, and N. L. Hickmon, 2007a: Multiresolution ensemble forecasts of an observed tornadic thunderstorm system, Part II. Storm scale experiments. *Mon. Wea. Rev.*, **135,** 759–782.

Kong, F., and Coauthors, 2007b: Preliminary analysis on the real-time storm-scale ensemble forecasts produced as a part of the NOAA hazardous weather testbed 2007 spring experiment. *22nd Conf. Wea. Anal. Forecasting/18th Conf. Num. Wea. Pred.*, Salt Lake City, Utah, Amer. Meteor. Soc., Paper 3B.2.

Kong, F., and Coauthors, 2009: A real-time storm-scale ensemble forecast system: 2009 Spring Experiment. *Preprints, 23nd Conf. on Weather Analysis and Forecasting/19th Conf. Num. Wea. Pred.,* Omaha, Nebraska, Amer. Meteor. Soc., Paper 16A3.

Kong, F., and Coauthors, 2010: Evaluation of CAPS multi-model storm-scale ensemble forecast for the NOAA HWT 2010 Spring Experiment. *25th Conf. Severe Local Storms*, Amer. Meteor. Soc., Paper P4.18.

Kong, F., and Coauthors, 2011a: CAPS Multi-Model Storm-Scale Ensemble Forecast for the NOAA HWT 2010 Spring Experiment. *24th Conf. Wea. Forecasting/20th Conf. Num. Wea. Pred.,* Amer. Meteor. Soc., Paper 457.

Kong, F., and Coauthors, 2011b: Storm-scale ensemble forecasting for the NOAA hazardous weather testbed. *Extended Abstract, 6th European Conference on Severe Storms*, Palma de Mallorca, Spain. [Available online at http://www.essl.org/ECSS/2011/programme/abstracts/171.pdf.]

Kumar, A., F. Chen, D. Niyogi, J. G. Alfieri, M. Ek, and K. Mitchell, 2011: Evaluation of a photosynthesis-based canopy resistance formulation in the Noah land-surface model. *Boundary-Layer Meteor.*, **138**, 263–284, doi: 10.1007/s10546-010-9559-z.

Kurkowski, N. P., D. J. Stensrud, and M. E. Baldwin, 2003: Assessment of implementing satellite-derived land cover data in the Eta model. *Wea. Forecasting*, **18**, 404–416, doi: http://dx.doi.org/10.1175/1520-0434(2003)18<404:AOISDL>2.0.CO;2.

LeMone, M. A., M. Tewari, F. Chen, J. G. Alfieri, and D. Niyogi, 2008: Evaluation of the Noah land surface model using data from a fair-weather IHOP_2002 day with heterogeneous surface fluxes. *Mon. Wea. Rev.*, **136**, 4915–4941, doi: http://dx.doi.org/10.1175/2008MWR2354.1.

LeMone, M. A., F. Chen, M. Tewari, J. Dudhia, B. Geerts, Q. Miao, R. L. Coulter, and R. L. Grossman, 2010: Simulating the IHOP_2002 fair-weather CBL with the WRF-ARW-Noah modeling system. Part I: Surface fluxes and CBL structure and evolution along the eastern track. *Mon. Wea. Rev.*, **138**, 722–744, doi: http://dx.doi.org/10.1175/2009MWR3003.1.

Leoncini, G., R. S. Plant, S. L. Gray, and P. A. Clark, 2010: Perturbation growth at the convective scale for CSIP IOP18. *Q J. R. Meteor. Soc.*, **136**, 653–670.

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. Wea. Rev.*, **138**, 1587–1612.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21**, 289–307.

Lorenz, E. N., 1975: Climatic predictability. In *The Physical Basis of Climate and Climate Modelling*, WMO GARP Publ. Ser. No. 16, pp. 132–36. Geneva: World Meteorological Organization, 265 pp.

Lorenz, E. N., 1993: *The essence of chaos.* Univ. of Washington Press, 227 pp.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418.

Leutbecher, M., and T. N. Palmer, 2008: Ensemble forecasting. *J. Comp. Phys.*, **227**, 3515–3539.

Lewis, J. M., 2005: Roots of ensemble forecasting. *Mon. Wea. Rev.*, **133**, 1865–1885.

Mahrt, L., and M. Ek, 1984: The Influence of Atmospheric Stability on Potential Evaporation. *J. Climate Appl. Meteor.*, **23**, 222–234, doi: http://dx.doi.org/10.1175/1520-0450(1984)023<0222:TIOASO>2.0.CO;2.

Mahrt, L., and H.-L. Pan, 1984: A two-layer model of soil hydrology. *Boundary-Layer Meteorol.*, **29**, 1–20.

Markowski, P., C. Hannon, and E. Rasmussen, 2006: Observations of convection initiation "failure" from the 12 June 2002 IHOP deployment. *Mon. Wea. Rev.*, **134**, 375–405. doi: http://dx.doi.org/10.1175/MWR3059.1.

Marshall, C. H., K. C. Crawford, K. E. Mitchell, and D. J. Stensrud, 2003: The impact of the land surface physics in the operational NCEP Eta model on simulating the diurnal cycle: Evaluation and testing using Oklahoma mesonet data. *Wea. Forecasting*, **18**, 748–768, doi: http://dx.doi.org/10.1175/1520-0434(2003)018<0748:TIOTLS>2.0.CO;2.

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

Mason, P., and D. Thomson, 1992: Stochastic backscatter in large eddy simulations of boundary layers. *J. Fluid Mech.*, **242**, 51–78.

Mass, C. F., D. Owens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.*, **83**, 407–430.

Mellor, G.L. and T. Yamada, 1982: Development of a turbulence closure model for geophysical fluid problems. *Rev. Geophys. Space Phys.*, **20**, 851–875.

Milbrandt, J. A. and M. K. Yau, 2005: A multimoment bulk microphysics parameterization. Part II: A proposed three-moment closure and scheme description. *J. Atmos. Sci.*, **62**, 3065–3081.

Miller, J., M. Barlage, X. Zeng, H. Wei, K. Mitchell, and D. Tarpley, 2006: Sensitivity of the NCEP/Noah land surface model to the MODIS green vegetation fraction data set. *Geophys. Res. Lett.*, **33**, L13404, doi:10.1029/2006GL026636.

Mittermaier, M. P., 2007: Improving short-range high-resolution model precipitation forecast skill using time-lagged ensembles. *Quart. J. Roy. Meteor. Soc.*, **133**, 1487–1500.

Molinari, J. and M. Dudek, 1992: Parameterization of convective precipitation in mesoscale numerical models: A critical review. *Mon. Wea. Rev.*, **120**, 326–344.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: methodology and validation. *Quart. J. Roy. Meteor. Soc*., **122**, 73–119.

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. Wea. Rev.*, **137**, 991–1007.

Morrison, H., and J. Milbrandt, 2011: Comparison of two-moment bulk microphysics schemes in idealized supercell thunderstorm simulations. *Mon. Wea. Rev.*, **139**, 1103–1130.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293.

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

Nastrom, G. D., and K. S. Gage, 1985: A climatology of atmospheric wavenumber spectra of wind and temperature observed by commercial aircraft. *J. Atmos. Sci.*, **42**, 950–960. doi: http://dx.doi.org/10.1175/1520-0469(1985)042<0950:ACOAWS>2.0.CO;2

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

Noilhan, J., and S. Planton, 1989: A Simple Parameterization of Land Surface Processes for Meteorological Models. *Mon. Wea. Rev.*, **117**, 536–549, doi: http://dx.doi.org/10.1175/1520-0493(1989)117<0536:ASPOLS>2.0.CO;2.

Nutter, P., D. Stensrud, and M. Xue, 2004: Effects of coarsely resolved and temporally interpolated lateral boundary conditions on the dispersion of limited-area ensemble forecasts. *Mon. Wea. Rev.*, **132**, 2358–2377.

Palmer, T. N., R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. J. Shutts, M. Steinheimer, and A. Weisheimer, 2009: Stochastic parametrization and model uncertainty. ECMWF technical memo 598. [Available online at http://www.ecmwf.int/sites/default/files/elibrary/2009/11577-stochastic-parametrization-and-model-uncertainty.pdf.]

Pan, H.-L., and L. Mahrt, 1987: Interaction between soil hydrology and boundary layer development. *Boundary-Layer Meteorol.*, **38**, 185–202.

Pielke, R. A., 2001: Influence of the spatial distribution of vegetation and soils on the prediction of cumulus convective rainfall. *Rev. Geophys.*, **39**, 151–177, doi:10.1029/1999RG000072.

Pleim, J. E., 2007: A combined local and nonlocal closure model for the atmospheric boundary layer. Part I: Model description and testing. *J. Appl. Meteor. Clim.*, **46**, 1383–1395, doi: http://dx.doi.org/10.1175/JAM2539.1.

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

Rabin, R. M., S. Stadler, P. J. Wetzel, D. J. Stensrud, and M. Gregory, 1990: Observed effects of landscape variability on convective clouds. *Bull. Amer. Meteor. Soc.*, **71**, 272–280.

Ralph, F. M., and Coauthors, 2005: Improving short-term (0-48 h) cool-season quantitative precipitation forecasts. *Bull. Amer. Meteor. Soc.*, **86**, 1619–1632.

Reisner, J., R. M. Rasmussen, and R. T. Bruintjes, 1998: Explicit forecasting of supercooled liquid water in winter storms using the MM5 mesoscale model. *Q. J. R. Meteor. Soc.*, **124**, 1071–1107.

Richardson, D. S., 2001: Skill and economic value of the ECMWF ensemble prediction system, *Q. J. R. Meteorol. Soc*., **126**, 649–668.

Roberts, N. M. and H. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97.

Robock, A., and Coauthors, 2003: Evaluation of the North American land data assimilation system over the southern Great Plains during the warm season. *J. Geophys. Res.*, **108**, 8846, doi:10.1029/2002JD003245.

Romine, G. S., C. S. Schwartz, 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. Wea. Rev.*, **142**, 4519–4541.

Ropnack, A., A. Hense, C. Gebhardt, and D. Majewski, 2013: Bayesian model verification of NWP ensemble forecasts. *Mon. Wea. Rev.*, **141**, 375–387.

Rotunno, R., J. B. Klemp, and M. L. Weisman, 1988: A theory for strong, long-lived squall lines. *J. Atmos. Sci.*, **45**, 463–485.

Rotunno, R., and C. Snyder, 2008: A generalization of Lorenz's model for the predictability of flows with many scales of motion. *J. Atmos. Sci.*, **65**, 1063–1076.

Sathye, A., G. Bassett, K. Droegemeier, M. Xue, and K. Brewster, 1996: Experiences using high performance computing for operational storm scale weather prediction, In:

Concurrency: Practice and Experience, special issue on Commercial and industrial Applications on High Performance Computing. John Wiley & Sons, Ltd., 731-740.

Schwartz, C. S., and Coauthors, 2010: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*, **25**, 263–280.

Segele, Z. T., D. J. Stensrud, I. C. Ratcliffe, and G. M. Henebry, 2005: Influence on a hailstreak on boundary layer evolution. *Mon. Wea. Rev.*, **133**, 942–960, doi: http://dx.doi.org/10.1175/MWR2897.1.

Shutts, G. J. 2005. A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Quart. J. R. Meteor. Soc.* **612**, 3079–3102.

Shutts, G. J., and T. N. Palmer, 2007: Convective forcing fluctuations in a cloud-resolving model: Relevance to the stochastic parameterization problem. *J. Climate,* **20**, 187–202.

Skamarock, W. C., 2004: Evaluating mesoscale NWP models using kinetic energy spectra. *Mon. Wea. Rev.*, **132**, 3019–3032.

Skamarock, W. C., and Coauthors, 2008: A description of the advanced research WRF version 3. NCAR technical note NCAR/TN-475+STR. 113 pp.

Smirnova, T. G., J. M. Brown, and S. G. Benjamin, 1997: Performance of different soil model configurations in simulating ground surface temperature and surface fluxes. *Mon. Wea. Rev.*, **125**, 1870–1884, doi: http://dx.doi.org/10.1175/1520-0493(1997)125<1870:PODSMC>2.0.CO;2.

Smirnova, T. G., J. M. Brown, S. G. Benjamin, and D. Kim, 2000: Parameterization of cold-season processes in the MAPS land-surface scheme. *J. Geophys. Res.*, **105**, 4077–4086, doi:10.1029/1999JD901047.

Smirnova, T. G., J. M. Brown, S. G. Benjamin, and J. S. Kenyon, 2016: Modifications to the rapid update cycle land surface model (RUC LSM) available in the weather research and forecast (WRF) model. *Mon. Wea. Rev.*, in press.

Snook, N., and M. Xue, 2008: Effects of microphysical drop size distribution on tornadogenesis in supercell thunderstorms. *J. Geo. Res. Lett.*, **35**, 5 pp. doi: 10.1029/2008GL035866.

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. Wea. Rev.*, **128**, 2077–2107.

Stensrud, D.J., and N. Yussouf, 2003: Short-range ensemble predictions of 2-m temperature and dewpoint temperature over New England. *Mon. Wea. Rev.*, **131**, 2510–2524.

Stensrud, D. J., and N. Yussouf, 2007: Reliable probabilistic quantitative precipitation forecasts from a short-range ensemble forecasting system. *Wea. Forecasting*, **22**, 3–17.

Sutton, C., T. M. Hamill, and T. T. Warner, 2006: Will perturbing soil moisture improve warm-season ensemble forecasts? A proof of concept. *Mon. Wea. Rev.*, **134**, 3174–3189.

Tennant, W. J., G. J. Shutts, A. Arribas, and S. A. Thompson, 2011: Using a stochastic kinetic energy backscatter scheme to improve MOGREPS probabilistic forecast skill. *Mon. Wea. Rev.*, **139**, 1190–1206.

Testud, J., S. Oury, R. A. Black, P. Amayenc, and X. Dou, 2001: The concept of "normalized" distribution to describe raindrop spectra: A tool for cloud physics and cloud remote sensing. *J. Appl. Meteor.*, **40**, 1118–1140.

Theis, S. E., A. Hense, and U. Damrath, 2005: Probabilistic precipitation forecasts from a deterministic model: A pragmatic approach. *Meteor. Appl.*, **12**, 257–268.

Thompson, P., 1957: Uncertainty of initial state as a factor in predictability of large-scale atmospheric flow patterns. *Tellus*, **9**, 275–295.

Thompson, G., P. R. Field, R. M. Rasmussen, and W. D. Hall, 2008: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Mon. Wea. Rev.*, **136**, 5095–5115.

Tong, M., and M. Xue, 2008: 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. Wea. Rev.*, **136**, 1630–1648.

Toth, Z, Y. Zhu, and T. Marchok, 2001: The use of ensembles to identify forecasts with small and large uncertainty. *Wea. Forecasting*, **16**, 465–477.

Tracton, M. S. and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Wea. Forecasting*, **8**, 379–398.

Trier, S. B., F. Chen, and K. W. Manning, 2004: A study of convection initiation in a mesoscale model using high-resolution land-surface initiation conditions. *Mon. Wea. Rev.*, **132**, 2954–2976, doi: http://dx.doi.org/10.1175/MWR2839.1.

Trier, S. B., F. Chen, K. W. Manning, M. A. LeMone, and C. A. Davis, 2008: Sensitivity of the PBL and precipitation in 12-day simulations of warm-season

convection using different land surface models and soil wetness conditions. *Mon. Wea. Rev.*, **136**, 2321–2343, doi: http://dx.doi.org/10.1175/2007MWR2289.1.

Trier, S. B., M. A. LeMone, F. Chen, and K. W. Manning, 2011: Effects of surface heat and moisture exchange on ARW-WRF warm-season precipitation forecasts over the central United States. *Wea. Forecasting*, **26**, 3–25, doi: http://dx.doi.org/10.1175/2010WAF2222426.1.

Trier, S. B., G. S. Romine, D. A. Ahijevych, R. J. Trapp, R. S. Schumacher, M. C. Coniglio, and D. J. Stensrud, 2015: Mesoscale thermodynamic influences on convection initiation near a surface dryline in a convection-permitting ensemble. *Mon. Wea. Rev.*, **143**, 3726–3752, doi: http://dx.doi.org/10.1175/MWR-D-15-0133.1.

Vie., B., O. Nuissier, and V. Ducrocq, 2011: Cloud-resolving ensemble simulations of Mediterranean heavy precipitating events: Uncertainty on initial conditions and lateral boundary conditions. *Mon. Wea. Rev.*, **139**, 403–423.

Wandishin, M. S., S. L. Mullen, D. J. Stensrud, and H. E. Brooks, 2001: Evaluation of a short-range multimodel ensemble system. *Mon. Wea. Rev.*, **129**, 729–747.

Wang, X., and C. H. Bishop, 2003: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.*, **60**, 1140-1158.

Wang, X., and C. H. Bishop, 2005: Improvement of ensemble reliability with a new dressing kernel. *Q. J. R. Meteor. Soc.*, **131**, 965–986.

Warner, T. T., R. A. Peterson, and R. E. Treadon, 1997: A tutorial on lateral boundary conditons as a basic and potentially serious limitation to regional numerical weather prediction. *Bull. Amer. Meteor. Soc.*, **78**, 2599–2617.

Weisman, M. L., W. C. Skamarock, and J. B. Klemp, 1997: The resolution dependence of explicitly modeled convective systems. *Mon. Wea. Rev.*, **125**, 527–548.

Weygandt, S. S., and Coauthors, 2011: Evaluation of the National Center for Environmental Prediction (NCEP) implementation version of the Rapid Refresh and its skill in providing short-term guidance for aviation hazards. *Preprints, 15th Conf. on Aviation, Range, and Aerospace Meteorology*, Los Angeles, CA, Amer. Meteor. Soc., 5.4. [Available online at https://ams.confex.com/ams/14Meso15ARAM/webprogram/Paper191213.html.]

Xue, M., D.-H. Wang, J.-D. Gao, K. Brewster, and K. K. Droegemeier, 2003: The Advanced Regional Prediction System (ARPS), storm-scale numerical weather prediction and data assimilation. *Meteor. Atmos. Phys.*, **82**, 139–170.

Xue, M. and W. J. Martin, 2006: A high-resolution modeling study of the 24 May 2002 dryline case during IHOP. Part II: Horizontal convective rolls and convective initiation. *Mon. Wea. Rev.*, **134**, 172–191. doi: http://dx.doi.org/10.1175/MWR3072.1.

Xue, M., and Coauthors, 2007: CAPS realtime storm-scale ensemble and high-resolution forecasts as part of the NOAA Hazardous Weather Testbed 2007 spring experiment. *22nd Conf. Wea. Anal. Forecasting/18th Conf. Num. Wea. Pred.*, Amer. Meteor. Soc., CDROM 3B.1.

Xue, M., and Coauthors, 2008: CAPS realtime storm-scale ensemble and high-resolution forecasts as part of the NOAA Hazardous Weather Testbed 2008 Spring Experiment. *24th Conf. Several Local Storms*, Savannah, GA, Ameri. Meteor. Soc., Paper 12.2. [Available online at http://twister.ou.edu/papers/Xue_24SLS_conf_2008SpringExperiment.pdf.]

Xue, M., and Coauthors, 2009: CAPS realtime multi-model convection-allowing ensemble and 1-km convection-resolving forecasts for the NOAA Hazardous Weather Testbed 2009 Spring Experiment. *23nd Conf. Wea. Anal. Forecasting/19th Conf. Num. Wea. Pred.*, Omaha, NB, Amer. Meteor. Soc., Paper 16A.2.

Xue, M., and Coauthors, 2011: CAPS Realtime Storm Scale Ensemble and High Resolution Forecasts for the NOAA Hazardous Weather Testbed 2010 Spring Experiment. *24th Conf. Wea. Forecasting/20th Conf. Num. Wea. Pred.*, Amer. Meteor. Soc., Paper 9A.2.

Xue, M., F. Kong, K. A. Brewster, K. W. Thomas, J. Gao, Y. Wang, and K. K. Droegemeier, 2013: Prediction of convective storms at convection-resolving 1 km resolution over continental United States with radar data assimilation: An example case of 26 May 2008 and precipitation forecasts from spring 2009. *Adv. Meteor.*, **2013,** Article ID 259052, doi:10.1155/2013/259052.

Yussouf, N., and D. J. Stensrud, 2012: Comparison of single-parameter and multiparameter ensembles for assimilation of radar observations using the ensemble Kalman filter. *Mon. Wea. Rev.*, **140**, 562–586. doi: http://dx.doi.org/10.1175/MWR-D-10-05074.1

Zhang, J., and Coauthors, 2011: National Mosaic and Multisensor QPE (NMQ) system: Description, results, and future plans. *Bull. Amer. Meteor. Soc.*, **92**, 1321–1338, doi:10.1175/2011BAMS-D-11-00047.1.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, K. Mylne, 2002: The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc.*, **83**, 73–83.

Zilitinkevich, S. S., 1995: Non-local turbulent transport: Pollution dispersion aspects of coherent structure of convective flows. *Air Pollution III – Volume I. Air Pollution Theory and Simulation*, Computational Mechanics Publications, 53–60.

Appendix A – List of abbreviations

ACM2	Asymmetric convective model version 2
ARW	Advanced Research WRF
BS	Brier score
BSS	Brier skill score
CAPE	Convective available potential energy
CAPS	Center for Analysis and Prediction of Storms
CASBS	Cellular automaton stochastic backscatter scheme
CIN	Convective inhibition
CONUS	Contiguous (lower 48) United States
ECMWF	European Centre for Medium-range Weather Forecasts
EFS	Ensemble forecast system (see also EPS)
EPS	Ensemble prediction system (also see EFS)
ETS	Equitable threat score
FBS	Fractions Brier score
FSS	Fractions skill score
HMUH	Hourly-maximum updraft helicity
HWT	Hazardous Weather Testbed
KE	Kinetic energy
LFC	Level of free convection
LSM	Land-surface model
MADIS	Meteorological Assimilation Data Ingest System
MCS	Mesoscale convective system

METAR Meteorological Terminal Air Report

MLCAPE Mixed-layer CAPE

MLCIN Mixed-layer CIN

M-Y Milbrandt-Yau

MYJ Mellor-Yamada-Janjić

MYNN Mellor-Yamada-Nakanishi-Niino

NAM North American Mesoscale

NCEP National Center for Environmental Prediction

NETS Neighborhood ETS

NICS National Institute of Computational Science

NMQ National mosaic and multi-sensor QPE

NOAA National Oceanic and Atmospheric Administration

NSF National Science Foundation

NWP Numerical weather prediction

OFB Outflow boundary

OSCER University of Oklahoma Supercomputing Center for Education and Research

PBL Planetary boundary layer

POD Probability of detection

POFD Probability of false detection

PoP Probability of precipitation

PSD Particle size distribution

PQPF Probabilistic QPF

PX	Pleim-Xiu
QNSE	Quasi-normal scale elimination
QPE	Quantitative precipitation estimate
QPF	Quantitative precipitation forecast
RAP	Rapid Refresh
RHS	Right hand side
RMSE	Root mean square error
ROC	Receiver (or relative) operating characteristic
RUC	Rapid Update Cycle
SBCAPE	Surface-based CAPE
SBCIN	Surface-based CIN
SFE	Spring forecasting experiment
SKEB	Stochastic kinetic energy backscatter
SREF	Short-range ensemble forecast system
SSBS	Spectral stochastic kinetic energy backscatter
SSEF	Storm-scale ensemble forecast system
USGS	United States Geological Survey
WDM6	WRF double moment 6-class
WSM6	WRF single moment 6-class
WRF	Weather Research and Forecasting
XSEDE	Extreme Science and Engineering Discovery Environment
YSU	Yonsei University