UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

OPTIMAL DESIGN OF A MULTI-SCALE ENSEMBLE SYSTEM FOR CONVECTIVE SCALE PROBABILISTIC FORECASTS: DATA ASSIMILATION AND INITIAL CONDITION PERTURBATION METHODS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

AARON JOHNSON
Norman, Oklahoma
2014

OPTIMAL DESIGN OF A MULTI-SCALE ENSEMBLE SYSTEM FOR CONVECTIVE SCALE PROBABILISTIC FORECASTS: DATA ASSIMILATION AND INITIAL CONDITION PERTURBATION METHODS

A DISSERTATION APPROVED FOR THE
SCHOOL OF METEOROLOGY

BY

_____
Dr. Xuguang Wang, Chair

_____
Dr. Michael Richman

_____
Dr. Ming Xue

_____
Dr. David Stensrud

_____
Dr. Yang Hong

_____
Dr. Sridhar Radhakrishnan

Acknowledgements

This research was made possible thanks to the encouragement and guidance of my advisor, Dr. Xuguang Wang, and other committee members, Dr. Michael Richman, Dr. Ming Xue, Dr. David Stensrud, Dr. Yang Hong and Dr. Sridhar Radhakrishnan. The development of reflectivity assimilation in the GSI software was facilitated by collaboration with Dr. Jacob Carley at NCEP. Quality controlled NEXRAD radar observations were obtained through collaboration with Dr. Louis Wicker and Dr. Chris Karstens at NSSL.

This research was primarily supported by NSF awards AGS-1046081 and AGS-1359703. Some of the computing was performed at the OU Supercomputing Center for Education and Research (OSCER) at the University of Oklahoma. David Akin, Senior System Administrator at OSCER, provided valuable contributions to debugging and running some of the software on OSCER machines. This work also used the Extreme Science and Engingeering Discovery Environment (XSEDE), which is supported by NSF grant ACI-1053575, and the Yellowstone machine (ark:/85065/d7wd3xhc) at NCAR's Computational and Information Systems Laboratory, sponsored by NSF. Ensemble forecasts generated during the NOAA HWT Spring Experiments were produced under the support from the NOAA CSTAR program (Grant NA17RJ1227) and NSF grant AGS-0802888.

Table of Contents

List of Tables

List of Figures

Abstract

This dissertation research was undertaken to better understand the optimal design of convection-permitting storm-scale ensemble forecast (SSEF) systems that resolve features ranging from synoptic to convective scales. The focus of the research is on the data assimilation (DA) and initial condition (IC) perturbation methods, both of which are uniquely affected by the multi-scale interactions inherent in an SSEF system. There are four components to this research. First, a GSI-based DA system is implemented in the multi-scale scenario with observations ranging from synoptic scale rawinsonde to convective scale radar observations. The GSI-based 3DVar and EnKF techniques are also compared to each other in this multi-scale context. Second, the systematic sensitivities of convection forecasts to different simple methods of IC perturbation are evaluated. Third, Observation System Simulation Experiments (OSSES) are conducted using ensemble analyses generated with the GSI-based EnKF to understand the impacts of different methods of generating more complex flow-dependent multi-scale IC perturbations. Fourth, the impacts of inconsistencies between the initial and lateral boundary condition (LBC) perturbations are evaluated as well as the impacts of model and physics errors in non-OSSE real-data experiments.

In the first part of this research, the multi-scale GSI-based EnKF and 3DVar techniques are systematically compared to each other to better understand the impacts of their differences on the analyses at multiple scales and the subsequent convective scale probabilistic forecasts. Averaged over ten diverse cases, 8h forecasts of hourly accumulated precipitation initialized using GSI-based EnKF are more skillful than those initialized using GSI-based 3DVar, both with and without storm-scale radar DA. The

advantage from radar DA persists for ~5h using EnKF, but only ~1h using 3DVar. A case study of an upscale growing MCS is also examined. The better EnKF-initialized forecast is attributed to more accurate analyses of both the mesoscale environment and the storm scale features. The mesoscale location and structure of a warm front is more accurately analyzed using EnKF than 3DVar. Furthermore, storms in the EnKF multi-scale analysis are maintained during the subsequent forecast period. However, storms in the 3DVar multi-scale analysis are not maintained and generate excessive cold pools. Therefore, while the EnKF forecast with radar DA remains better than the forecast without radar DA throughout the forecast period, the 3DVar forecast quality is degraded by radar DA after the first hour. Diagnostics revealed that the inferior analysis at meso- and storm-scales for the 3DVar is primarily due to the lack of flow-dependence and coherent cross-variable correlation, respectively, in the 3DVar static background error covariance.

In the second part of this research, multi-scale precipitation forecast sensitivities are examined for two events and systematically over 34 events out to 30-h lead time using Haar Wavelet decomposition of hourly accumulated precipitation. The impacts of two small scale IC perturbation methods are compared to the larger scale IC and physics perturbations included in an experimental convection-allowing ensemble. For an event where the forecast precipitation is driven primarily by a synoptic scale baroclinic disturbance, small scale IC perturbations result in little precipitation forecast perturbation energy on medium and large scales, compared to larger scale IC and physics (LGPH) perturbations after the first few forecast hours. However, for an event where forecast convection at the initial time grows upscale into a Mesoscale Convective

System (MCS), small scale IC and LGPH perturbations result in similar forecast perturbation energy on all scales after about 12 hours. Averaged over 34 forecasts, the small scale IC perturbations have little impact on large forecast scales while LGPH accounts for about half of the error energy on such scales. The impact of small scale IC perturbations is also less than, but comparable to, the impact of LGPH perturbations on medium scales. On small scales, the impact of small scale IC perturbations is at least as large as the LGPH perturbations. The spatial structure of small scale IC perturbations also affects the evolution of forecast perturbations, especially at medium scales. For these random homogeneous small scale IC perturbations, there is little additional impact of the small scale IC perturbations when added to LGPH. Additional study of more realistic flow-dependent IC perturbations, and their impacts on ensemble forecast skill in addition to deterministic forecast sensitivity, are therefore motivated.

In the third part of this research, the impacts of multi-scale flow-dependent IC perturbations (MULTI) for SSEFs are investigated using perfect model OSSEs. The MULTI perturbations are compared to downscaled IC perturbations from a larger scale ensemble (LARGE). Forecasts initialized at different stages of the upscale growth of an MCS case study are first used to qualitatively understand the impacts of the IC perturbation methods. Scale-dependence of the results is assessed by evaluating two-hour storm-scale reflectivity forecasts in 0-48km neighborhoods separately from hourly accumulated precipitation forecasts in mesoscale neighborhoods with a 48-km radius. For the reflectivity forecasts over small neighborhood radii (0-8km), the small scales of IC perturbation, resolved in MULTI but not LARGE, are advantageous for about 1h. For reflectivity forecasts at larger radii and for mesoscale precipitation forecasts, the

differences in IC perturbations on scales resolved by both MULTI and LARGE dominate the forecast skill. The MULTI IC perturbations are more consistent with the analysis uncertainty than the LARGE IC perturbations in the vicinity of the developing MCS. However, an area of spurious convection away from the observed MCS contains unrealistically large mid-level moisture perturbations for MULTI that can have the effect of enhancing the spurious convection. The relative importance of these differences between MULTI and LARGE, and their effects on forecast skill, depends on when during the MCS upscale growth process the forecasts are initialized.

The perfect-model OSSE case study is also extended to 11 diverse cases. The mesoscale precipitation forecasts from MULTI are systematically more skillful than LARGE at 1h and ~5-9h lead times. This is due to the smaller magnitude mesoscale IC perturbations near analyzed convective systems for MULTI that are more consistent with the analysis uncertainty than for LARGE. This difference also leads to systematically more skillful reflectivity forecasts for MULTI than LARGE using radii >4km. The reflectivity forecasts using radii of 0-4km are systematically more skillful for MULTI than LARGE during the first hour due to the presence of the small scale IC perturbations. The small scale IC perturbations also systematically contribute to further improving the MULTI mesoscale precipitation forecasts after ~5h.

In the final part of this research, two considerations for operational application of the multi-scale IC perturbation methods are investigated. First, the impact of inconsistencies between the multi-scale IC perturbations and mesoscale LBC perturbations is evaluated. Spurious pressure waves originating at the LBCs result from this inconsistency. However, unlike previous studies with a larger resolution difference

between then inner and outer domains and with different DA methods on each domain, significant impacts on convective scale probabilistic forecast skill are not found with the multi-scale GSI-based DA system. Second, real-data experiments with model error are used to further understand the practical implications of the OSSE results. In real-data experiments, LARGE is generally more skillful than MULTI except for reflectivity forecasts at short lead times of ~30-90 minutes, depending on spatial scale. This is because the larger magnitude mesoscale IC perturbations in LARGE compensate for unrepresented model errors. The flow-dependent small-scale IC perturbations are even more important for storm-scale reflectivity forecasts in the ensemble with unrepresented model error than in the perfect-model OSSEs.

Chapter 1: Introduction

1.1 Background and motivation

Convection-permitting (i.e., grid spacing of 1~4 km and no cumulus parameterization) forecasts have shown advantages over coarser resolution models for many different convective-scale forecasting applications. Such applications include the climatological diurnal cycle of convective precipitation events (Clark et al. 2007), mesoscale quantitative precipitation forecasting (Clark et al. 2009, 2012; Johnson and Wang 2012; Duc et al. 2013) and short lead time high impact severe weather events (Stensrud et al. 2009; Yussouf et al. 2013). Convection-permitting models have therefore been widely used in storm-scale ensemble forecasting (SSEF) systems (e.g., Clark et al. 2012). However, compared to global and mesoscale ensemble systems there has been relatively little systematic study of how to optimally design SSEFs (e.g., Clark et al. 2010, 2011; Johnson et al. 2011a,b). A few early studies have investigated the impacts of ensemble size (Clark et al. 2011) and the impacts of different sources of diversity on the ensemble spread (Clark et al. 2010; Johnson et al. 2011b). More extensive research on the optimal design of SSEF systems is still needed.

The optimal design of SSEF systems remains largely unknown, although coarser resolution ensembles have been relatively well studied. For example, medium range (~1 week) synoptic scale (~100 km grid spacing) ensembles have been studied for about two decades (Buizza and Palmer 1995; Toth and Kalnay 1997; Houtekamer et al. 1996; Wang and Bishop 2003; Wang et al. 2004). Short-range (~1-3 days) mesoscale (~10-20 km) ensembles have also been the focus of many past studies (Du et al. 1997; Stensrud

et al. 1999; Marsigli et al. 2001; Xu et al. 2001; Grimit and Mass 2002; Eckel and Mass 2005; Lu et al 2007; Li et al. 2008; Berner et al. 2011). However, the optimal design of SSEFs may be quite different than that of coarser resolution ensembles (Hohenegger and Schär 2007b).

Convective precipitation forecasting is an inherently multi-scale challenge because of the broad range of spatial scales impacting the initiation and evolution of convective systems (Lorenz 1969; Perkey and Maddox 1985; Zhang et al. 2007; Rotunno and Snyder 2008; Johnson et al. 2015). Convective systems are not only strongly influenced by the larger scale environment within which they occur, but also in turn impact the larger scale environment itself (e.g., Perkey and Maddox 1985). This sensitivity to features on multiple scales, and their interactions, has important implications for the design of SSEF systems at convection-permitting resolution. For example, since an accurate forecast depends on accurate initial condition (IC) analyses for features on all resolvable spatial scales, appropriate multi-scale data assimilation (DA) systems are needed. Furthermore, since some amount of IC error is inevitable, the ensemble IC perturbations should accurately sample the effects of the multi-scale IC analysis errors on the forecast uncertainty. There is a natural link between ensemble forecasting and ensemble DA. Ensemble DA provides an analysis ensemble that can provide ensemble IC perturbations. Meanwhile, the ensemble forecasts initialized from such perturbations can provide flow dependent background error covariance information to an ensemble DA system.

1.1a Data assimilation methods for SSEFs

The accuracy of convective scale precipitation forecasts depends not only on convective scale processes, but also on the synoptic and mesoscale environment and interactions across multiple scales (e.g., Lorenz 1969; Perkey and Maddox 1985; Zhang et al. 2007; Rotunno and Snyder 2008). A unique challenge for storm-scale DA is to properly estimate the atmospheric state on such a broad range of spatial scales. Observations resolving synoptic-, meso- and convective-scale features are therefore needed. The forecast model must also permit convective scale motions over a large enough domain to also resolve synoptic scale features.

Early studies assimilating convective scale radar data used a homogeneous ambient environment derived from a representative atmospheric sounding (e.g., Snyder and Zhang 2003; Dowell et al. 2004, 2011; Caya et al. 2005; Aksoy et al. 2009,2010). Some studies used coarser resolution model analyses and forecasts as background fields for storm scale DA, often adopting different independent DA methods for each (Xiao and Sun 2007; Stephan and Schraff 2008; Zhao et al. 2008; Dixon et al. 2009; Snook et al. 2011; Caron 2013; Brousseau et al. 2014; Chang et al. 2014; Simonin et al. 2014; Wheatley et al. 2014). A few studies have applied three-dimensional variational (3DVar) techniques iteratively with successively smaller length scales of background error covariance to generate multi-scale analyses (e.g., Crook and Sun 2002; Dixon et al. 2009; Xie et al. 2010, Schenkman et al. 2011). More recently, some studies have used ensemble based DA methods such as the Ensemble Kalman Filter (EnKF; Evensen 2003; Hamill 2006) to provide multi-scale analyses by assimilating observations resolving synoptic to convective scale features (Zhang et al. 2009; Yussouf et al. 2013;

Thompson 2014, Sobash and Wicker 2014). In such studies, the same ensemble based DA system is used to analyze both the storm scale features and the meso- to synoptic scale environment.

One advantage of ensemble based DA over 3DVar is the potential for more accurate spatial and cross-variable correlations between model state and observed variables. This is made possible by the flow dependent ensemble based background error covariance. There has not yet been a systematic comparison of the variational and ensemble based methods in the context of multi-scale DA where scales ranging from several km to 1000s of km are resolved by both the model and the observations. Past studies comparing 3DVar with ensemble based DA focus either on large scale (i.e., relatively coarse resolution requiring cumulus parameterization) or convective scale (e.g., radar observation) DA alone. For example, mesoscale to global scale (grid spacing 10s to 1000s of km) studies have shown EnKF and other ensemble based DA techniques to produce both more accurate analyses and forecasts than 3DVar, especially in data sparse regions (e.g., Meng and Zhang 2008, 2011; Whitaker et al. 2008, 2009; Wang et al. 2008ab, 2011, 2013; Yang et al. 2009; Buehner et al. 2010; Zhang, M. et al. 2011; Wang, H. et al. 2013; Schwartz and Liu 2014). Convective scale (grid spacing of several km) studies comparing 3DVar and ensemble based DA have been more limited in terms of the amount of studies, the types of radar observations assimilated and the diversity of cases studied. For example, Potvin et al. (2013) compared 3DVar and EnKF supercell analyses from radial wind observations. Li et al. (2012) compared tropical cyclone forecasts initialized from 3DVar and hybrid ensemble-3DVar analyses, also using radial wind observations. Caya et al. (2005) compared 4DVar and EnKF for

convective scale radar DA in a perfect model OSSE framework for an isolated supercell case. Carley (2012) showed that forecasts initialized from hybrid ensemble-3DVar analyses outperformed those from 3DVar for convective-scale radar DA with a single case that featured upscale growth of supercells into a multicellular mode. Carley (2012) used the Nonhydrostatic Multiscale Model on the B grid (NMMB) and the Ensemble Transform technique, rather than EnKF, to obtain the ensemble part of the hybrid background error covariance. Different from these early studies, the present study compares 3DVar and EnKF in the multi-scale scenario with radar radial velocity and reflectivity observations using ten diverse mid-latitude convection cases.

1.1b Initial condition perturbation methods for SSEFs

Limited predictability of warm season precipitation forecasts has been demonstrated by low deterministic forecast skill (Fritsch and Carbone 2004), theoretical arguments (Thompson 1957; Lorenz 1963), sensitivity to small perturbations (e.g., Hohenegger et al. 2006, Hohenegger and Schär 2007a, 2007b, Zhang et al. 2003, 2006), and sensitivity to model and physics differences (e.g., Zhang and Fritsch 1988; Zhang et al. 2006; Johnson et al. 2011a,b; Johnson and Wang 2012, 2013). The ability to resolve small scale features associated with rapid non-linear error growth limits the predictability of convection-permitting forecasts even more than that of coarser resolution forecasts (Elmore et al. 2002; Walser et al. 2004; Hohenegger et al. 2006; Hohenegger and Schär 2007a, 2007b; Zhang et al. 2003, 2006). Predictability studies at convection-permitting resolution have been limited to a small number of forecasts, rather than systematic evaluation over a period of many forecasts. Understanding

perturbation growth is important for ensemble design because ensemble perturbations are intended to approximate the growth of errors and hence the forecast uncertainty (Leith 1974; Toth and Kalnay 1997).

Hohenegger and Schär (2007a) found similar convective precipitation forecast sensitivity to different perturbation methods after about 11 hours for a case study. However, it is not known if these results are characteristic of other cases with different background flow and/or a different role of topography. Other studies have demonstrated large differences in predictability for different events. For example, Zhang et al. (2006) showed reduced sensitivity to small scale IC perturbations for a warm season heavy precipitation event compared to a cold season large scale cyclone event. Walser et al. (2004) and Hohenegger et al. (2006) further found that some warm season cases in the Alpine region characterized by stratiform precipitation exhibited greater predictability than some cases characterized by deep moist convection. However, it was also found that deep convective cases can exhibit higher predictability, depending on other factors such as the presence of topography and the residence time of the perturbations in convectively unstable regions. Done et al. (2012) have also related different aspects of predictability on two case studies to whether convection is in statistical equilibrium with large scale forcings. The characteristics of the evolution of different types of perturbations have yet to be systematically studied over an extended period of many convection-permitting forecasts.

While IC errors on multiple scales contribute to the forecast error, it is not clear how to optimally design corresponding multi-scale IC perturbations for SSEFs. Similar to regional mesoscale ensembles (e.g., Wang et al. 2014), IC perturbations for SSEFs

are typically generated by either downscaling perturbations from a coarser resolution ensemble without small scale IC perturbations (e.g., Hohenegger et al. 2008; Xue et al. 2010a,b; Zhang et al. 2010 ; Peralta et al. 2012 ; Schwartz et al. 2014; Kühnlein et al. 2014), or generating multi-scale IC perturbations directly on the forecast grid using cycled ensemble-based DA (e.g., Vie et al. 2011 ; Snook et al. 2011; Harnisch et al. 2014). Multi-scale IC perturbations have been shown to be more effective than downscaled global ensemble perturbations for regional mesoscale ensembles (e.g., Wang et al. 2014). However, evaluation of the relative advantages of these IC perturbation methods for SSEFs has been very limited (e.g., Harnisch et al. 2014; Kuhnlein et al. 2014). While it may be computationally expensive to generate an ensemble of IC perturbations at convection-permitting resolution, IC perturbation methods appropriate for coarser resolution ensembles may be particularly ill-suited for SSEFs (Hohenegger and Schar 2007a). It remains unclear what the advantages of generating the IC perturbations at the full model resolution are for SSEFs and how to optimally do so in the context of convective precipitation forecasting.

Small scale IC perturbations can have significant impacts on convection-permitting ensemble spread as a result of rapid propagation and upscale growth (Hohenegger et al. 2006; Hohenegger and Schär 2007a,b; Zhang et al. 2003, 2006; Leoncini et al. 2010 ; Chapter 3). However, initial studies have suggested that the added benefit of small scale IC perturbations may be very limited when larger scale perturbations are already present (e.g., Kong et al. 2007). Durran and Gingrich (2014) have even suggested that explicitly added small scale IC perturbations have no practical importance because of rapid downscale propagation of larger scale perturbation energy.

While the predictability study in Chapter 4 and Durran and Gingrich (2014) use random homogenous small scale IC perturbations, the forecast sensitivity can depend on the spatial structure of the small scale IC perturbations (Hohenegger and Schär 2007a; Chapter 3). Therefore, the impact of more realistic flow-dependent small scale IC perturbations that sample the fastest growing errors remains an open question in the context of multi-scale IC perturbations for SSEFs.

While high resolution, multi-scale IC perturbations may have forecast advantages, some studies have shown a negative impact of the inconsistency with the coarser resolution perturbations at the lateral boundaries (Caron 2013; Wang et al. 2014). Such inconsistencies can be alleviated by blending the multi-scale IC perturbations with the largest scales of the outer domain IC/LBC perturbations (e.g., Caron 2013). Therefore the impact of IC/LBC perturbation inconsistency is also evaluated in the context of the ensemble forecast system used in this study.

1.2 Dissertation overview

The overall goal of this research is to better understand the optimal design of DA and IC perturbation methods for SSEFs, in the context of convective scale probabilistic forecasts. A multi-scale DA system is demonstrated and used to systematically compare 3DVar and EnKF techniques in the multi-scale DA context. The impacts of different aspects of IC perturbation methods for SSEFs are then evaluated in a series of experiments with increasing complexity.

A GSI-based 3DVar and EnKF DA system is extended to the multi-scale assimilation of observations ranging from synoptic scale rawinsonde to convective scale

radar observations in Chapter 2. The EnKF and 3DVar techniques are systematically compared to each other to better understand the impacts of differences between the DA techniques on the analyses at multiple scales and the subsequent convection forecasts.

The sensitivity of hourly accumulated precipitation forecasts to large scale IC and physics perturbations is systematically compared to the sensitivity to different types of simple random homogeneous small scale IC perturbations in Chapter 3.

Ensemble analyses provided by the multi-scale GSI-based EnKF system are used to generate more complex mesoscale and multi-scale flow-dependent IC perturbations in a perfect model Observation System Simulation Experiment (OSSE) case study in Chapter 4. Different aspects of the IC perturbation methods are evaluated to qualitatively understand the impacts on SSEF skill for different convective forecasting applications.

The OSSE study is extended to 11 diverse convectively active cases to obtain a more robust objective understanding of the systematic impacts of the multi-scale IC perturbation method in Chapter 5.

The impacts of inconsistencies between the IC and LBC perturbations are investigated in Chapter 6.

Finally, in Chapter 7 the IC perturbation experiments from the OSSE study are repeated using the real data analyses and forecasts from Chapter 2. The real-data experiments show the impacts of model error, which is absent in the OSSEs, on the results.

A summary and discussion of conclusions is presented in Chapter 8.

Chapter 2: Multi-scale GSI-based data assimilation: Comparison of EnKF and 3DVar.

2.1 Introduction

In this chapter, a multi-scale Gridpoint Statistical Interpolation (GSI)-based data assimilation (DA) system is used to better understand the differences between EnKF and 3DVar analyses in the context of multi-scale DA, and the impact of such differences on subsequent convective scale precipitation forecasts. Here, multi-scale DA refers to the assimilation of observations from networks that have been designed to sample different scales of motion ranging from synoptic-scale rawinsonde observations to convective scale radar observations. The cases selected for systematic evaluation represent convective organization on scales ranging from discrete cellular convection to supercells to organized Mesoscale Convective Systems (MCS). The cases also include forcing mechanisms on a range of scales such as synoptic scale waves and fronts, mesoscale features such as drylines and storm scale features such as cold pools. A case study of an upscale growing MCS is also evaluated in greater detail to further understand the systematic differences.

The GSI-based DA system development and configuration are described in section 2.2 while the forecast events and observation data are presented in section 2.3. Systematic results over 10 cases are presented in section 2.4 and the results for a case study are presented in section 2.5. Section 2.6 contains a summary and discussion of conclusions.

2.2 System development and configuration

2.2a Forecast model configuration

The Weather Research and Forecast (WRF) Advanced Research WRF (ARW) model (Skamarock et al. 2005) is a highly scalable numerical prediction system that is widely used in both research and operational forecasting applications. The WRF-ARW version 3.2 (Skamarock and Klemp 2007) is used for the experiments in the present chapter and chapters 4-7. An outer domain is configured with 12 km grid spacing and 50 vertical terrain-following levels over a 326x259 grid point domain (Fig. 2.1). The physics configuration includes the Mellor-Yamada-Janic boundary layer scheme with Eta surface exchange parameterization (Janjic´ 1994; 2001), Noah land surface model (Ek et al. 2003), Rapid Radiative Transfer Model for General circulation models longwave (Iacono et al. 2008) and Goddard shortwave radiation (Tao et al. 2003), WRF Single Moment 6-class (WSM6) microphysics (Hong and Lim 2006), and Grell-3 cumulus parameterization (Grell and Dévényi 2002). An inner nested domain is configured similarly, except with 4 km grid spacing over a 346x277 grid point domain (Fig. 2.1) and no cumulus parameterization.

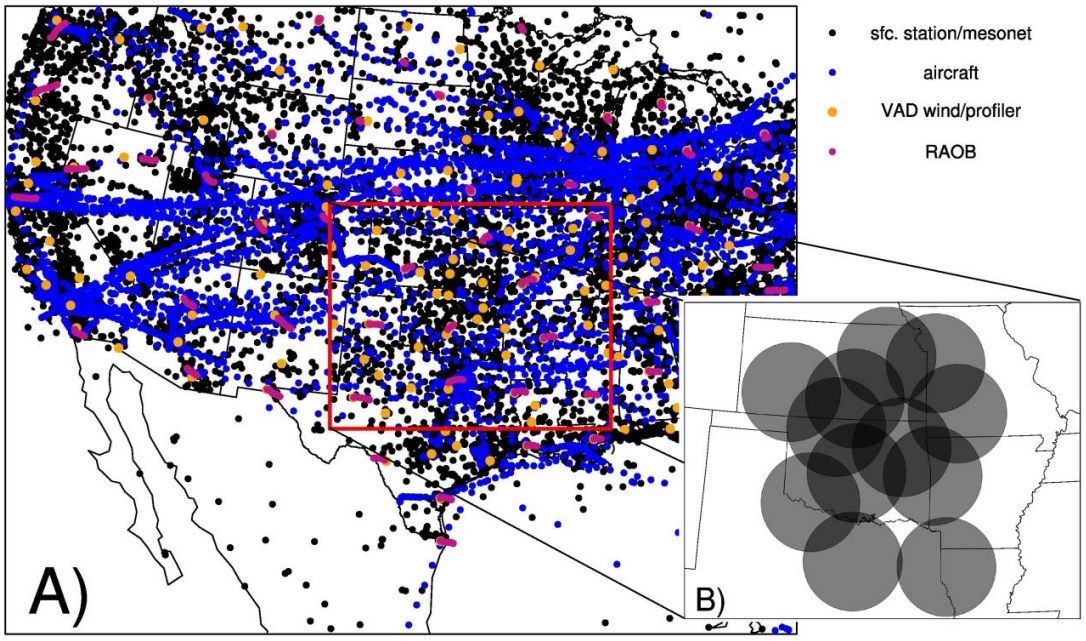

Figure 2.1: Location of (a) outer and (b) inner domains for the 20 May 2010 case study. A representative distribution of assimilated observation types is also shown.

2.2b Extension of GSI-based 3DVar and EnKF for direct radar DA

A GSI-based hybrid EnKF-3DVar system has been implemented operationally as part of the Global Forecast System (GFS) at the National Center for Environmental Prediction (NCEP). The newly implemented hybrid system improved both global forecast and hurricane forecast applications (Hamill et al. 2011; Wang et al. 2013; Wang and Lei 2014). The GSI-based hybrid system has also been integrated with other regional modeling systems such as the North American Mesoscale (NAM) model using the NMMB, the Rapid Refresh (RAP) using the WRF-ARW, and the Hurricane WRF (HWRF). These deterministic regional NCEP operational models currently use the ensemble covariance from the GFS ensemble in their hybrid DA systems. The GSI-based 3DVar and EnKF systems are here extended by further developing the convective scale radar DA capability for the WRF-ARW model to enable multi-scale DA. Since

the processing of observations, quality control and observation forward operators are all unified under the GSI framework, the GSI-based system allows a clean and direct comparison for understanding the differences between the multi-scale 3DVar and EnKF techniques. The system also provides an opportunity to study the impacts of different methods of generating multi-scale SSEF IC perturbations from the EnKF analysis ensemble in Chapters 4-7. In this subsection, the extension for GSI-based 3DVar is first described followed by the extension for GSI-based EnKF.

GSI-based 3DVar combines the first guess background forecast and assimilated observations by variational minimization of a cost function (e.g., Wu et al. 2002). The cost function includes penalty terms for the difference between the analysis and the observations, relative to the observation error covariance, and for the difference between the analysis and the background forecast, relative to the background error covariance. In 3DVar, the background error covariance is predefined and quasi-static. 3DVar therefore requires specification of a static background error covariance which affects how the observation information is spread out into the analysis. Only the radar radial wind can be assimilated during variational minimization in the operational GSI-based 3DVar. The GSI-based 3DVar is here extended to directly assimilate radar reflectivity observations by implementing additional control variables, forward observation operator, and background error statistics. Consistent with the WSM6 microphysics scheme, the new control variables added to the GSI variational minimization are the mixing ratios of rain, snow and graupel hydrometeors. The logarithm is first applied to these hydrometeor mixing ratio control variables to reduce non-Gaussianity of the error statistics and to minimize the errors associated with the

linearization and adjoint of the reflectivity observation operator. Carley (2012) also extended the GSI variational minimization with hydrometeor control variables for the Ferrier microphysics scheme (Ferrier et al. 2002, 2011) and conducted experiments with the NMMB model.

An observation forward operator consistent with WSM6 is also introduced into the GSI-based 3DVar system for reflectivity DA. The observation operator defines the relationship between model control variables (hydrometeors) and the observed quantity (reflectivity). The reflectivity is a function of the rain, snow and graupel hydrometeor mixing ratios that also depends on the background temperature. The observation operator follows Dowell et al. (2011) except that, in addition to snow, hail is also classified as either wet or dry based on the background temperature as in Tong and Xue (2005). The contribution to reflectivity from rain and snow/graupel is also set to zero if the background temperature is less than -5 C and greater than 5C, respectively. This step avoids unrealistic increments that add snow/graupel at very warm levels or rain at very cold levels (Gao and Stensrud 2012).

One of the challenges of 3DVar for radar reflectivity assimilation is the specification of an appropriate static background error covariance model. The background errors of hydrometeor variables can be highly correlated with errors in other model variables (Michel et al. 2011). However, including such cross-variable correlations in the static background error covariance model in a computationally efficient manner during variational minimization remains a challenge. Therefore, cross-variable correlations between hydrometeor and other control variable errors have typically been neglected in the static covariance in previous variational assimilation of

reflectivity observations (e.g., Caya et al. 2005; Carley 2012; Wang, H. et al. 2013). This approach is also used in the present study. Future work will explore the inclusion of such cross variable correlations associated with reflectivity assimilation through the use of ensemble covariance in the variational framework, following a similar approach as the hybrid ensemble-variational DA method (e.g. Wang et al. 2008a).

Here, the static 3DVar background error covariance for reflectivity observation assimilation is defined as follows. The amplitudes of the background error variances for the hydrometeor mixing ratios are defined as a function of height and hydrometeor type. The height dependence of the background error variance for each hydrometeor type is the variance of 5 minute ensemble forecasts initialized at 0000 UTC 20 May 2010, averaged over the convectively active region. The amplitude is further tuned to obtain subjectively reasonable looking increments and analyses when used to assimilate reflectivity observations in test cases. The spatial correlation of the hydrometeor background errors is determined by tuning the spatial correlation of the static background error covariance that is used by the mesoscale NCEP NAM model to assimilate conventional water vapor observations. The horizontal correlation length scale is reduced by a factor of 20 and the vertical length scale is reduced by a factor of 4. These values are also chosen to minimize objective and subjective errors during test cases. The static background error covariance for radar radial wind assimilation is also tuned so that the spatial scale is reduced using the same factors as for reflectivity.

The GSI-based EnKF is based on the Ensemble Square Root Filter (EnSRF) of Whitaker and Hamill (2002). Like the GSI-based 3DVar, the GSI software performs the observation quality control and applies the observation operators to the model first-

guess fields. The GSI-based EnKF has the option to take into account the four-dimensional ensemble covariance within the assimilation window to assimilate asynchronous observations. This EnKF code has been efficiently parallelized following Anderson and Collins (2007) and directly interfaced with the GSI by using the GSI's observation operators, pre-processing and quality control for operationally assimilated data. In the GSI-based EnKF currently implemented operationally for NCEP GFS, it does not contain radar data assimilation.

In this study, the GSI-based EnKF is further extended to include the assimilation of radar data through three new developments. First, the ensemble observation priors are extended to include both radar radial wind and reflectivity. This is accomplished by applying the GSI observation operators developed for GSI 3DVar on the first guess ensembles and ingesting the resulting observation priors into EnKF. Second, the option to include rain, snow and graupel hydrometeor mixing ratios as state variables is added to the EnKF code. Third, the EnKF WRF interface is extended to read the ensemble first guess and update the ensemble analysis of these new state variables.

2.2c 3DVar configuration

For the 3DVar experiments, specification of a static background error covariance is needed. In this study, the outer domain has comparable resolution to the operational NAM. The background error covariance from the regional NCEP NAM (NCAR 2011) is therefore adopted for non-radar observations assimilated on the outer domain. The background error covariance for assimilation of radar observations is constructed and tuned as described above in section 2.2b.

2.2d EnKF configuration

For the EnKF experiments, each forecast step contains a 40-member ensemble forecast and each analysis step provides a 40-member analysis ensemble. The 12-km grid ensemble at the very beginning of the DA cycles is created by adding random perturbations to the operational NCEP GFS analysis. These perturbations are drawn so that their covariance is equal to the static covariance in the WRF 3DVar (Wang et al. 2008a). The same method is used to perturb the 12 km outer domain LBCs. The outer domain ensemble provides the initial and lateral boundary conditions for the inner domain ensemble.

The general EnKF/EnSRF theory and equations have been described in many papers (e.g., Whitaker and Hamill 2002; Whitaker et al. 2008). One challenge in applying EnKF is the treatment of system errors associated with the sampling errors and misrepresentation of model errors. Covariance localization and inflation are commonly used to treat such deficiencies. Optimal methods and parameters for covariance localization and inflation are application dependent. The details of the methods used for the present study are outlined below.

Two methods of posterior covariance inflation are used starting from the values found for Whitaker and Hamill (2012). The parameters are tuned to minimize the first guess errors during the 20 May case study. The first guess errors for the other cases are similar in magnitude to the 20 May case that is used for tuning, suggesting that the parameters are also generally appropriate for other cases. The first inflation method is a height-dependent multiplicative inflation that is applied uniformly across the domain to

all ensemble perturbations. This multiplicative inflation is intended to account for model errors that are not represented in the ensemble (Whitaker and Hamill 2012), such as the errors associated with physical parameterizations. The amount of multiplicative inflation at the surface is 15% every 3 hours on the outer domain or, equivalently, ~0.4% every 5 minutes on the inner domain. The multiplicative inflation smoothly tapers to ~9% at 200mb and ~3% at the 50mb model top to avoid excessive spread near the model top (Fig. 2.2), similar to Zhu et al. (2013). The second inflation method is Relaxation To Prior Spread (RTPS; Whitaker and Hamill 2012) which inflates the posterior ensemble spread to a fraction, α, of the prior ensemble spread. The RTPS accounts for excessive spread reduction during the assimilation of observations resulting from sampling errors in the ensemble approximation of the Kalman gain. Thus the RTPS inflation is greatest where there are many observations and is absent where there are no observations. For our experiments, a value for α of 0.95 is used to inflate the posterior ensemble spread to 95% of the prior ensemble spread for both the inner domain and outer domain. On the outer domain, the average consistency ratio of observation first-guesses (i.e., ensemble spread divided by ensemble mean error) for wind, temperature and water vapor is 0.87 at the end of DA, indicating a reasonably well-tuned system. On the inner domain the consistency ratios are stable and, similar to other studies (e.g., Aksoy et al. 2009; Dowell et al. 2011; Sobash and Stensrud 2013), indicate some ensemble under-dispersion with values ~0.75 for both radial velocity and reflectivity. However, tuning tests show degradation of accuracy when the inflation is further increased (not shown).

Covariance localization is used to minimize the impact of sampling errors in the ensemble covariance which are greatest where the actual correlation is small, such as at large distances (Sobash and Stensrud 2013). The covariance localization is applied using the Gaspari and Cohn (1999) function with a cutoff radius also tuned to minimize the first guess errors during the 20 May case study. For the assimilation of synoptic/mesoscale observations on the outer domain, the localization is also height and variable dependent (Zhu et al. 2013). For the synoptic/mesoscale observation assimilation, the horizontal localization is set to 700 km at the surface and increases by a factor of 1.5 at the model top (Fig. 2.2). Vertical localization increases from 0.275 to 0.55 scale height (natural log of pressure) for temperature and moisture and increases from 0.55 to 1.1 scale height for wind (Fig. 2.2). For the inner domain storm-scale DA, constant covariance localization length scales are used. For the inner domain DA, tests of horizontal covariance localization showed very little sensitivity of the forecast to the cutoff radius. Among the range of radii tested, a value of 20 km (i.e., 5 grid points) showed a slight improvement over other values such as 16 and 12 km. The chosen value of 20 km is comparable to the 18 km radius that was found to work well in Sobash and Stensrud (2013) for a similar 50-member ensemble. However, this value is somewhat larger than that used in many early studies of EnKF with radar data. Such early studies focused primarily on isolated supercells whereas Sobash and Stensrud (2013) suggest that the larger localization radius is beneficial for other convective modes, such as cell mergers and MCS cases. The grid spacing of 4 km also necessitates a larger radius than used in past studies with 1~3km grid spacing. The vertical localization for radar reflectivity and velocity assimilation is 1.1 in scale height units.

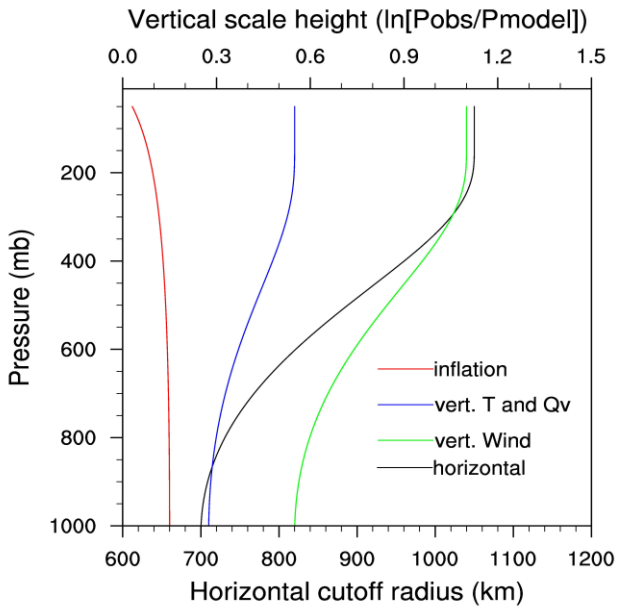

Figure 2.2: Covariance localization length scales used for EnKF for assimilation of synoptic/mesoscale observations on the outer domain in the horizontal (bottom axis, black line) and vertical (top axis, blue and green lines for T/Qv and wind respectively) directions. Also shown is the vertical profile of the constant multiplicative inflation (red line), using the top horizontal axis..

2.2e DA cycling configuration

An important consideration for cycled multi-scale DA with observations and model that resolve many different scales of motion is the choice of cycling frequency. Peña and Kalnay (2004) suggest that the cycling frequency can be consistent with either the larger scale baroclinically driven or smaller scale convective instability driven modes of error growth, but not both. This is because the atmosphere contains fast-growing (i.e., moist convective instability) errors that saturate at a lower amplitude than slower-growing (i.e., synoptic scale baroclinic instability) errors. Here, the storm scale and synoptic/mesoscale observations are assimilated with different cycling frequencies, chosen to correspond to the approximate error growth rates of features observed by each. A 3-hour cycling interval is used to assimilate the synoptic/mesoscale

observations on the outer domain and a 5 minute cycling interval is used to assimilate the storm scale radar observations on the inner domain (Fig. 2.3).

Synoptic/mesoscale observations within the three hour window centered on the analysis time are assimilated. For both 3DVar and EnKF, FGAT (First Guess at Appropriate Time) is used by outputting first-guess fields at 30-min increments until 1.5 hours after the analysis time. The first guess fields are then compared to the observations at the observation time using linear interpolation in time. In addition, for EnKF asynchronous assimilation is adopted, where the ensemble covariance at the observation time determines the increment at the analysis time.

The second to last outer domain analysis at 2100 UTC is used to initialize the first inner domain 5 minute forecast for the inner domain DA cycles (Fig. 2.3). For each member, the forecast from the 2100 UTC outer domain analysis also provides the inner domain lateral boundary condition. Radar observations within 5- minute windows are assimilated every five minutes (synchronously) until the final analysis time at 0000 UTC. Before the final analysis of radar observations at 0000 UTC, the synoptic/mesoscale observations are also assimilated on the inner domain. This final step reduces the inconsistency between the outer and inner domain analyses of the synoptic/mesoscale environment. The different cycling intervals for the mesoscale and storm-scale DA are chosen based on the different approximate error growth rates on the different spatial scales. This configuration is therefore expected to result in IC perturbations that are flow-dependent and fast growing on multiple scales (Peña and Kalnay 2004).

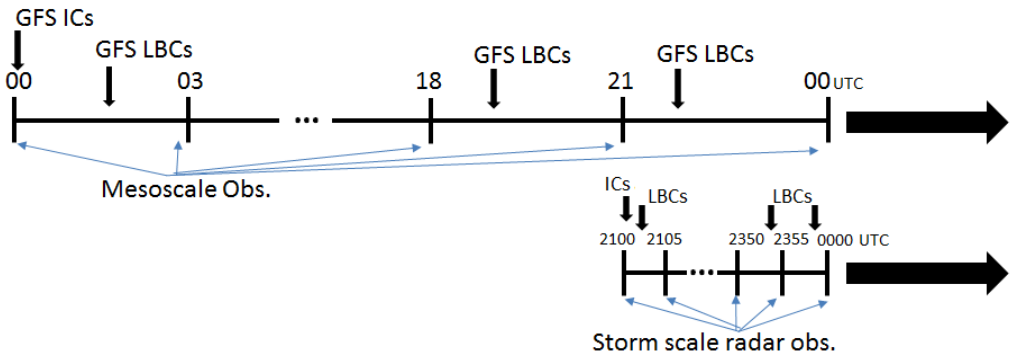

Figure 2.3: Schematic of data assimilation configuration showing the 3 hour cycling of mesoscale outer domain DA, driven by NCEP GFS IC/LBCs, and 5 minute cycling of storm scale inner domain radar DA, driven by the mesoscale analysis cycles. A forecast is then initialized from the 0000 UTC analyses.

2.3 Forecast events and observation data

2.3a Forecast events

Ten convectively active events from May 2010 are selected for the evaluation of the GSI-based multi-scale DA system (Table 2.1). Given the computationally intensive nature of the experiments, the number of cases (10) is chosen to provide a minimally adequate sample size for establishing statistically significant results while keeping the computational demands of the project manageable. In order to ensure robust results, diverse cases including examples of both discrete cellular convection and organized MCSs are used. The cases also include diverse forcing and organizing mechanisms such as strong upper level shortwaves and surface cold fronts, slow moving or stationary frontal zones with relatively weak large-scale ascent, and convective storm outflows. Since the focus of each convective episode was in a slightly different geographic location for each case, the center of the model domain is relocated for each case (Table 2.1).

Table 2.1: Convectively active days from May 2010 that are selected for use in Chapters 2 and 4-7. The columns are, from left to right, analysis time, domain center latitude (degrees N), and domain center longitude (degrees E).

Analysis Time	Center latitude	Center longitude
00 UTC 11 May	37.5	-98.0
00 UTC 13 May	35.5	-100.0
00 UTC 15 May	32.0	-101.5
00 UTC 17 May	34.0	-95.0
00 UTC 18 May	29.0	-101.0
00 UTC 19 May	36.0	-102.2
00 UTC 20 May	35.5	-100.0
00 UTC 21 May	34.0	-94.5
00 UTC 25 May	37.5	-102.5
00 UTC 26 May	37.5	-102.5

In addition to the systematic results, a case study of the forecasts initialized at 0000 UTC 20 May 2010 is selected for more in depth analysis. The 20 May case is selected because the convective cells present at 0000 UTC grew upscale into an MCS during the forecast period. This makes the forecast particularly sensitive to both the storm scale analysis and larger scale environment because of the multi-scale nature of such convective systems (e.g., Perkey and Maddox 1985) and upscale growth of small errors on this case (Chapter 3). At 0000 UTC 20 May there was a broad and slow moving trough aloft with an embedded shortwave rounding its base (Fig. 2.4a). At the surface, a weak surface low propagated from central Oklahoma into western Missouri between 0000 UTC 20 May and 0000 UTC 21 May without substantial intensification (Fig. 2.4b,d,f). By 0600 UTC 20 May, cellular convection from the previous evening (Fig. 2.5a,b) was organizing into an MCS in eastern Oklahoma, Arkansas and Missouri that dissipated by 1200 UTC (Fig. 2.5c). The remnant outflow boundary was the focus for additional convection that developed the following afternoon (Fig. 2.5d,e).

Stratiform precipitation also developed by 1200 UTC from southeastern Nebraska to southeastern Missouri (Fig. 2.5c) and weakened later in the day (Fig. 2.5d,e).

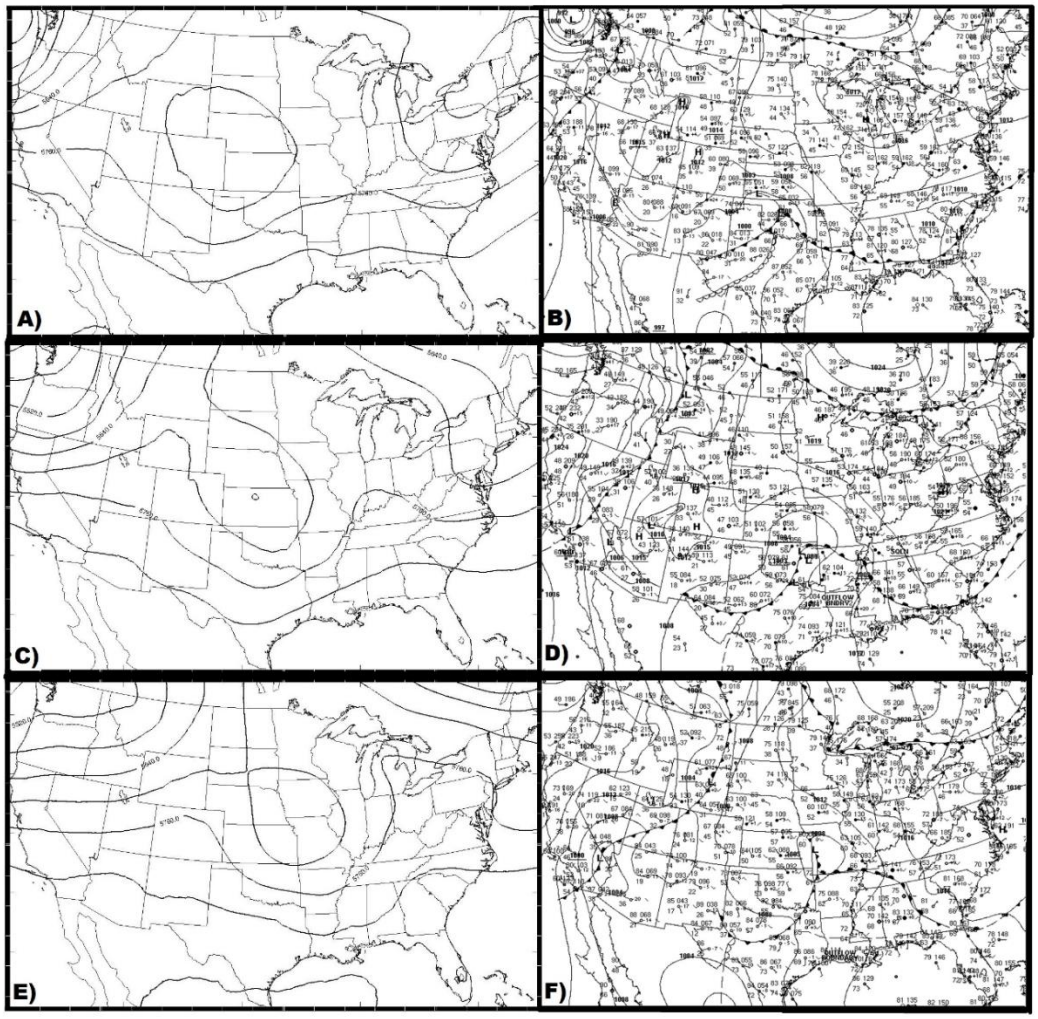

Figure 2.4: Synoptic scale conditions at (a), (b) 0000 UTC 20 May, (c), (d) 1200 UTC 20 May and (e), (f) 0000 UTC 21 May. In (a), (c) and (e), 500 hPa geopotential height of the control member forecast initialized at 0000 UTC 10 May is shown. In (b), (d) and (f) the mean sea level pressure, surface fronts and surface observations from the Hydrometeorological Prediction Center surface analysis archive are shown (http://www.hpc.ncep.noaa.gov/html/sfc_archive.shtml).

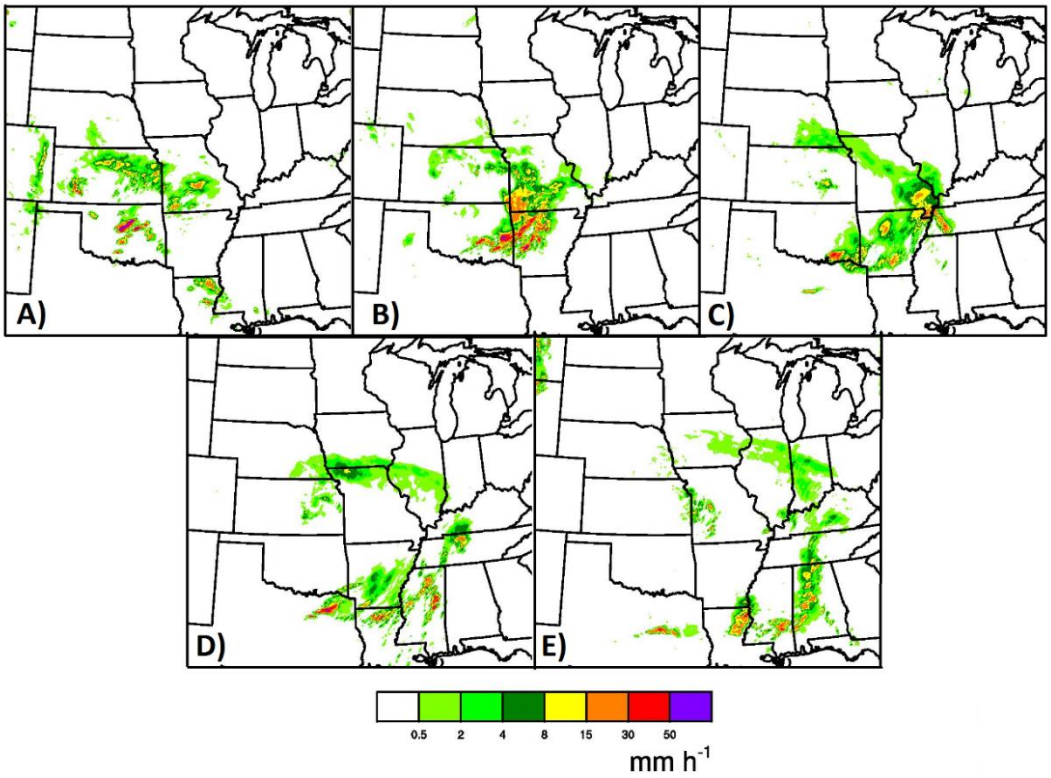

Figure 2.5: Observed hourly accumulated precipitation at (a) 0100 UTC 20 May 2010, (b) 0600 UTC, (c) 1200 UTC, (d) 1800 UTC and (e) 0000 UTC 21 May

2.3b Observation data

Since both the 3DVar and EnKF techniques are GSI-based, the same observations and quality control methods are used for both techniques. The synoptic/mesoscale observation data are obtained from the Climate Forecast System Reanalysis project at the NOAA Operational Model Archive and Distribution System (NOMADS). These observations include rawinsonde, surface station, surface mesonet, Aircraft Communications Addressing and Reporting System (ACARS), and NOAA wind profiler observations (Fig. 2.1).

Radar observations of reflectivity and radial velocity are obtained from the NEXRAD level 2 data archived at the National Climate Data Center (NCDC) and

quality controlled using the Warning Decision Support System – Integrated Information (WDSSII; Lakshmanan et al. 2007a) software (www.wdssii.org). For the reflectivity data, the neural network based w2qcnn utility within WDSSII (Lakshmanan et al. 2007b, 2010) is used to remove non-meteorological echoes. The reflectivity data are then thresholded at 5 dBZ such that all data less than or equal to 5 dBZ are considered "no precipitation" observations (Aksoy et al. 2009). Velocity data are then dealiased using a two-dimensional dealiasing algorithm (Jing and Wiener 1993). Velocity data are also thresholded based on the reflectivity data, such that the velocity data are omitted where the reflectivity is less than or equal to 5 dBZ (Aksoy et al. 2009). Additional gross error checks are also performed within GSI. Velocity observations are rejected if the difference from the background value is greater than 30 ms^{-1} which is only likely to occur in cases of extreme aliasing that was missed during pre-processing. Reflectivity observations are not rejected based on the observation-background difference because very large differences may not indicate bad observation data, especially during the early DA cycles. The observation error of radar velocity and reflectivity is assumed to be 2 m s^{-1} and 5 dBZ, respectively.

The precipitation forecasts are verified against radar-derived quantitative precipitation estimates from the National Severe Storms Laboratory Q2 product (Zhang, J. et al. 2011).

2.4 Results aggregated over 10 forecasts

Ten diverse cases are used to provide a robust evaluation of the impact of the differences between multi-scale GSI-based EnKF and 3DVar. The cases include a

variety of different forcing mechanisms and convective modes. In this study, "mesoscale analysis" refers to the analysis generated by the outer domain DA which assimilates observations that only resolve mesoscale and larger features using the 12-km grid. The "multi-scale" analysis refers to the result of the inner domain DA which includes further storm scale radar DA using the mesoscale analysis as a background field. In addition to providing a background environment consistent with the storm scale features to be assimilated, the mesoscale environment also interacts with the storm scale features during the DA and forecast periods to affect the overall forecast evolution (e.g., Perkey and Maddox 1985). Therefore, the mesoscale analyses are first evaluated to distinguish the impacts of the synoptic/mesoscale environment differences on the subsequent forecasts in section 2.4a, followed by evaluation of the multi-scale analyses in section 2.4b.

2.4a Mesoscale analysis evaluation

Since the true atmospheric state is unknown and approximated by the analyses, the quality of such analyses is evaluated based on the similarity of a subsequent forecast to independent observations. For example, the first guess errors of the short-term forecasts during the DA period are commonly used to evaluate DA systems. Here, the first guess errors are averaged only over the last 5 cycles to allow the EnKF to spin up reasonable estimates of the background error covariance. This approach also emphasizes the end of the DA period which is consistent with the focus on the final 0000 UTC analysis time. The first guess errors during the mesoscale DA on the 12-km grid are generally smaller for EnKF than for 3DVar (Fig. 2.6), with the exception of

temperature between 500 and 700 mb, and above 100 mb. The difference is fairly uniform with height for wind (Fig. 2.6b) and is most pronounced at low levels for temperature and moisture (Fig. 2.6a,c). The differences in first guess errors are statistically significant at most levels for wind and moisture (Fig. 2.6b,c). The limited advantage of EnKF over 3DVar for temperature at certain levels may be due to systematic model biases in temperature at those levels. A more pronounced warm bias for EnKF than 3DVar was noted both between 500 and 700 mb and above 100 mb (not shown). A WRF model temperature bias near the model top (i.e., above 100 mb) was also documented by Wee et al. (2012). EnKF may not correct these systematic biases as well as 3DVar because such biases would be common to all members and therefore lead to an under-estimate of the background error based on the ensemble variance. While further covariance inflation may improve EnKF performance at such levels, there is an overall negative effect of additional inflation when all levels and variables are considered. The smaller first-guess errors using GSI-based EnKF suggest that the synoptic/mesoscale environment is analyzed more accurately than using GSI-based 3DVar.

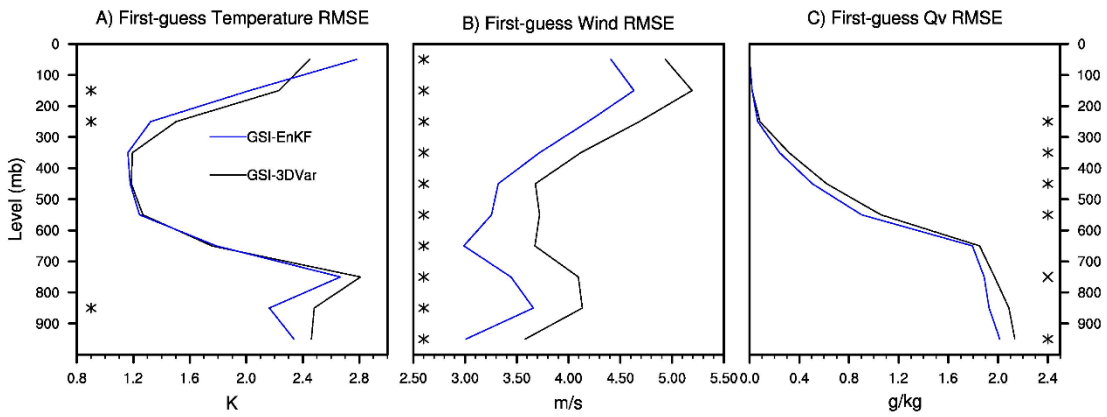

Figure 2.6: RMSE of outer domain first guess (i.e., 3 hr forecast) of (a) temperature, (b) wind, and (c) water vapor mixing ratio observations averaged over the inner domain region, excluding the first 3 DA cycles while the ensemble covariance spins up (i.e., last 5 cycles only) and averaged over all 10 cases. Markers indicate a significant difference between the two lines at the 90% level (crosses) or 95% level (asterisks). Statistical significance is determined using permutation resampling (Hamill 1999).

Free forecasts are also run out to 8 h lead time on the 4 km grid by interpolating the mesoscale analyses to the 4 km grid using the WRF ndown utility, without radar DA. The better mesoscale analyses for EnKF than 3DVar are also reflected in the Equitable Threat Score (ETS) of such precipitation forecasts (Fig. 2.7; dashed lines). The results discussed herein are similar using the Hiedke Skill Score and Neighborhood Probability Brier Skill Score (not shown). The precipitation forecasts initialized from the mesoscale analyses show the contribution of the synoptic/mesoscale environment analysis to the forecast skill differences. Although the 3DVar precipitation forecasts are more skillful than the EnKF precipitation forecasts for the first couple of hours without radar DA (i.e. mesoscale/synoptic DA only), the EnKF forecasts become more skillful starting at about forecast hour 3 (Fig. 2.7; dashed lines). The differences in skill are consistent with past studies showing 3DVar to fit to observations better than EnKF at

the analysis and very short forecast times but with faster error growth during the forecast for 3DVar (e.g., Wang et al. 2008b; Li et al. 2012). The initially closer fit to observations for 3DVar may be important for more quickly spinning up the initial storms. The difference in skill during the first two hours may also be related to a smoothing of features in the ensemble mean used for the EnKF background. A smoothing of features associated with focused convergence can slow the spin up of convection. The better performance of the EnKF forecasts at later lead times after the short spin-up period suggests that the larger scale environment is more supportive of the actual convective evolution in the EnKF analyses than the 3DVar analyses.

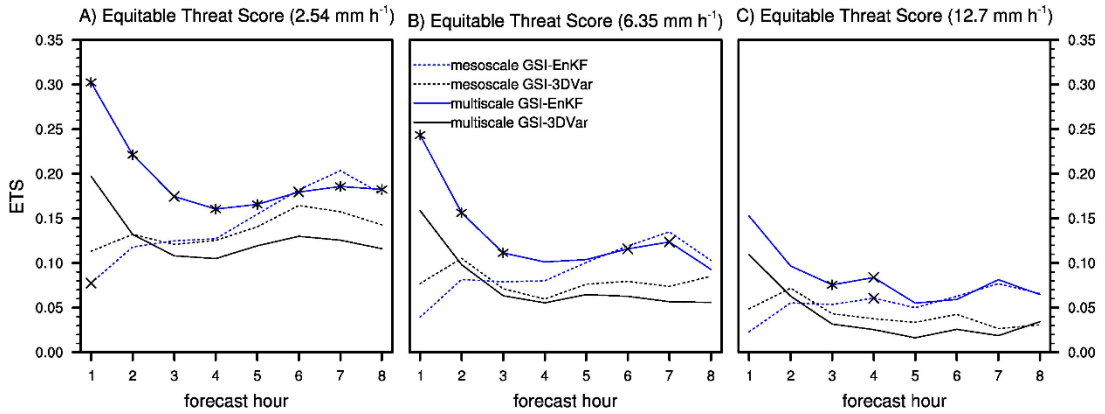

Figure 2.7: Equitable threat score of hourly accumulated precipitation forecasts initialized from the outer domain mesoscale EnKF analysis (blue dashed), inner domain multi-scale EnKF analysis (blue solid), outer domain mesoscale 3DVar analysis (black dashed) and inner domain multi-scale 3DVar analysis, using a threshold of (a) 2.54 mm h^{-1}, (b) 6.35 mm h^{-1} and (c) 12.7 mm h^{-1}. Markers along the solid (dashed) blue line indicate a significant difference from the solid (dashed) black line. Marker types are determined as in Fig. 2.6. Panel (d) shows the observation frequency for all three thresholds.

2.4b Multi-scale analysis evaluation

The mesoscale analyses drive the inner domain radar DA which adds storm-scale features to provide the final multi-scale analyses. The short lead time (5 min) first guess forecasts during radar DA are used to evaluate the analysis of storm scale features. First guess errors are again averaged over the last part of the DA period (here, 12 cycles which is 1 h of DA) to avoid the initial spin up period and focus on the final analysis time. The 4 km domain EnKF first guess errors are significantly smaller than the 3DVar first guess errors for radial velocity at all levels (Fig. 2.8a) and for reflectivity at many levels (Fig. 2.8b). The exceptions for reflectivity include the near surface level and the 500-600 mb level (Fig. 2.8b). Therefore, on average the storm scale details are also better analyzed with GSI-based EnKF than GSI-based 3DVar.

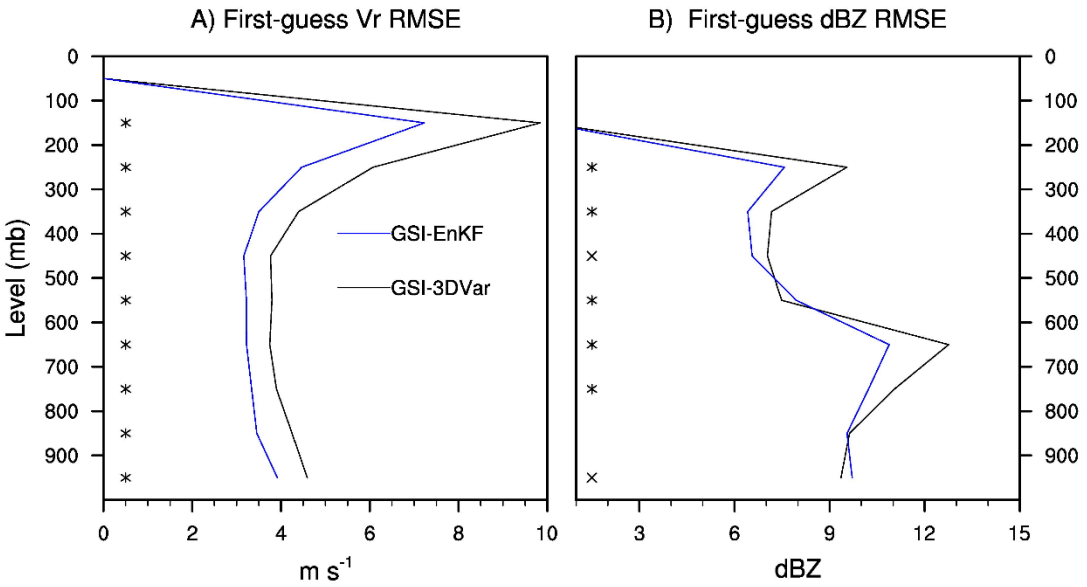

Figure 2.8: As in Fig. 2.6, except for (a) radial velocity and (b) reflectivity during inner domain radar DA, averaged over the 10 cases for the last 12 DA cycles to emphasize the time period close to the 0000 UTC analysis time.

Longer (8 h) forecasts are used to evaluate the combined influence of the multi-scale analysis of storm-scale and mesoscale features. The 8 h forecasts initialized from the multi-scale analyses are also more skillful using GSI-based EnKF than using GSI-based 3DVar. The difference is generally statistically significant except for the highest threshold at several times, likely due to the smaller sample size at the higher threshold which is a rarer event (Fig. 2.7; solid lines). The more pronounced difference in skill indicates that all scales of motion contributing to the convective evolution, not only the mesoscale environment, are better analyzed with GSI-based EnKF than GSI-based 3DVar. At the later lead times the skill of forecasts initialized from the multi-scale analyses (Fig. 2.7; solid lines) is generally similar to the skill of forecasts initialized from the mesoscale analyses (Fig. 2.7; dashed lines). This shows that at later lead times the differences in the mesoscale environment contribute increasingly to the difference in forecast skill. This contrasts with the dominant impact of the storm scale analysis at early lead times.

The impact of storm scale radar DA on the precipitation forecasts, compared to the mesoscale analyses without radar DA, is also notably different for EnKF and 3DVar. The impact of the better storm scale analyses for EnKF is to increase the skill of the precipitation forecast for ~4-5h (Fig. 2.7). However, for 3DVar the precipitation skill is only improved by radar DA during the first hour and is then degraded compared to the mesoscale analysis at later times (Fig. 2.7; black dashed vs black solid). This shows that the multi-scale EnKF analyses lead to forecasts that more realistically maintain the storm scale features and their interaction with the larger scale environment, compared to the multi-scale 3DVar analyses.

2.5 20 May 2010 case study

A case study is used to better understand the systematic differences between the GSI-based EnKF and 3DVar multi-scale analyses and subsequent forecasts. This case of an upscale growing MCS is selected for evaluation of multi-scale analyses because of strong sensitivity to analysis perturbations on all spatial scales (Chapter 3). The mesoscale analyses are again evaluated first to distinguish the impact of differences in the synoptic/mesoscale environment from the impact of the storm scale analyses.

2.5a Mesoscale analysis evaluation

The mesoscale analyses in this case lead to generally smaller first guess errors during the outer domain DA period for GSI-based EnKF than for GSI-based 3DVar (Fig. 2.9). The EnKF advantage is seen for all variables at most levels, with the exception of temperature near the model top (Fig. 2.9a). This exception is likely related to the WRF temperature bias, as mentioned in section 2.9a. This case is thus representative of the systematic results in that the GSI-based EnKF provides a better analysis of the synoptic/mesoscale environment than the GSI-based 3DVar.

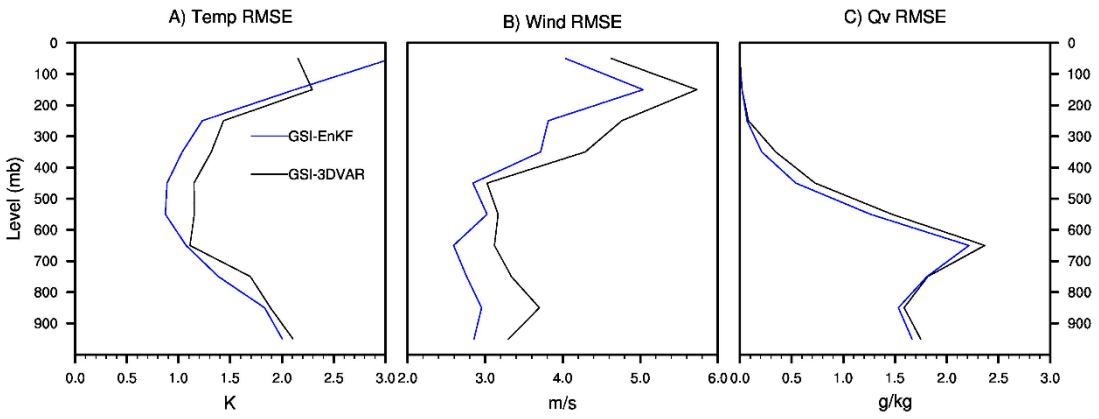

Figure 2.9: As in Fig. 2.6, except only for the 20 May 2010 case study.

The impact of the more accurately analyzed synoptic/mesoscale environment for EnKF on the precipitation forecasts is seen in the Neighborhood Probability (NP; Schwartz et al. 2010) forecasts initialized from the mesoscale analyses (Fig. 2.10). The 3DVar and EnKF forecasts both develop an MCS in about the same location as observed during the second forecast hour (Fig. 2.10b,j). However, for both 3DVAR (Fig. 2.10i) and EnKF (Fig. 2.10a) the forecast MCS takes more than an hour to "spin up", resulting in under-prediction of precipitation during the 1st hour and a slight westward displacement of the NP maximum relative to the observed MCS at later forecast times (e.g., Fig. 2.10c,k). Furthermore, the MCS in the 3DVar forecast starts to dissipate several hours too early (Fig. 2.10n,o,p). During this time spurious convection develops in southeastern Oklahoma and becomes dominant as it moves into southwestern Arkansas (Fig. 2.10j-p). The spurious precipitation develops along a northwest-to-southeast oriented warm front in southeastern Oklahoma. The development and dominance of this spurious precipitation in the 3DVar forecast is a result of the poorer synoptic/mesoscale environment analysis for 3DVar. In particular, the warm front is displaced to the southwest and shows a more pronounced wind shift and convergence in the 3DVAR analysis than the EnKF analysis. The warm front including location, wind shift and enhanced surface temperature gradient is much better analyzed by the EnKF analysis than the 3DVar analysis. The difference is still present in the final multi-scale analysis after radar DA (Fig. 2.11), showing the importance of the mesoscale environment analysis.

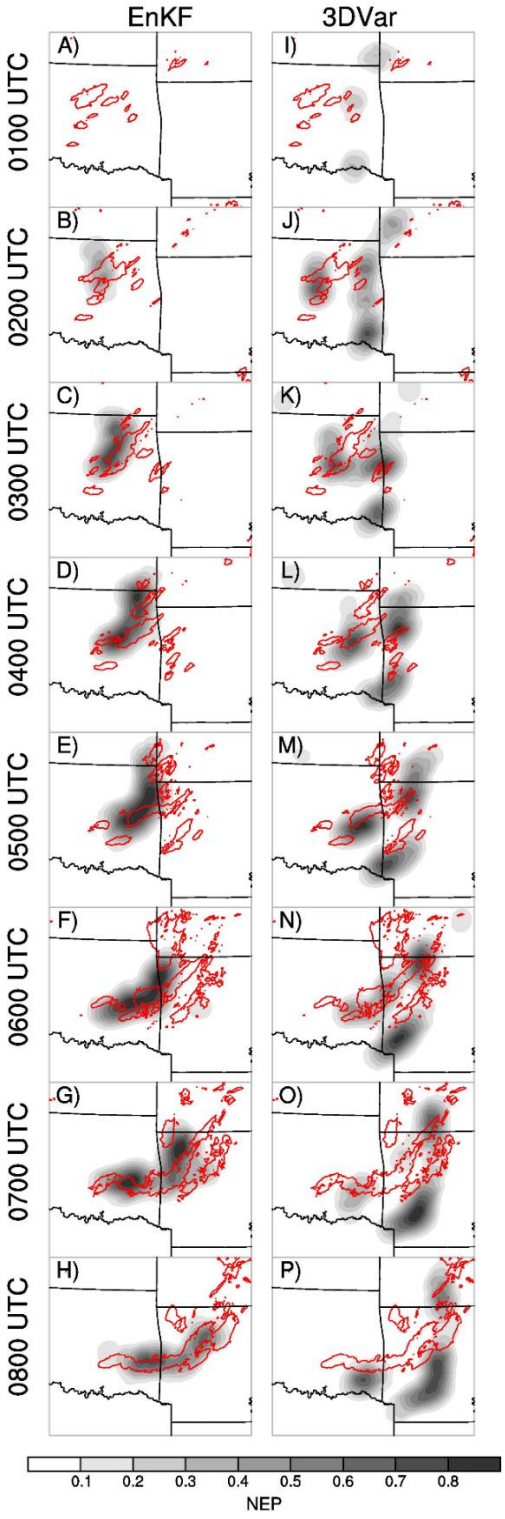

Figure 2.10: Neighborhood probability forecast of hourly accumulated precipitation exceeding 12.7 mm h^{-1} (shaded) and observation contour of 12.7 mm h^{-1} (red contour). Forecasts are initialized from downscaled outer domain (a)-(h) EnKF and (i)-(p) 3DVar analyses.

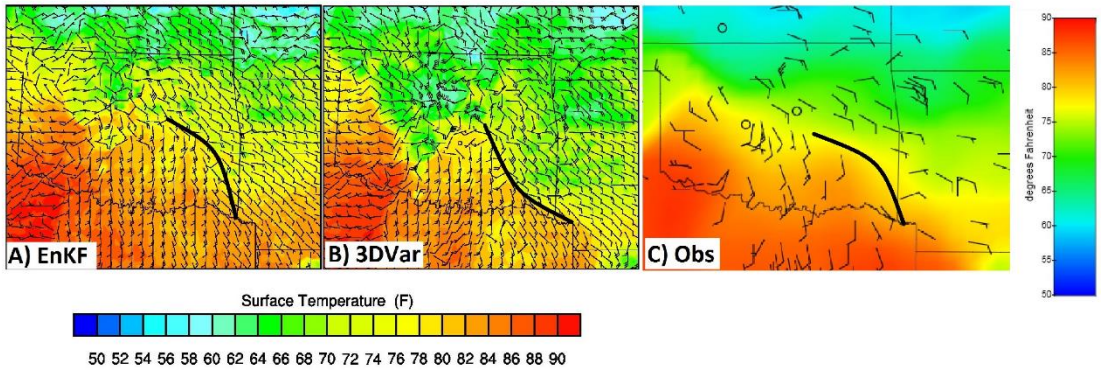

Figure 2.11: Surface temperature and wind at 0000 UTC 20 May 2010 from (a) EnKF inner domain analysis, (b) 3DVar inner domain analysis and (c) objective analysis of Oklahoma Mesonet observations using "WeatherScope" software downloaded from [http://www.mesonet.org/index.php/weather/weatherscope]. The color bar for panel (c) is at the far right.

Subjective evaluation of the mesoscale analyses throughout the DA period was conducted to better understand how the differences between the 3DVar and EnKF methods contributed to the above differences in forecasts initialized from the 0000 UTC mesoscale analyses. The southwestward displacement of the warm front analyzed by 3DVar, compared to the warm front analyzed by EnKF, first appears in the 1500 UTC background forecast (Fig. 2.12a,b; thick black line). Low level clouds developed north of the warm front between 1200 and 1500 UTC in the 3DVAR forecast only (not shown). The clouds prevented surface warming and impeded the northward advancement of the warm front during subsequent cycles. This is evident at 1500 UTC by the cooler temperatures north of the warm front wind shift and the more pronounced temperature gradient in the 3DVar background field (Fig. 2.12a), compared to the EnKF background field (Fig. 2.12b). The difference is particularly pronounced along the Arkansas-Oklahoma border (Fig. 2.12e,f; area in red circle). The observation innovations (observation minus first guess) in this area are generally positive by several

degrees K for 3DVar (Fig. 2.12e). However, the resulting 3DVar increment (analysis minus background) is only positive in southern Arkansas and is neutral and even negative along the Arkansas-Oklahoma border despite the positive innovations nearby (Fig. 2.12c). This shows that the 3DVar increment did not correct the background errors in the warm front location and temperature gradient in this area.

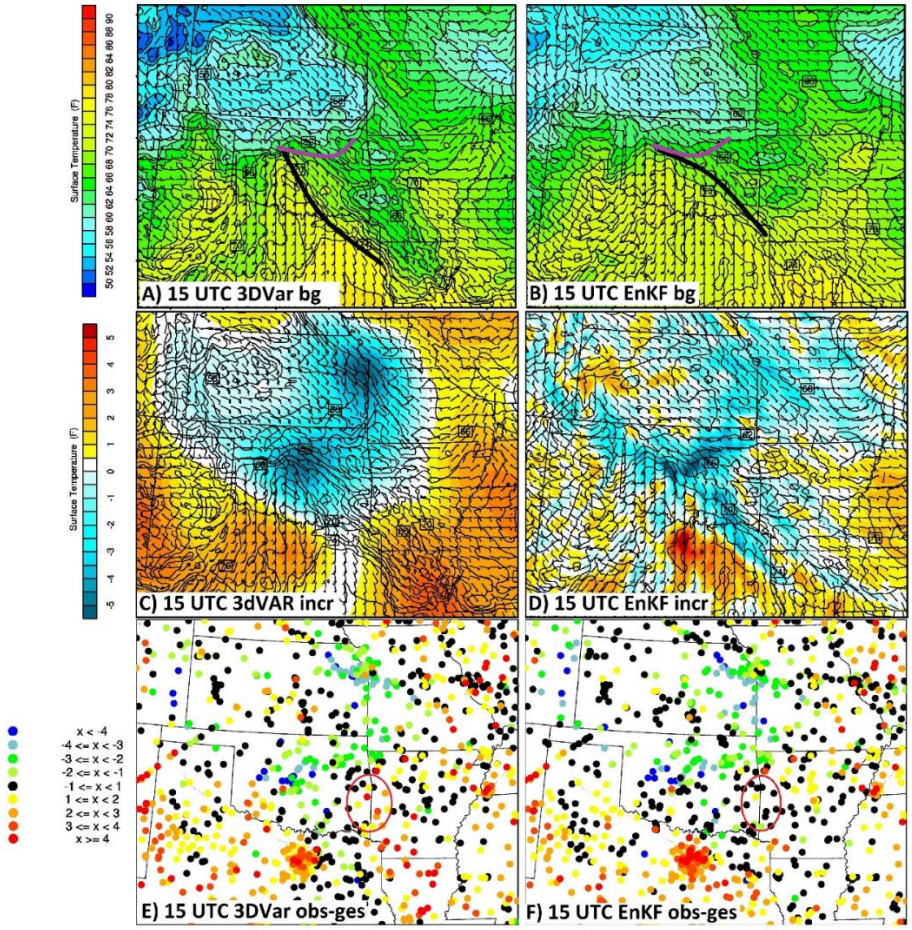

Figure 2.12: 15 UTC 20 May 2010 outer domain analysis, plotted over the region of the inner domain, for (a) 3DVar background temperature (shading; F) and wind (barbs; knots), (b) as in (a) except for EnKF, (c) 3DVar increment (analysis minus background) for temperature (shading; F) with first guess temperature contours and wind barbs overlaid, (d) as in (c) except for EnKF, (e) 3DVar surface temperature innovations (observation minus background; K), and (f) as in (e) except for EnKF. The thick black and purple lines in (a),(b) represent approximate locations of the warm front and cold pool boundary, respectively. The red circle in (e),(f) highlights an area where the too-cold 3DVar background forecast was not correct by the 3DVar data assimilation.

The inconsistency between the observation innovation and analysis increment along the Arkansas-Oklahoma border for 3DVar is attributed to the static background error covariance. Although the background error covariance length scales were tuned to perform well on average the spatial correlation does not reflect the shape and spatial extent of the relevant mesoscale features such as cold pools and warm fronts for this particular case.

Both 3DVar and EnKF are warmer than the observations in northern and central Oklahoma (Fig. 2.12e,f). This background error is a result of inadequate resolution of an MCS and associated cold pool (Fig. 2.12a,b; thick purple line) in northern Oklahoma on the outer domain. Both 3DVar and EnKF also have negative innovations in southeastern Oklahoma just south of the warm front (Fig. 2.12e,f). The corresponding EnKF increments are focused along the east-west oriented temperature gradient in northern Oklahoma, effectively enhancing and shifting southward the cold pool boundary (Fig. 2.12d). The EnKF increments are also elongated along the wind shift and temperature gradient in southeastern Oklahoma associated with the warm front (Fig. 2.12d). In contrast to the flow-dependent shape and localized spatial scale of the EnKF increments, the 3DVar surface temperature increments are isotropic and too large in scale for the mesoscale cold pool and warm front. As a result, the 3DVar increment in Oklahoma does not show spatial structure corresponding to these features (Fig. 2.12c). Due to the relatively large number of observations in central Oklahoma the impact of the negative 3DVar increment extends to the Arkansas-Oklahoma border (Fig. 2.12e). This limits the ability of the sparse observations along the Arkansas-Oklahoma border to adequately correct the too cold background forecast in this area.

Thus, the 3DVar increments do not sufficiently correct the error in the location of the warm front, along which spurious precipitation develops in the subsequent forecast.

In summary, the EnKF with flow-dependent background error covariance provides more physically reasonable analysis increments to correct the mesoscale first guess than 3DVar. As a result, a more accurate synoptic/mesoscale analysis is produced by EnKF at the end of the DA period (i.e., 0000 UTC).

2.5b Multi-scale analysis evaluation

The mesoscale analyses evaluated in the previous sub-section provide the background for the inner domain storm scale radar DA. In this section, the differences between GSI-based 3DVar and EnKF for the resulting multi-scale analyses are evaluated. Also representative of the systematic results, the storm scale radar DA for this case shows consistently smaller reflectivity and velocity first guess errors for EnKF than for 3DVar (Fig. 2.13). This indicates a better analysis of the storm scale features for EnKF than 3DVar, in addition to the synoptic/mesoscale features evaluated above.

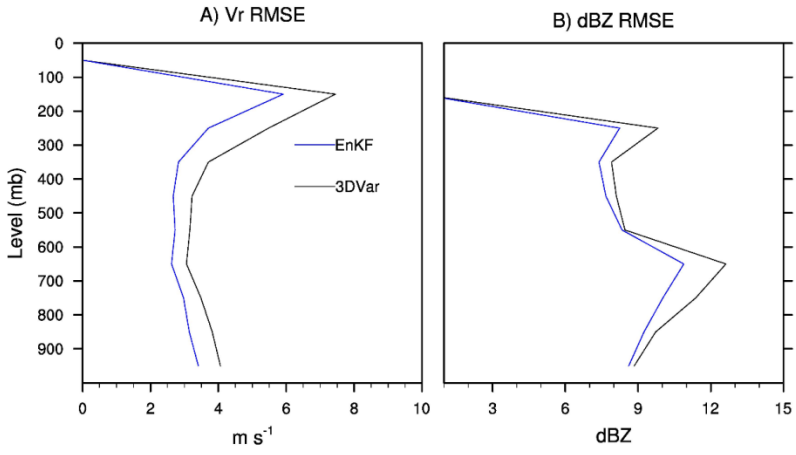

Figure 2.13: As in Fig. 2.8, except for the 20 May 2010 case study.

Forecasts initialized from the multi-scale analyses (Fig. 2.14) further reveal differences in the impact of the storm scale analyses on the subsequent precipitation forecasts, compared to the mesoscale analyses (Fig. 2.10). The storm scale radar DA results in an improved forecast during the first hour for both EnKF and 3DVar because of the reduced spin up time for the MCS (Fig. 2.14a,i). The reduced spin up time is a result of the convective systems already being present at the initialization time. For EnKF, the subjective improvement resulting from storm scale radar DA persists through the 8 hour forecast period (Fig. 2.14a-h). However, for 3DVar the initial storms are not maintained in the forecast and the spurious precipitation in eastern Oklahoma and western Arkansas is further enhanced (Fig. 2.14i-p). Further diagnostics show that the enhancement of spurious convection is due primarily to convergence resulting from the cold pools emanating from the convection in central Oklahoma (Fig. 2.11). At later lead times the subjective differences between the forecasts initialized with multi-scale analyses (Fig. 2.14h,p) are similar to the differences between the forecasts initialized with mesoscale analyses (Fig. 2.10h,p). This shows the increasing impact of the mesoscale analysis at later lead times, compared to the impact of the storm scale analysis at earlier lead times. The overall result for 3DVar is that after an initial improvement during the first hour, the forecast is similar or even degraded by the radar DA, especially at forecast hours 2-6. This contrasts with EnKF which shows a subjectively improved forecast at all lead times resulting from the radar DA. This result is also consistent with the systematic impacts of the storm scale radar DA for 3DVar discussed in the previous section (e.g., Fig. 2.7).

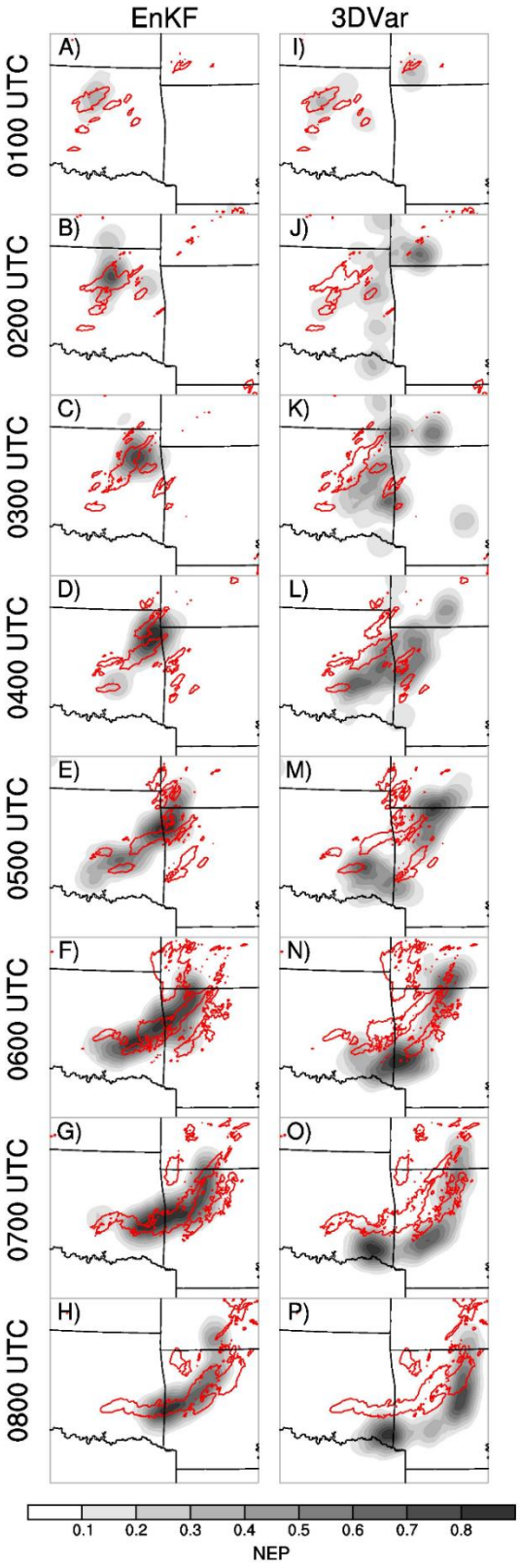

Figure 2.14: As in Fig. 2.10, except for forecasts initialized from the multi-scale analyses with further radar data assimilation.

The generation of excessive and unrealistic cold pools in the 3DVar analysis is illustrated by the early evolution of the initial supercell in western Oklahoma at ~2100 UTC. Both the EnKF and 3DVar techniques are able to increase reflectivity associated with this storm at 2105 UTC that is missing in the first guess (Figs. 2.15a,b and 2.16a,b). However, only the EnKF analysis also adjusts other fields such as vertical velocity, temperature and humidity to create a weak updraft and more saturated environment along with the added reflectivity (Fig. 2.15a,b,c,d). As a result, a buoyant updraft develops and maintains the convection during the subsequent cycles for EnKF (e.g., Fig. 2.15e,f,g,h). However, for 3DVar the added reflectivity does not correspond to increased humidity in the vicinity of the increased reflectivity (Fig. 2.16b). Therefore, a weak downdraft forms during the subsequent cycle as a result of precipitation loading and evaporative cooling (Fig. 2.16g). During subsequent cycles the subsidence results in column stabilization with net warming above ~3km due to adiabatic descent and net cooling of up to 1-2 K within just 10 minutes at lower levels where the evaporation of precipitation dominates (Fig. 2.16c,g,k).

In summary, the excessive storm scale cold pools in the 3DVar multi-scale analyses are a result of the lack of coherent cross-variable correlation in the static background error covariance for storm scale reflectivity assimilation. When hydrometeors are added to unsaturated locations, without corresponding increments to the dynamic and thermodynamic variables, much of the added hydrometeors may evaporate, creating or enhancing evaporative cooling and downdrafts. The storm scale differences between the EnKF and 3DVar analyses dominate the subsequent precipitation forecasts for several hours. At later lead times (e.g., ~5-8h), the mesoscale

differences discussed in the previous sub-section have a greater impact on the precipitation forecasts than the storm scale differences.

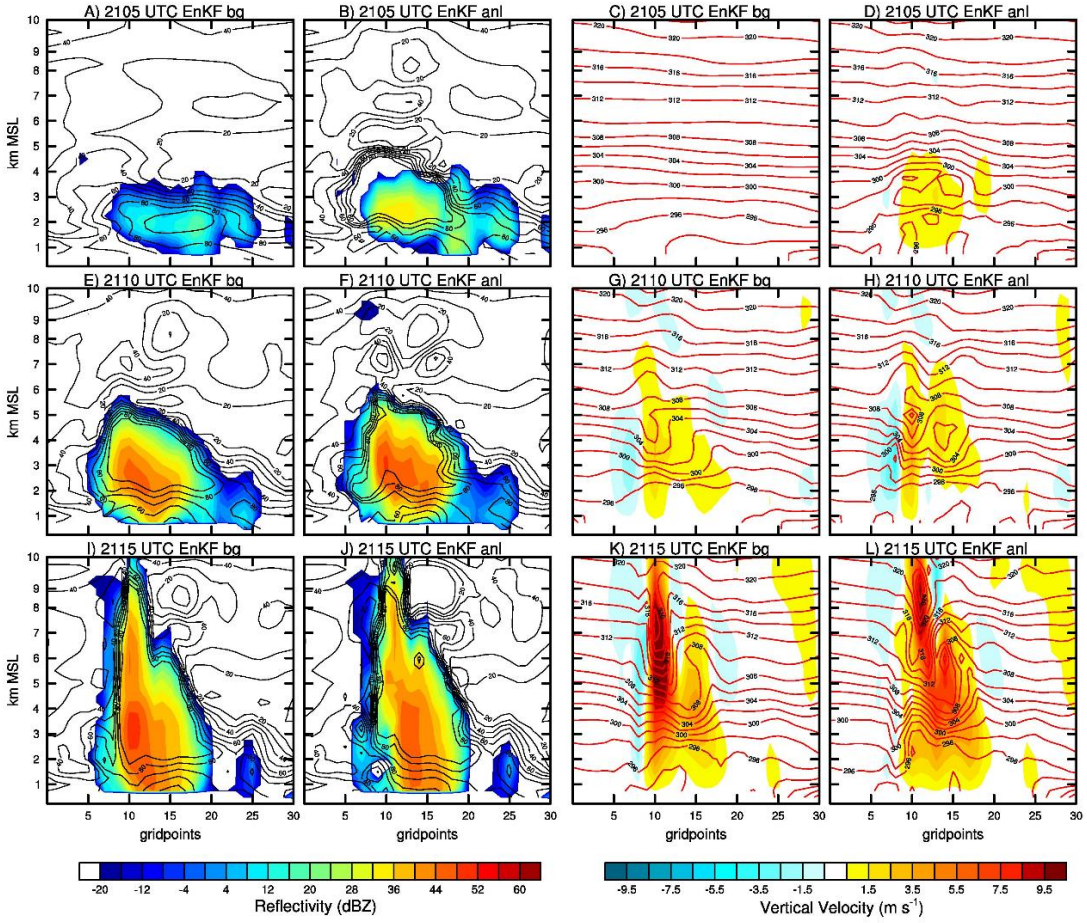

Figure 2.15: Cross section through a supercell in western OK (not shown) as it is assimilated during the first 3 radar DA cycles using GSI-EnKF. Panels (a)-(c) show the first guess reflectivity (shading) and relative humidity (contours) at 2105, 2110 and 2115 UTC, panels (d)-(f) are as in (a)-(c), except for the analysis fields, panels (g)-(i) are as in (a)-(c), except for vertical component of wind (shading) and potential temperature (contours) and panels (j)-(l) are as in (g)-(i), except for the analysis fields.

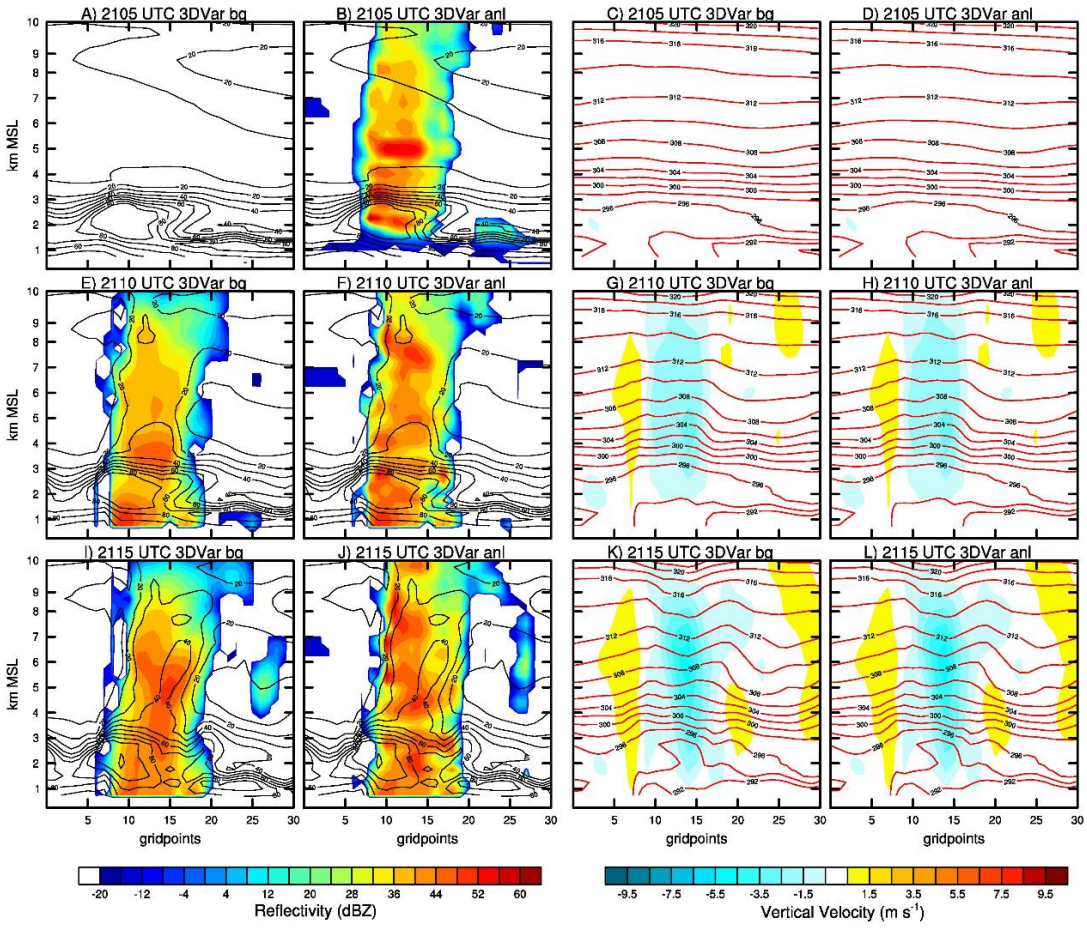

Figure 2.16: As in Fig. 2.15, except for GSI-3DVar.

2.6 Summary and conclusions

The accuracy of storm scale precipitation forecasts depends not only on processes at the storm scale but also on the mesoscale and synoptic scale environment supporting them. Therefore, accurate forecasts for convective scales require DA systems to properly estimate the atmospheric state on multiple scales. In order to perform multi-scale DA, the GSI-based DA system, including both 3DVar and EnKF, is extended to directly assimilate radar observations, in addition to the capability to assimilate synoptic/mesoscale observations. In this Chapter, the newly extended multi-scale GSI-based DA system is used to compare 3DVar and EnKF in the context of

multi-scale DA where scales ranging from convective scales to synoptic scales are resolved by both the model and the observations. The purpose of such a comparison is to facilitate understanding of how the differences among DA techniques lead to analysis differences at different scales and their subsequent impact on storm scale precipitation forecasts.

The comparison of GSI-based EnKF and 3DVar is performed systematically over 10 diverse convectively active cases in the central United States. The goal is a robust evaluation of the differences between the EnKF and 3DVar techniques for producing analyses at multiple scales. The mesoscale analyses obtained from assimilation of synoptic/mesoscale observations on the outer domain provide estimates of the synoptic/mesoscale environment for the storm scale radar DA. Multi-scale analyses result from the further storm scale assimilation of radar observations on the inner domain. Comparison of forecasts initialized from the mesoscale and multi-scale analyses differentiates the impacts of the different spatial scales on the subsequent precipitation forecasts.

The convection-permitting precipitation forecasts initialized from the multi-scale analyses are more skillful with GSI-based EnKF than GSI-based 3DVar for two reasons. First, precipitation forecasts initialized from the mesoscale analyses become more skillful with EnKF than 3DVar after about 3h. This suggests that the synoptic/mesoscale environment is more accurately analyzed by EnKF than 3DVar. Second, the improved forecast skill at early lead times resulting from the further inner domain storm scale radar DA lasts about five hours for EnKF and only one hour for 3DVar. This suggests that the analysis of storm-scale features is also more accurate

using EnKF than 3DVar. After the first hour the forecast initialized from the 3DVar analysis is actually degraded by the storm scale DA (Fig. 2.7). The greater benefit of the storm scale DA at early lead times for EnKF, together with the forecast degradation and inferior synoptic/mesoscale environment at later lead times for 3DVar, explains the systematically better forecasts initialized from the GSI-based EnKF multi-scale analyses compared to the GSI-based 3DVar multi-scale analyses.

A case study of upscale growth of cellular convection into an MCS during the evening/overnight hours of 19/20 May 2010 is used to qualitatively understand the systematic differences. Convection-permitting forecasts initialized from the EnKF mesoscale analyses show a subjectively better MCS forecast than forecasts initialized from the 3DVar mesoscale analyses. The forecast difference is largely due to a difference in the warm front analysis including location, temperature gradient and convergence. These differences in the mesoscale analyses occur because of the lack of flow-dependence in the static 3DVar background error covariance. Storm scale radar DA on top of the mesoscale DA (i.e., multi-scale DA) alleviates under-forecasting of precipitation during the first hour for both 3DVar and EnKF. However, only the EnKF forecast properly maintains the initial storms, leading to a subjective improvement over the forecast initialized from the mesoscale analysis throughout the forecast period. The initial 3DVar storms quickly collapse and generate unrealistically strong cold pools as a result of the lack of cross-variable correlations in the static background error covariance for hydrometeors. Further diagnostics revealed that the reflectivity observations assimilated with 3DVar are successfully able to correct errors in the precipitation hydrometeor fields. However, corresponding increments to vertical velocity,

temperature and humidity are not obtained. This results in substantial evaporation of hydrometeors that are added to sub-saturated regions in the background field. The evaporative cooling generates excessive cold pools during both the DA and forecast periods for 3DVar. Both the storm scale and mesoscale analysis differences contribute to the better EnKF forecast initialized from the multi-scale analysis. Consistent with the systematic results, the storm scale analyses dominate the precipitation forecasts at early lead times while the synoptic/mesoscale environment analyses dominate at later times. These results support the hypothesis that skillful convective scale precipitation forecasts require effective multi-scale DA methods.

Further development of the static background error covariance for radar reflectivity DA with 3DVar is clearly needed. This study also suggests that if ensemble estimates of the background error covariance are affordable, then using ensemble-based covariance in variational radar DA systems provides a straightforward solution. This method is commonly referred to as hybrid DA (Wang et al. 2008a,b, 2013b). Even the static covariance constructed with the simple method in this study has some useful aspects for reflectivity DA. For example, the initial ensemble downscaled from mesoscale analyses may have very small or zero variance of hydrometeors, limiting the impact of the assimilated reflectivity observations with EnKF. While studies such as Dowell et al. (2011) have alleviated this issue by adding random noise where observations indicate precipitation should be occurring, making use of the 3DVar static covariance model provides an alternative method (Carley 2012). Compared to EnKF, the static covariance model more quickly and effectively adds reflectivity that is completely absent from the first guess field. This is evident in larger RMS first guess

errors for reflectivity during the first few forecast cycles for EnKF than for 3DVar (not shown). Furthermore, EnKF can take several cycles for physically reasonable cross-variable correlations with the hydrometeors to develop in the flow-dependent background error covariance (Tong and Xue 2005). As also noted in Caya et al. (2005), this spin up time motivates additional research on hybrid methods to take advantage of both the reduced spin up time of a static background error covariance, and the improved forecast performance of the EnKF flow-dependent background error covariance.

Chapter 3: Multi-scale characteristics and evolution of perturbations: Dependence on background flow and method of perturbation

3.1 Introduction

The characteristics of the evolution of different types of perturbations have yet to be systematically studied over an extended period of many convection-allowing forecasts. The purpose of this chapter is to provide such a systematic evaluation of convection-permitting perturbation evolution and characteristics. Furthermore, two case studies are evaluated to expand on the types of flow regimes considered in past case studies. In contrast to the Mesoscale Alpine Program cases studied by Walser et al. (2004) and Hohenegger et al. (2006), this study focuses on the Great Plains of the United States where topography plays a less dominant direct role, severe convective weather is more frequent and intense (Brooks et al. 2003), and the latitude is farther south from the main belt of the westerlies.

Given the range of scales that can be resolved using convection-permitting resolution, the growth and interaction of perturbations on different scales is of particular interest. The multi-scale evolution of perturbations in a convection-permitting model has been considered in even fewer case studies than convection-permitting forecast predictability in general and has also not been considered systematically (Zhang et al. 2003, 2006; Walser et al. 2004; Luo and Zhang 2011). Therefore, the present study focuses on the multi-scale impacts of the perturbations, using Harr wavelet decomposition.

A few additional deterministic forecasts were generated by the Center for Analysis and Prediction of Storms (CAPS) during the 2010 National Oceanographic and Atmospheric Administration Hazardous Weather Testbed (NOAA HWT) Spring Experiment (Kong et al. 2010, Xue et al 2010, Clark et al 2012) to complement the CAPS Spring Experiment real-time SSEF system. These additional forecasts were designed to study the sensitivity to small scale IC perturbations which were not included in the CAPS SSEF design. This study of the multi-scale sensitivity to small scale IC perturbations has three main goals. The first goal is to determine the sensitivity to small scale IC perturbations, relative to the sensitivity to larger scale IC and physics perturbations that are already included in the SSEF design. The second goal is to compare two possible methods of generating such small scale IC perturbations. The third goal is to compare the perturbations in the existing ensemble to one method of combining small scale IC perturbations with the large scale IC and physics perturbations. These goals are addressed using two case studies with different background flows and systematic evaluation of all 34 available cases. Since the existing method of perturbation actually includes multiple perturbation sources (IC and physics), additional forecasts were later generated for the two case studies, with physics perturbations excluded, to aid interpretation of the results.

The WRF model configuration and scale decomposition and perturbation methods are described in section 3.2. Results are presented in section 3.3. Section 3.4 contains conclusions and a discussion.

3.2 Methods

3.2a WRF model configuration

The forecasts evaluated in this chapter were generated with 4 km grid spacing at 0000 UTC on 34 weekdays from 3 May to 18 June 2010 by CAPS for the 2010 NOAA HWT Spring Experiment (Xue et al. 2010a,b; Kong et al. 2010). These forecasts provide a readily available data set for an initial evaluation of the sensitivity of the precipitation forecasts to various perturbation methods. The control (i.e., unperturbed) forecast used the WRF ARW model (Skamarock et al. 2005). The control forecast ICs were obtained from the operational National Centers for Environmental Prediction's North American Model (NCEP NAM) 0000 UTC NAM Data Assimilation System (NDAS; Rogers et al. 2009) analysis at 12 km grid spacing, interpolated to the 4 km WRF ARW grid. Additional radar and mesoscale observations were then assimilated using Advanced Regional Prediction System (ARPS) 3DVar and cloud analysis package (Xue et al. 2003; Gao et al. 2004; Hu et al. 2006). Radial velocity from over 120 radars in the Weather Surveillance Radar (WSR)-88D network, as well as surface pressure, horizontal wind, potential temperature, and specific humidity from the Oklahoma Mesonet, METAR (Meteorological Aviation Report), and Wind Profiler networks were assimilated with ARPS 3DVar. The ARPS cloud analysis package uses radar reflectivity along with surface data, Geostationary Operational Environmental Satellite (GOES) visible and 10.5 micron infrared data to estimate hydrometeor species and adjust in-cloud temperature and moisture (Hu et al., 2006). The control forecast was configured with the Thompson et al. (2008) microphysics scheme, the Mellor-Yamada-Janic (Janjic´ 1994) boundary layer scheme, the Rapid Radiative Transfer

Model longwave radiation scheme (Mlawer et al. 1997), the Goddard shortwave radiation (Tao et al. 2003) scheme and the NCEP-Oregon State University-Air Force-NWS Office of Hydrology (NOAH; Ek et al. 2003) land surface model. The vertical turbulent mixing was represented in the boundary layer scheme and sub-grid scale horizontal turbulence mixing was represented by Smagorinsky parameterization. No additional numerical diffusion was applied.

3.2b Forecast perturbation method

In the general design of the SSEF during the 2010 HWT Spring Experiment, perturbations that sample model and physics uncertainty as well as IC and Lateral Boundary Condition (LBC) perturbations derived from the Short Range Ensemble Forecast system (SREF; Du et al. 2009) are included. Since the SREF was run at grid spacing of 32-45 km (corresponding to a wavelength of 64-90 km; Du et al. 2009), SREF perturbations are on scales much larger than the SSEF model resolution. Therefore the perturbations from SREF do not include small scales (i.e., order of tens of kilometers). Methods to generate perturbations on multiple scales, ranging from the synoptic to the convective scales, have yet to be systematically studied. During the 2010 Spring Experiment additional forecasts were generated with small scale IC perturbations as a first step to help guide development of practical methods of sampling errors across multiple scales in a SSEF system. For each perturbation method described below, one perturbed deterministic forecast was generated and compared to the control member forecast.

Six methods of perturbation are investigated in this study. Perturbations RAND (random) and RECRS (recursive filter) are designed to simulate random small scale errors in the initial state. Perturbation LGPH (large scale and physics) is designed to simulate the medium and large scale (i.e., order of hundreds and thousands of kilometers, respectively) IC errors and model physics errors that are currently sampled in the CAPS SSEF system. Perturbation LGPH_RECRS (large scale and physics with recursive filter) is a combination of the LGPH and RECRS perturbation methods. For the two case studies, two additional perturbations are evaluated. Perturbations LG and LG_RECRS are identical to LGPH and LGPH_RECRS, respectively, except without any physics differences from the control member.

The RAND perturbation is obtained by adding spatially uncorrelated, Gaussian random numbers to the IC temperature and relative humidity (standard deviation of 0.5 K and 5%, respectively). The RECRS perturbation is obtained similarly, except with a recursive filter applied to the random perturbations to create spatially correlated perturbations with a 12 (3) km horizontal (vertical) de-correlation scale. The RAND perturbation is conceptually similar to the random perturbations of Hohenegger and Schär (2007a). The RECRS perturbation is conceptually similar to the Gaussian perturbation of Hohenegger and Schär (2007a), except RECRS is applied homogenously across the domain instead of only at a single location.

The LGPH IC perturbation is obtained from the difference between a 3 hour forecast of a SREF WRF-ARW member (labeled P1 in Du et al. 2009) and the corresponding SREF control member forecast. The SREF perturbations of u and v wind components, potential temperature, and specific humidity are rescaled to have a root

mean square value of 1 m s^{-1} , 0.5 K, and 0.02 g kg^{-1}, respectively. In addition to the IC perturbation, the LGPH forecast uses a different physics configuration than the control forecast to approximate physics errors. Unlike the control forecast (Section 3.2a), the LGPH perturbation uses Morrison et al. (2008) microphysics scheme, RUC land surface model (Benjamin et al. 2004) and Yonsei University (Noh et al. 2003) boundary layer scheme. The LGPH_RECRS perturbation is identical to LGPH except with additional recursive filtered random perturbations added in the same way as for the RECRS perturbation.

Although only temperature and humidity (and wind in the case of LGPH and LGPH_RECRS) are directly perturbed, results are evaluated in terms of precipitation differences. Thus, the focus is on the net effect, rather than the processes, of perturbation growth and evolution for the purpose of precipitation forecasting at convection-allowing resolution.

3.2c Scale decomposition method

Following Casati et al. (2004), precipitation fields are decomposed into components of different spatial scale using 2D Haar Wavelets with the Model Evaluation Tools package from the Developmental Testbed Center, available at http://www.dtcenter.org/met/users. The decomposition is defined over a 2^n by 2^n grid point domain for n>1. The original field is decomposed into its component on each of n+1 scales, and is equal to the sum of its components. The i^{th} component can be calculated as the difference between the original field averaged in boxes of 2^{i-1} by 2^{i-1} grid points and the original field averaged in boxes of 2^i by 2^i grid points for $1 \leq i \leq n$.

The $(n+1)^{th}$ component is the domain average value. Each component therefore represents the variation over a spatial scale of $4 * 2^{i-1}$ km from a larger scale average. Analogous to the more familiar Fourier decomposition, in the rest of the paper, the wavelet-decomposed spatial scales are referred to in terms of a corresponding wavelength. Thus, for example, the smallest resolvable scale of 4 km (e.g., Fig. 4.1b) corresponds to the smallest resolvable wavelength of 8 km. A verification domain (plotted in Fig. 3.3) of 512 by 512 grid points (2048 by 2048 km) within the larger forecast domain (shown in Fig. 3.2) of 1163 by 723 grid points (4652 by 2892 km) is used in this chapter. Further details of the wavelet decomposition are described in Casati et al. (2004). Precipitation forecast energy is defined as the square of the one-hour accumulated precipitation field, averaged over the verification domain. The energy on a particular scale is defined similarly, using only the component of the precipitation field on that scale. The error (or perturbation) energy is the square of the precipitation field difference between a forecast and the observations (or control forecast). The evolution of a perturbation, or difference, energy metric is a common method of quantifying sensitivity to forecast perturbations (e.g., Zhang et al. 2006, Hohenegger et al. 2006).

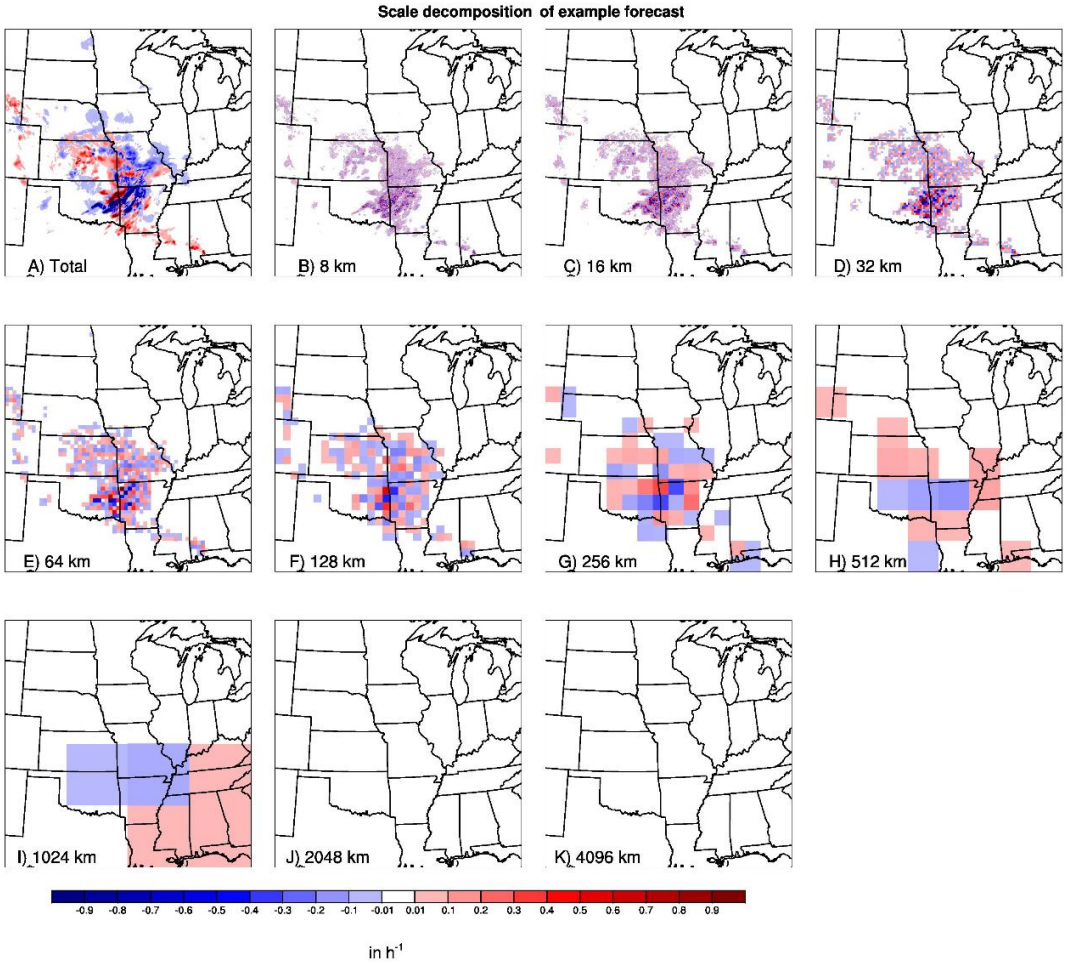

Figure 3.1: Difference between control forecast and observed 1-h accumulated precipitation, at 0600 UTC 20 May 2010 using forecast initialized at 0000 UTC 20 May 2010, showing (A) the total precipitation forecast and (B)-(K) the anomalies on each scale identified by the 2D Haar wavelet decomposition.

Figure 3.1 illustrates the 2D Haar wavelet decomposition of the difference between the 6h control forecast and corresponding observation of hourly accumulated precipitation on the 20 May case. The distribution of difference energy across scales is also found in Fig 3.8 (dashed cyan line). Objectively, there is a maximum of difference energy at 32-64 km wavelength scales and a smaller secondary maximum at the 256 km scale (Fig. 3.8). The total difference field (Fig. 3.1a) subjectively looks most similar to the difference fields on 32-64 km scales (Fig. 3.1d,e), suggesting that the high

amplitude, small-scale features on these scales account for most of the total difference. The subjectively apparent displacement of the MCS in Oklahoma and Arkansas (Fig. 3.1a) also corresponds to increased energy on the 256 km scale (Figs. 3.1g and 3.8).

For presentation of results it is convenient to reduce the number of considered scales from 10 to 3 by defining (in this chapter) the large scale as the sum of scales with wavelengths of 4096, 2048, and 1024 km, the medium scale as the sum of scales with wavelengths of 512, 256, 128 km and 64 km and the small scale as the sum of scales with wavelengths of 32, 16, and 8 km. The small scales are those that are too small to be represented with the current SREF-derived perturbations.

3.3 Characteristics of perturbation growth

The characteristics of the precipitation forecast perturbation evolution are evaluated using the change in perturbation energy with time in total and on the small, medium and large scales as well as the change in perturbation energy with spatial scale for selected fixed times. It should be noted that when perturbation characteristics are related to the background flow, the background flow refers to the control forecast upon which the perturbations were added, which may be different than the observations.

In theory, an optimal ensemble design would contain members that are equally plausible, and therefore equally skillful (Leith 1974). Although lower skilled members can still add value to an ensemble (Eckel and Mass 2005) and this study focuses on forecast sensitivity rather than forecast skill, the impact of the perturbations on forecast skill should also be considered when designing an ensemble system. Among the forecasts evaluated systematically in this study, only the physics perturbations at some

lead times (~2-5 h and ~22-27 h) and the RECRS perturbations during the first hour resulted in significant decreases in skill (not shown). The differences in skill resulting from physics perturbations are in large part related to differences in forecast bias. How to optimally sample model and physics error is still an open research question for SSEF design. The inclusion of LG and LG_RECRS perturbations in the case studies below helps to account for the impacts of different forecast biases. The early loss of skill resulting from recursive filter perturbations is a result of spurious precipitation that formed over large areas on many cases (not shown). This is clearly not desirable in an ensemble and it suggests that the spatial scales and amplitude of such perturbations should be more carefully studied before this perturbation method is used for ensemble forecasting.

The following case studies and season-average results address the three goals stated in section 3.1: a comparison of LGPH (and LG) with RAND and RECRS, a comparison of RAND with RECRS, and a comparison of LGPH_RECRS (and LG_RECRS) with LGPH (and LG).

3.3a 10 May 2010 case

A case study of forecasts initialized at 0000 UTC 10 May 2010 is selected because a synoptic scale baroclinic disturbance generated widespread precipitation. The convective evolution was determined primarily by large scale influences (e.g., fronts, jets and temperature advection). This event is also of interest because of a significant tornado outbreak that occurred in the southern Plains on the afternoon of 10 May (e.g., Palmer et al. 2011). At 0000 UTC 10 May there was an embedded

shortwave trough aloft over the western US and a broad ridge over the central US (Fig. 3.2a). There was southerly flow and a warm front in central Texas at the surface (Fig. 3.2b). By 0000 UTC the negatively-tilted shortwave had propagated to the central US, inducing surface cyclogenesis and an intersecting dryline, cold front and warm front in the southern Plains (Fig. 3.2c,d,e,f). An initial wave of observed scattered showers associated with the low-level warm advection developed in Arkansas and Missouri by 0600 UTC and moved eastward into Tennessee and northern Alabama by 1800 UTC (Fig. 3.3a,b,c). Convection also developed near the Kansas/Nebraska border by 1200 UTC, moving eastward into northern Missouri by 1800 UTC (Fig. 3.3b,c). At 0000 UTC 11 May more intense convection was occurring in the southern Plains (Fig. 3.3d).

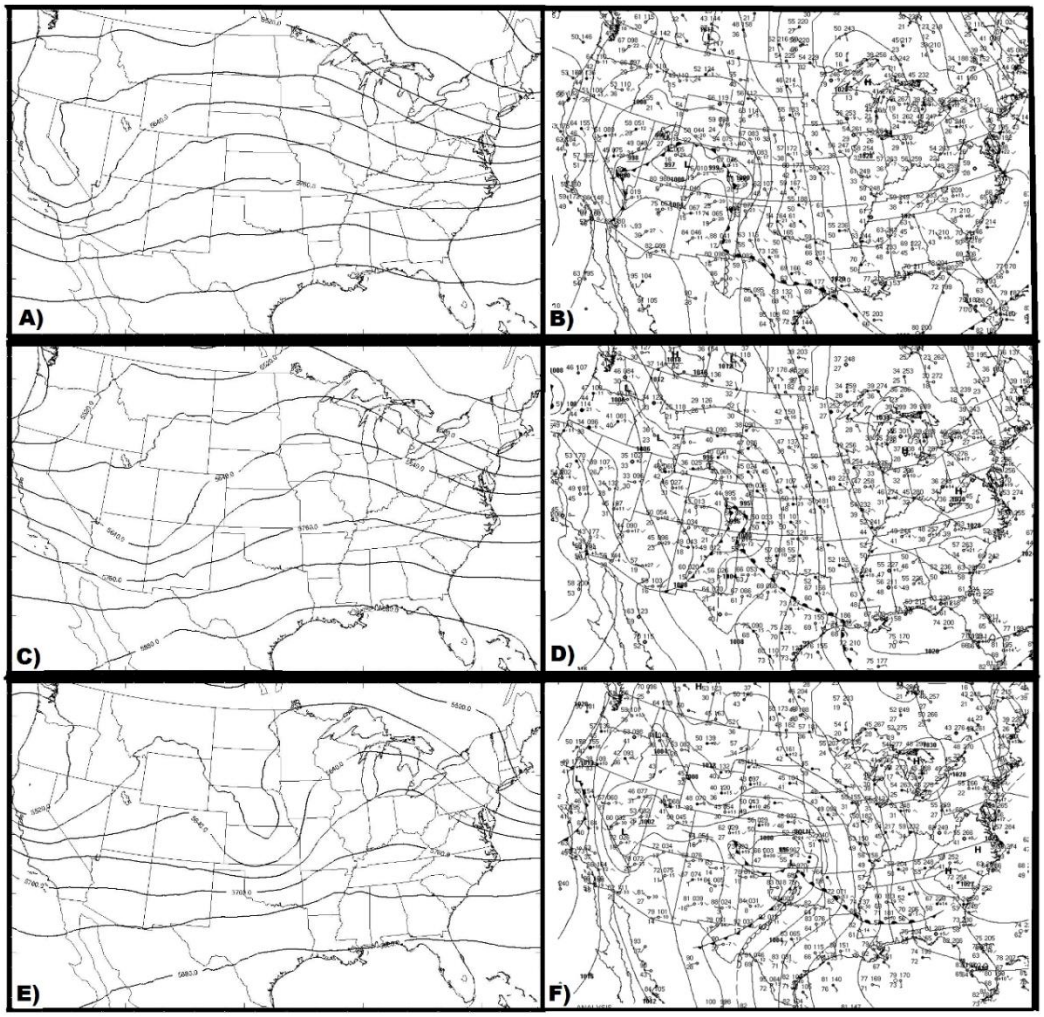

Figure 3.2: As in Fig. 2.4, except at (a),(b) 0000 UTC 10 May 2010, (c),(d) 1200 UTC 10 May and (e),(f) 0000 UTC 11 May.

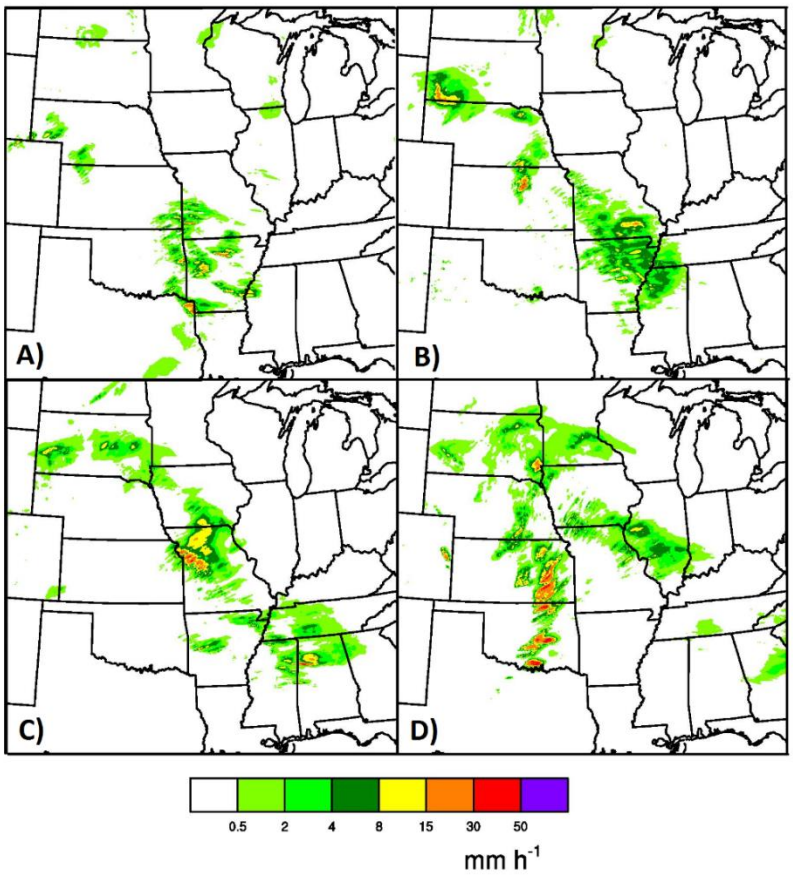

Figure 3.3: As in Fig. 2.5, except for (a) 0600 UTC 10 May 2010, (b) 1200 UTC 10 May, (c) 1800 UTC 10 May and (d) 000 UTC 11 May.

For this case, the control forecast predicted the initial wave of scattered showers, although with a southwestward displacement and with greater intensity than observed (Fig. 3.4a,b vs. Fig. 3.3a,b), as well as the development of convection along the Kansas-Nebraska border, although with more linear organization, weaker intensity and a slight northward displacement (Fig. 3.4b,c vs. Fig. 3.3b,c). The most prominent difference between the forecast and observation is the absence of the intense convection over the southern Plains at 0000 UTC (Fig. 3.4d vs. Fig. 3.3d). Storms eventually developed in the control forecast but they were several hours slower to develop than observed and did not extend as far south as observed (not shown).

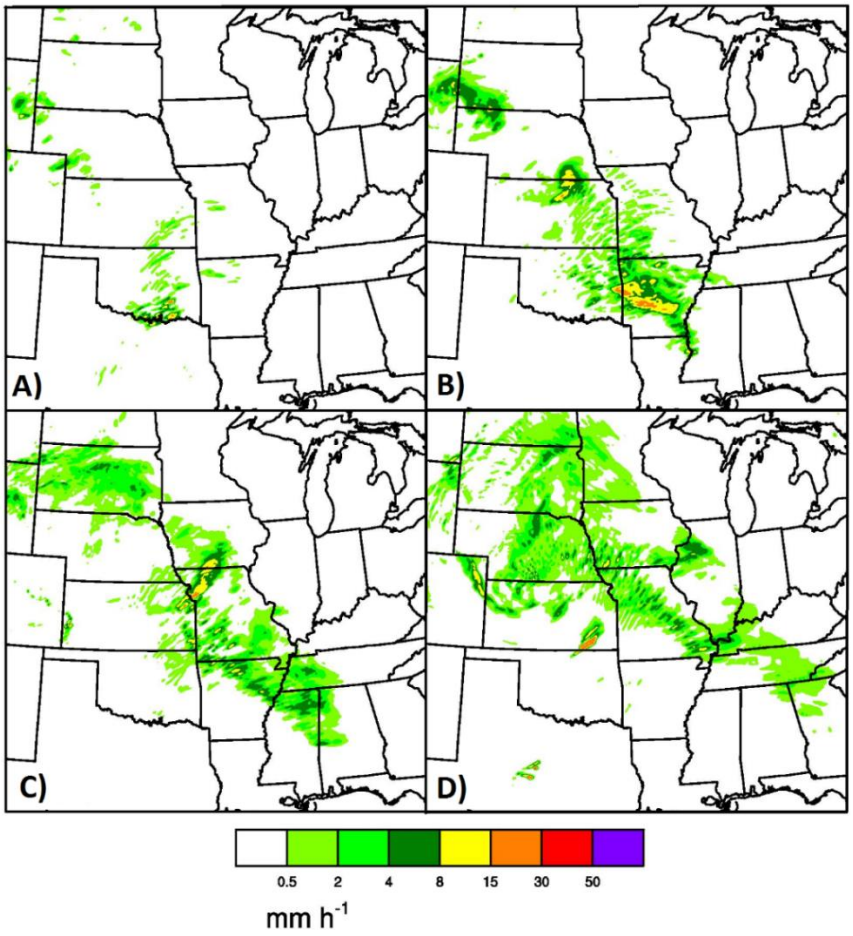

Figure 3.4: As in Fig. 3.3, except for the control forecast, rather than observations.

For the 10 May case, the control forecast error energy shows maxima in forecast error energy at lead times of about 10-15h and 24-27h (Fig. 3.5d). The general trend of two error energy maxima superimposed on an overall increasing trend is found on all scales (Fig. 3.5a-c). The magnitude of error energy is an order of magnitude greater on the medium and small scales than on the large scales. The first maximum occurs at a later time with decreasing scale ranging from ~9h on the large scale to ~15h on the small scale. The second maximum occurs at ~26h on all scales. In general, LGPH and LGPH_RECRS capture about half of the total error energy and RAND and RECRS capture about one quarter of the total error energy, which is dominated by the small and

medium scales (Fig. 3.5). Compared to the control forecast error energy, the perturbation energy for most lead times and methods is not only too small in magnitude (Fig. 3.5), but also too small in the spatial scale of maximum energy at all lead times for RAND and RECRS and through the 12h lead time for LGPH and LGPH_RECRS (Fig. 3.6). A particularly pronounced absence of medium scale perturbation energy with a scale of about 64-256 km at 24h for all perturbation methods, compared to forecast error, is consistent with Fig. 3.4. The medium scale storms in the southern Plains at this time (Fig. 3.3d) are absent in the corresponding forecast (Fig. 3.4d), contributing to the medium scale forecast error energy. However, all perturbation methods also missed these storms (not shown) so the perturbation energy does not reflect that particular forecast error.

The characteristics of the evolution of perturbation energy on different scales depend strongly on the method of perturbation for the 10 May case. LGPH shows more pronounced growth than RAND and RECRS for large and medium scales, but not for small scales (Fig. 3.5). When physics perturbations are excluded, LG perturbation energy is less than LGPH on medium and large scales at later lead times (Fig. 3.5). However, the qualitative comparison of RAND and RECRS to LGPH is consistent with the comparison to LG. RECRS shows an increase of perturbation energy over RAND on the medium scales and after ~20h on the small scales (Fig. 3.5b,c). When small scale perturbations are combined with the large scale IC and physics perturbations, LGPH_RECRS and LG_RECRS do not show much increase of perturbation energy compared to LGPH and LG, respectively (Fig. 3.5)

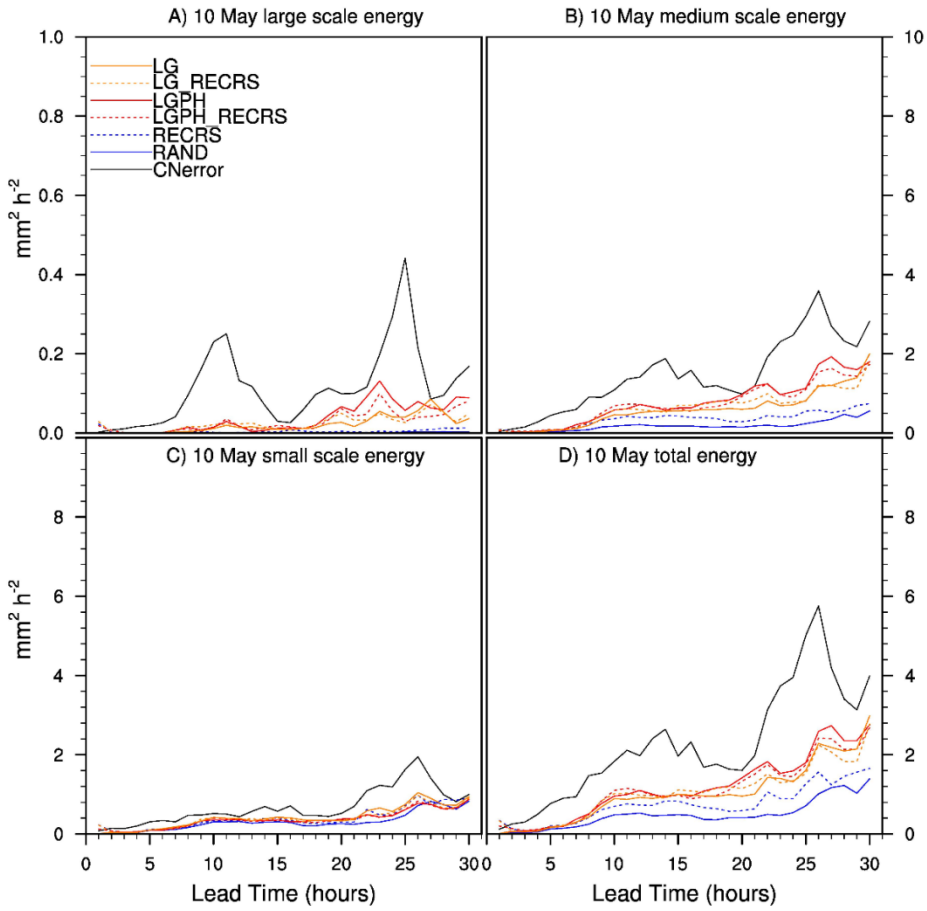

Figure 3.5: Average squared difference (i.e., energy) between control forecast and observed hourly accumulated precipitation (CNerror), and between each perturbed forecast and the control forecast, during the 10 May case for (a) large scales only, (b) medium scales only, (c) small scales only and (d) without any scale decomposition or filtering.

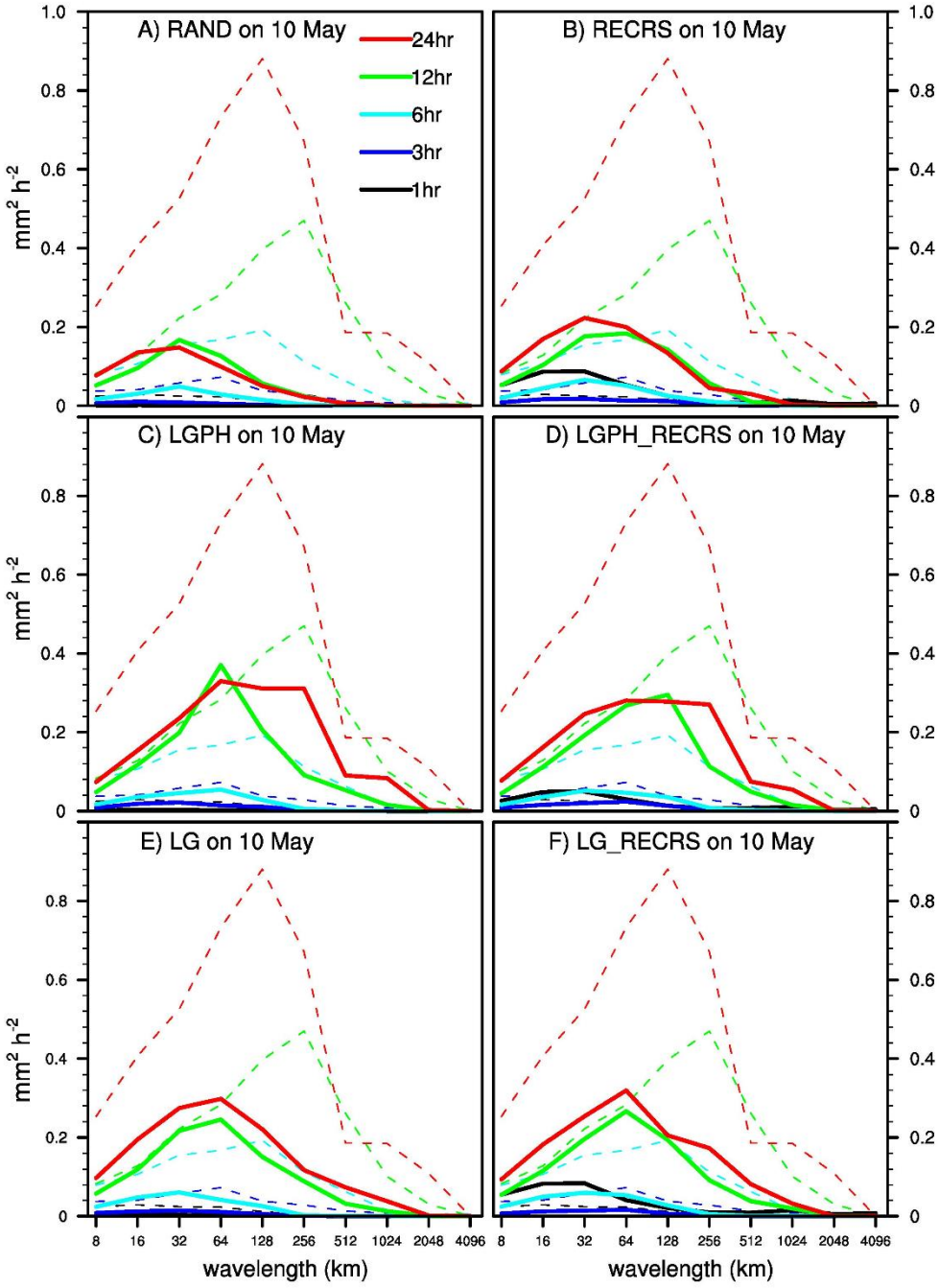

Figure 3.6: Perturbation energy as a function of wavelength for the 10 May case at lead times of 1, 3, 6, 12 and 24 h for (a) RAND, (b) RECRS, (c) LGPH, (d) LGPH_RECRS, (e) LG and (f) LG_RECRS. The CNerror energy is the dashed line in all panels.

The characteristics of perturbation growth are also seen in the perturbation energy spectra at selected lead times (Fig. 3.6). None of the perturbation methods generates much energy during the first 6h. The perturbation method affects both the spectral width and the wavelength of maximum energy of the resulting precipitation forecast perturbation. For example, at 12h the wavelength of maximum energy of 32 km for RAND (Fig. 3.6a) is smaller than the 64 km for LGPH (Fig. 3.6c) and is also smaller than the 64 km for RECRS (Fig. 3.6b). The LGPH spectrum at later (i.e., 12 and 24 h) lead times is broader than the spectra for RAND and RECRS (Fig. 3.6a,b,c), indicating perturbations across a wider range of scales. The RECRS spectrum is also broader than the RAND spectrum (Fig. 3.6a,b). The 128 km wavelength of maximum energy for LGPH_RECRS at 12 h (Fig. 3.6d) is even larger than the 64 km for LGPH (Fig. 3.6c). However, LG_RECRS also has a 64 km wavelength of maximum energy at 12 h (Fig. 3.6f).

The RECRS, LGPH_RECRS and LG_RECRS perturbation energy maxima at 16-32 km wavelength at 1h corresponds to the spurious small scale precipitation mentioned above. This spurious precipitation may be a result of adding unrealistically large perturbations on such scales, a lack of realistic coupling between the temperature and moisture observations, or some other imbalance resulting from the temperature and humidity perturbations in RECRS. The lack of spurious precipitation in the RAND perturbations may be a result of diffusion quickly reducing the amplitudes of the small scale perturbations when the perturbations are of grid scale.

In summary, for the 10 May case, the perturbation methods considered, especially small scale IC perturbations, do not reflect the forecast error magnitude or

temporal variability. The shape of the perturbation energy spectrum also does not reflect the shape of the forecast error energy spectrum for many lead times and perturbation methods. Compared to RAND and RECRS, LGPH and LG show greater medium and large scale perturbation growth, resulting in broader perturbation energy spectra with larger wavelength of maximum energy at some lead times. LG generally has less perturbation energy than LGPH at later lead times on medium and large scales. The difference between RAND and RECRS is seen mainly on medium scales with an increase of perturbation energy for RECRS, resulting in a broader perturbation energy spectrum at some lead times. The differences between LGPH_RECRS/LG_RECRS and LGPH/LG are generally small for most scales and lead times. The relative lack of medium and large scale forecast perturbations in RAND and RECRS suggests a relative insensitivity of this forecast on such scales to random small scale IC perturbations relative to larger scale IC and physics perturbations and, as shown below, relative to other cases.

3.3b 20 May 2010 case

In contrast to the 10 May case, the 20 May case is selected because early in the control forecast (i.e., about the first 12h) an MCS developed upscale from smaller scale convection present at the forecast initialization time. The details of this case are described in section 2.3. The control forecast reflects the upscale organization and intensification of convection, subsequent dissipation of the MCS, development of stratiform precipitation and regeneration of convection the following afternoon (Fig 3.7 vs. Fig. 2.5). However, the forecast MCS evolved a different structure than the

observed MCS (Fig. 3.7b vs. 2.5b). The coverage, timing and location of subsequent convection along the remnant outflow boundary was also qualitatively different than observed (Fig. 3.7d,e vs. 2.5d,e).

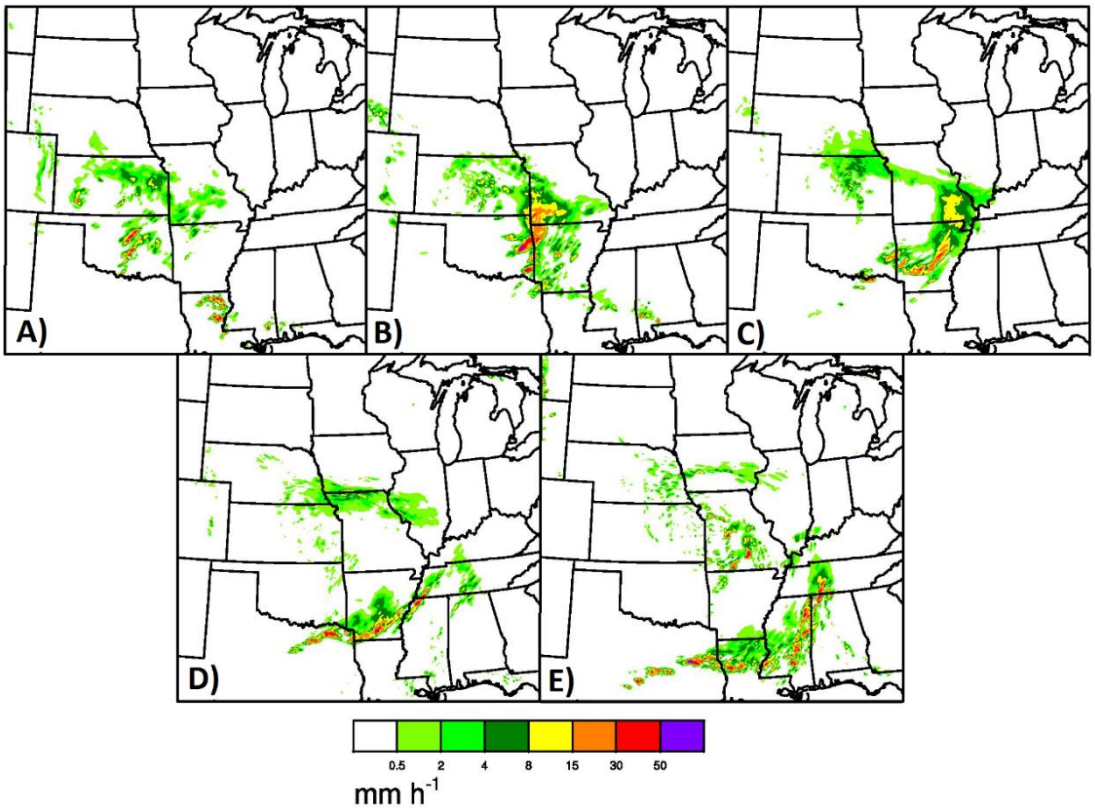

Figure 3.7: As in Fig. 2.5, except for the control forecast, rather than observations.

As in the 10 May case, the 20 May case also shows forecast error energy with a maximum at early lead times followed by a larger maximum at ~24-27h (Fig. 3.8d). In contrast to the 10 May case, the forecast error energy on 20 May does not clearly show a general increasing trend. This may be due to the already much larger error energy on the 20 May case than on the 10 May case at early lead times, especially on small and medium scales (Fig. 3.8b,c). Although the error energy during the first maximum is again under-represented by the forecast perturbations, the perturbation energy follows

the error energy more closely on this case during the second maximum than on the 10 May case. For hours 3 through 12, only RAND and LG (and RECRS at 12h) fail to capture the error energy maximum wavelengths of 32, 32-64 and 64 km at 3, 6 and 12h, respectively (Fig. 3.9). This contrasts with the 10 May case where all perturbation methods generated maximum error energy on smaller scales than the forecast error energy during the first 12 h. By 24h, all perturbation methods reflect the maximum error energy on the 64 km wavelength scale on the 20 May case.

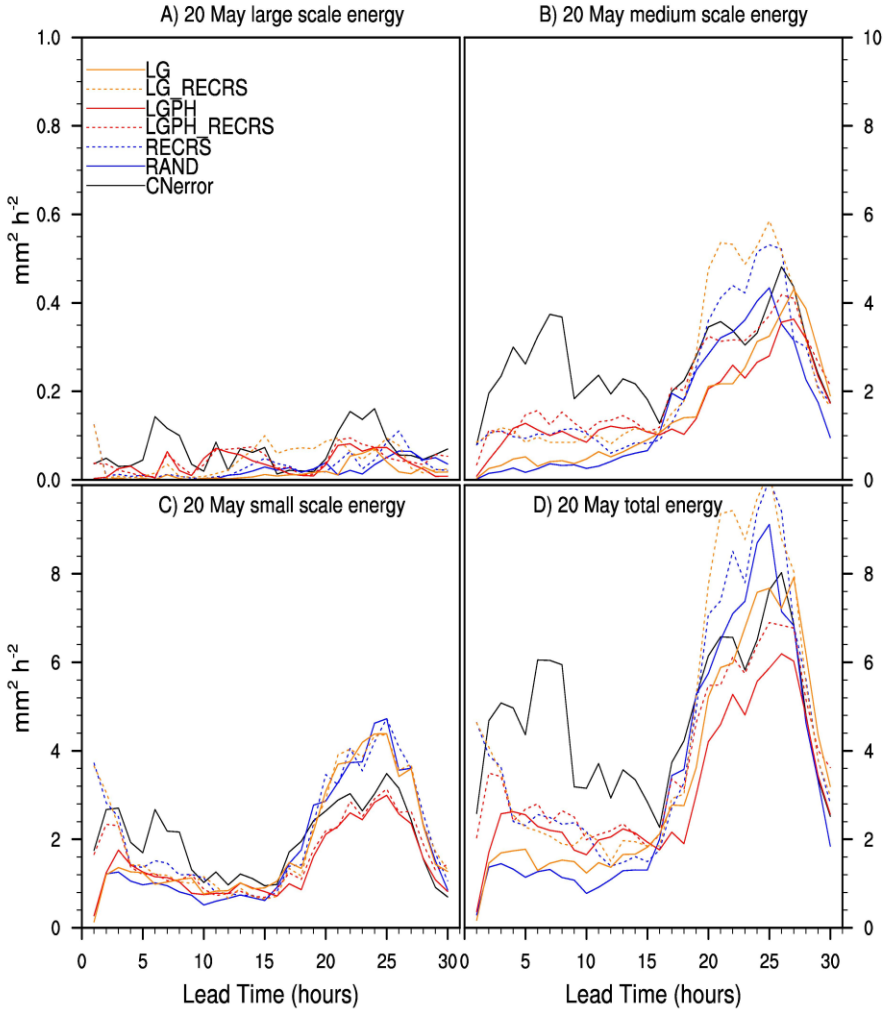

Figure 3.8: As in Fig. 3.5, except for the 20 May case.

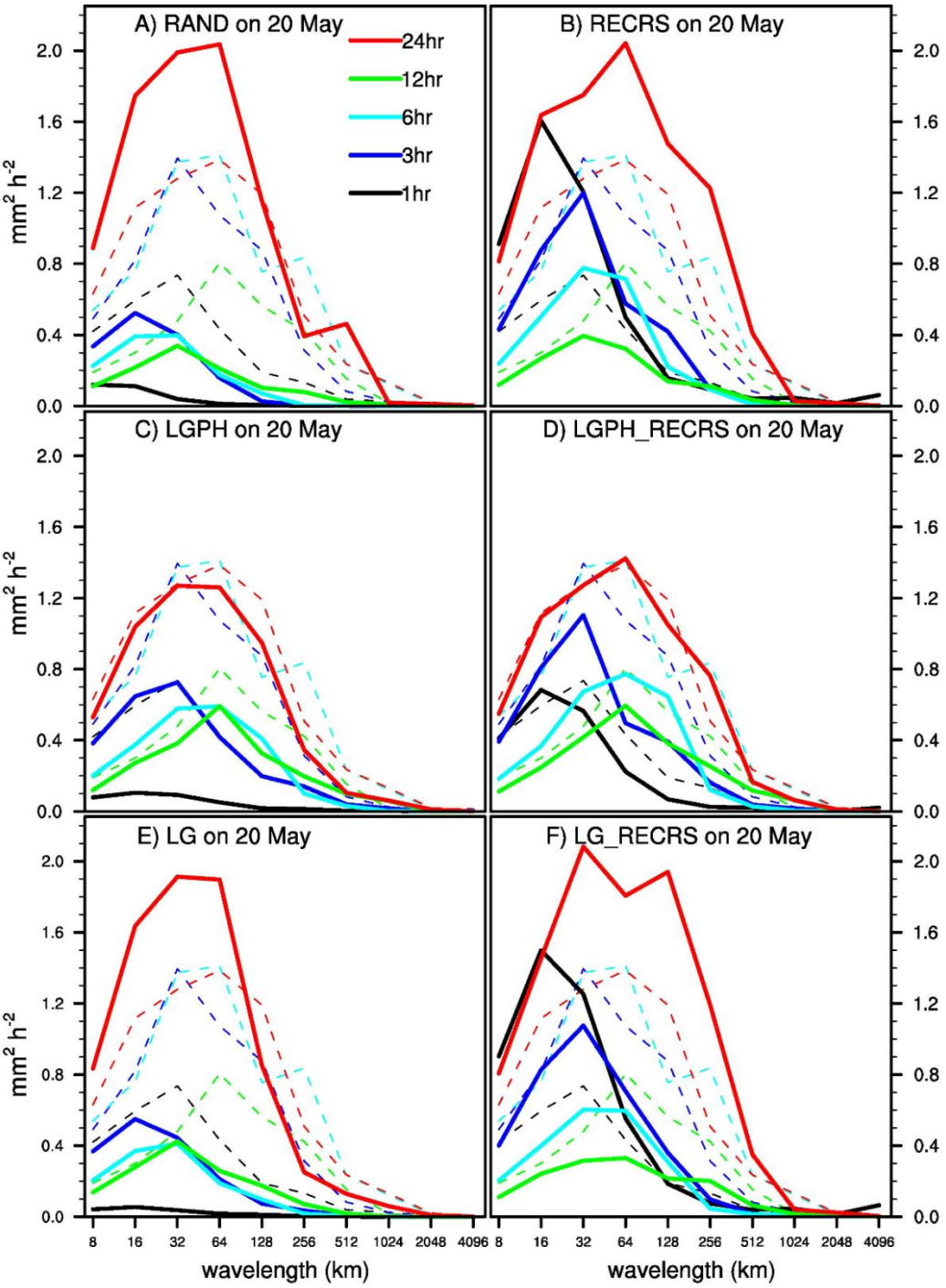

Figure 3.9: As in Fig. 3.6, except for the 20 May case.

The characteristics of the evolution of perturbation energy on the 20 May case are generally less dependent on the method of perturbation than on the 10 May case.

There is not a clear separation between LG/LGPH and RAND/RECRS on medium and large scales during most of the forecast period (Fig. 3.8). LGPH even shows less perturbation energy than RAND and RECRS on small scales at ~20-27 h. LGPH_RECRS is also less than RECRS alone at 1 h for large and small scales. Since LG_RECRS is more similar to RECRS at 1 h and LG is more similar to RAND at 20-27 h on small scales, these seemingly counter-intuitive results are due to a damping effect of the LGPH physics configuration. The physics configuration of LGPH also showed less systematic bias than RAND and RECRS at these lead times (not shown). It is not clear whether this damping effect is related to the differences in microphysics or boundary layer parameterization. The differences between LG and LGPH are most pronounced on medium scales at early lead times and small scales at later lead times for this case (Fig. 3.8b,c). LG and LGPH become similar after ~15h on the medium scales, suggesting that medium scale forecast sensitivity is dominated by the IC, rather than physics, perturbations at later lead times. LG and RAND have similar perturbation energy to each other and less than LGPH during the first 15h on the medium scales (Fig. 3.8b). During the early forecast hours RECRS has more perturbation energy than RAND on small and medium scales (Fig. 3.8b,c). In contrast to the 10 May case, this difference diminishes and RAND and RECRS become similar by about 10-12h. Also in contrast to the 10 May case, there are larger differences between LGPH_RECRS and LGPH and between LG_RECRS and LG on the 20 May case. LGPH_RECRS shows greater perturbation energy than LGPH at early lead times on small scales (Fig. 3.8c), most lead times on medium scales (Fig. 3.8b), and only the 1 h lead time, corresponding to regional variation in the spurious precipitation response to RECRS, on large scales

(Fig. 3.8a). These differences are even more pronounced when LG_RECRS is compared to LG.

Differences between the 10 and 20 May cases are further illustrated by the perturbation energy spectra which are also less sensitive to the perturbation method on 20 May than on 10 May (Fig. 3.9). Unlike the 10 May case, even the small scale RAND perturbation energy grows substantially during the first 3h (Fig. 3.9a). At 6 and 12 h the wavelength of maximum energy for LGPH and LGPH_RECRS is again larger than for RAND and RECRS (Fig., 3.9a,b,c,d). This difference is largely due to the physics perturbations since the LG and LG_RECRS spectra (Fig. 3.9e,f) at these times are much more similar to the RAND and RECRS spectra. The differences between RAND and RECRS spectra are generally limited to the first 6 h, when the impact of the early spurious precipitation has not yet diminished. The LGPH_RECRS and LG_RECRS spectra (Fig., 3.9e,f) also look very similar to the LGPH and LG spectra (Fig., 3.9c,d), respectively, after the first 6 h.

The different sensitivities of the 10 and 20 May cases to different perturbations are illustrated subjectively with representative RAND, LGPH and LG forecast perturbations at the 24h lead time (Fig. 3.10). On 10 May it is primarily the convective scale details of an incipient MCS over southeast Kansas, and the small scale features within the stratiform precipitation farther north that are substantially affected by the RAND perturbation (Fig. 3.10a). However, the LGPH perturbation alters the mesoscale structure of the stratiform precipitation region, and more dramatically changes the structure and location of the incipient MCS which is displaced ~100 km to the northwest (Fig. 3.10b). The LG perturbation also shows a displaced MCS but the

amount of displacement is much less than for LGPH (Fig., 3.10c). In contrast, even the mesoscale characteristics and location of the MCS forecast over the southern part of the domain on 20 May are substantially changed by the RAND perturbation (Fig. 3.10d) at least as much as the LG and LGPH perturbation (Fig. 3.10e,f).

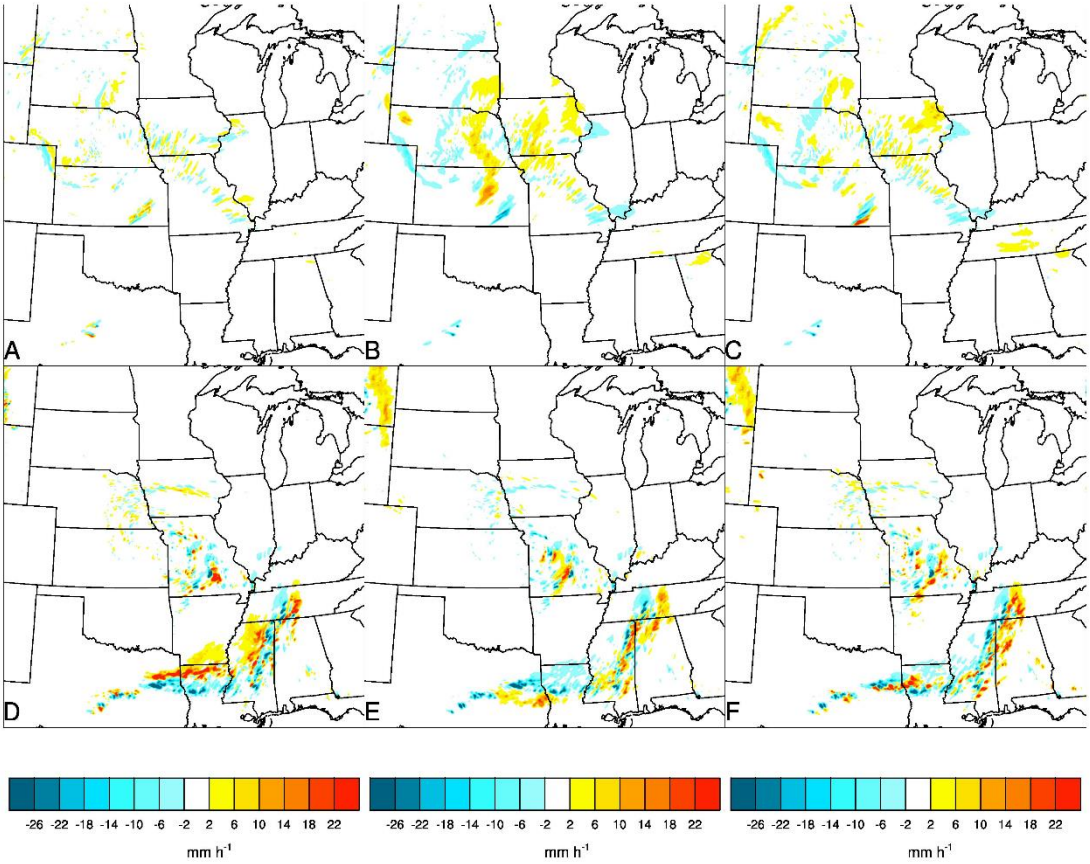

Figure 3.10: Forecast perturbations at the 24 h lead time (perturbed forecasts minus the control forecasts shown in Figs. 3.4d and 3.7e) for (a) RAND on the 10 May case, (b) LGPH on the 10 May case, (c) LG on the 10 May case, (d) RAND on the 20 May case, (e) LGPH on the 20 May case and (f) LG on the 20 May case.

In summary, the perturbation energy is again smaller than the error energy at early lead times but, unlike the 10 May case, is similar to the error energy after ~15h. Unlike the 10 May case, there is not a clear distinction between LGPH and

RAND/RECRS in terms of the magnitude of perturbation energy growth. Also unlike the 10 May case, RAND and RECRS become very similar on all scales after about 12 h. The distribution of perturbation energy across spatial scales was also generally more similar among the different perturbation methods on this case than on the 10 May case. The similar or greater energy of the RAND and RECRS perturbations compared to forecast errors and LGPH shows that small scale IC errors on this case contribute to the forecast uncertainty at least as much as the larger scale IC and physics errors. Differences between LGPH_RECRS and LGPH and between LG_RECRS and LG on this case suggest that LGPH_RECRS may be an effective method of combining the small scale IC perturbations with the SSEF design in certain situations.

3.3c Season average results

On average, the forecast error energy grows approximately linearly on the large scale with much less magnitude than on smaller scales (Fig. 3.11). On medium and small scales, the forecast error energy follows the diurnal cycle of convection, with maxima during the early forecast hours and during the following afternoon (Fig. 3.11b,c). The medium scale afternoon maximum of the second day persists into the evening while the small scale maximum decreases after ~23h (i.e., ~2300 UTC) (Fig. 3.11b,c). The total error energy temporal variability is dominated by the small and medium scales which have the largest magnitudes (Fig. 3.11c,d). All perturbation methods result in less total energy than the forecast errors (Fig. 3.11d). The under-estimation of forecast errors is most pronounced for medium and large scales and for the RAND and RECRS perturbations (Fig. 3.11a,b,c).

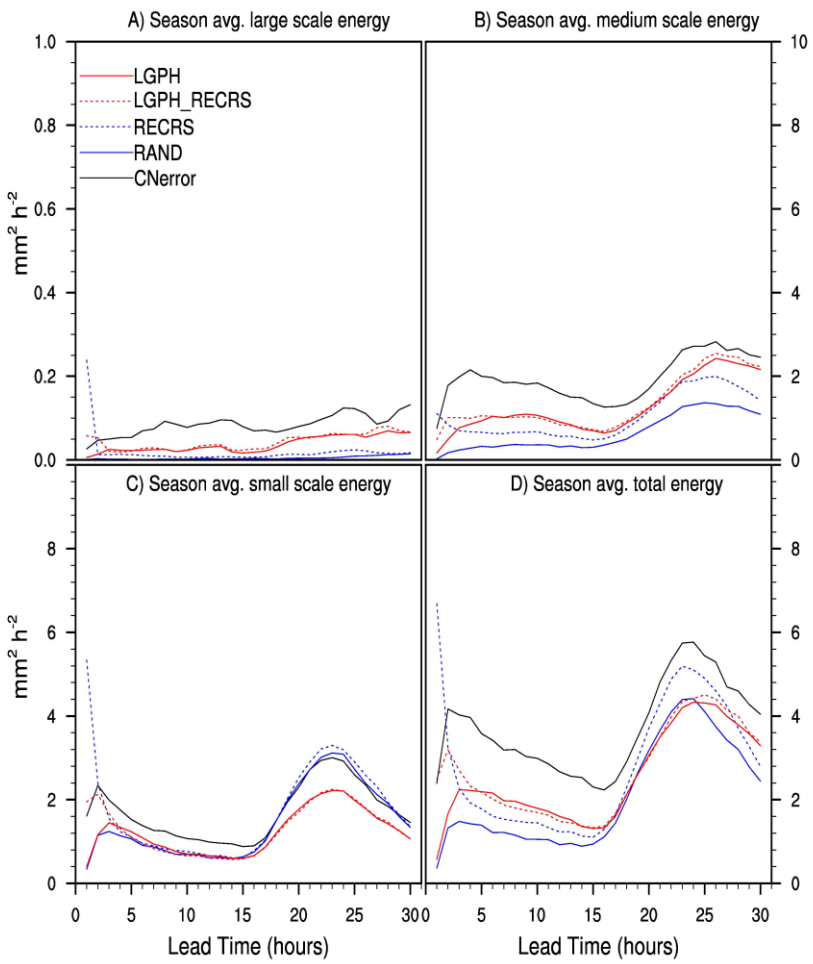

Figure 3.11: As in Fig. 3.5, except averaged over the entire experiment period.

Differences among the average perturbation energies in Fig. 3.11 are tested for statistical significance using one-sided permutation resampling (Hamill 1999) at the 95% confidence level (Table 3.1). On medium and large forecast scales, LGPH has significantly more perturbation energy than RAND and RECRS, except at early lead times due to the spurious precipitation of RECRS and except at 19-24h on the medium scale where the difference between LGPH and RECRS is not significant (Table 3.1). Only LGPH and LGPH_RECRS account for a substantial fraction of the error energy on large scales (Fig. 3.11a). On small scales LGPH is slightly, but significantly, greater

than RAND at 3-9h and is markedly less than RAND and RECRS at 16-30h (Fig. 3.11 and Table 3.1). The reduced LGPH perturbation energy compared to RAND and RECRS on small scales at 16-30h is a systematic result of the physics-related bias difference discussed for the 20 May case. Besides the first few hours, dominated by spurious precipitation for RECRS, significantly greater energy for RECRS than RAND is found at most lead times for large and medium scales and at several lead times (e.g., 20-23 and 25-27h) for small scales (Table 3.1). This difference is qualitatively most pronounced on the medium scales (Fig. 3.11b). On average, the medium scale differences between LGPH and RAND/RECRS are less pronounced than on the 10 May case. The RAND/RECRS medium scale perturbation energy is 50% or more of the LGPH perturbation energy on average at most lead times. This suggests systematic upscale growth of the small scale IC errors throughout the 30h forecast period. However, the differences between LGPH and LGPH_RECRS on average are generally small and/or not significant, again excluding early lead times dominated by spurious precipitation (Fig. 3.11 and Table 3.1).

The total average perturbation energy from all perturbation methods becomes similar after ~16h (Fig. 3.11d), 4 h later than the 11h time scale of insensitivity to the initial small scale perturbation method suggested by Hohenegger and Schär (2007a). The differences between RAND and RECRS perturbation energy, especially on the medium scales, throughout the forecast period suggests that the impact of the structure of small scale IC perturbations may persist longer into the forecast than expected.

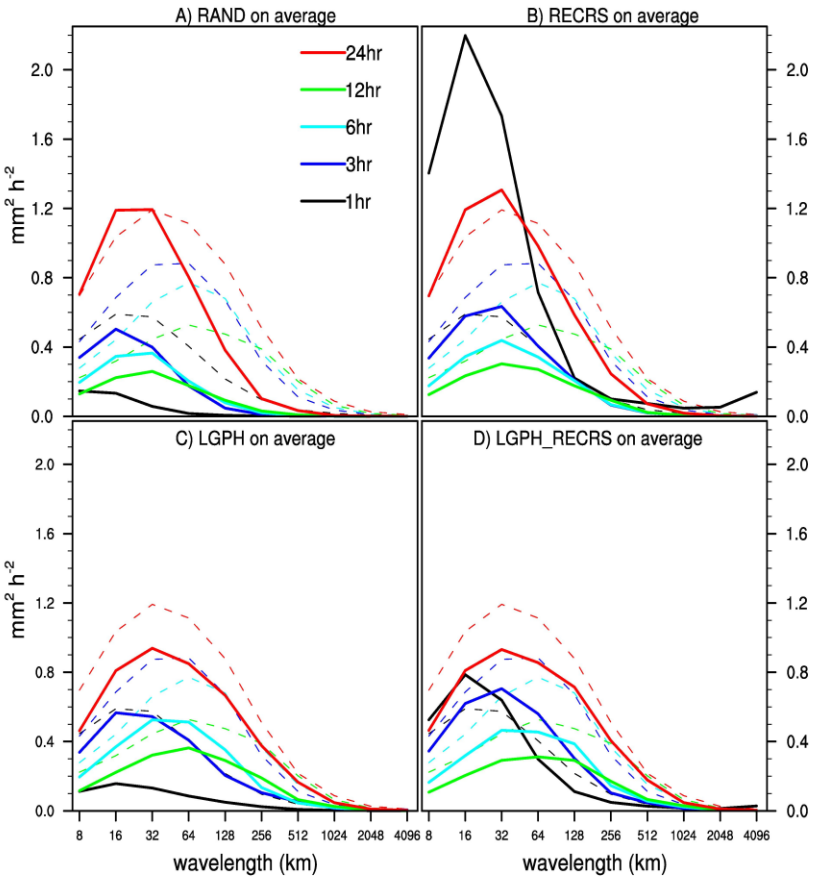

Figure 3.12: As in Fig. 3.8, except averaged over the entire experiment period and excluding panels (e) and (f) which were only generated for the two case studies.

Table 3.1: Statistical significance of the season-average differences in average perturbation (shown in Fig. 3.11) energy between each pair of perturbation methods. The pair of perturbations being compared is given at the top of columns 2-7. Each entry contains four results for large scale, medium scale, small scale and total, respectively. The Y indicates statistical significance at the 95% level using one-sided permutation resampling (Hamill 1999) and the N indicates no statistical significance.

Lead time	Rand vs recrs	Rand vs lgph_recrs	Rand vs lgph	Recrs vs lgph_recrs	Recrs vs lgph	Lgph vs lgph_recrs
1	Y,Y,Y,Y	Y,Y,Y,Y	Y,Y,Y,Y	Y,Y,Y,Y	Y,Y,Y,Y	Y,Y,Y,Y
2	Y,Y,Y,Y	Y,Y,Y,Y	Y,Y,N,Y	Y,Y,Y,N	N,Y,Y,Y	Y,Y,Y,Y
3	Y,Y,Y,Y	Y,Y,Y,Y	Y,Y,Y,Y	Y,Y,Y,Y	Y,N,Y,N	Y,Y,Y,Y
4	Y,Y,Y,Y	Y,Y,Y,Y	Y,Y,Y,Y	N,Y,Y,Y	Y,Y,Y,Y	Y,Y,N,N
5	Y,Y,N,Y	Y,Y,N,Y	Y,Y,Y,Y	Y,Y,N,Y	Y,Y,Y,Y	N,Y,Y,N
6	Y,Y,N,Y	Y,Y,N,Y	Y,Y,Y,Y	Y,Y,N,Y	Y,Y,Y,Y	N,N,Y,Y
7	Y,Y,N,Y	Y,Y,N,Y	Y,Y,Y,Y	Y,Y,Y,Y	Y,Y,N,Y	N,N,Y,Y
8	Y,Y,Y,Y	Y,Y,N,Y	Y,Y,Y,Y	Y,Y,Y,Y	Y,Y,N,Y	N,N,Y,Y
9	Y,Y,Y,Y	Y,Y,N,Y	Y,Y,Y,Y	Y,Y,Y,Y	Y,Y,N,Y	N.N,Y,Y
10	Y,Y,Y,Y	Y,Y,N,Y	Y,Y,N,Y	Y,Y,Y,Y	Y,Y,N,Y	N,N,Y,N
11	Y,Y,N,Y	Y,Y,N,Y	Y,Y,N,Y	Y,Y,N,Y	Y,Y,N,Y	N,N,Y,N
12	Y,Y,Y,Y	Y,Y,N,Y	Y,Y,N,Y	Y,Y,N,Y	Y,Y,N,Y	N,N,Y,Y
13	Y,Y,N,Y	Y,Y,N,Y	Y,Y,N,Y	Y,Y,Y,Y	Y,Y,N,Y	N,N,Y,N
14	Y,Y,N,Y	Y,Y,N,Y	Y,Y,N,Y	Y,Y,Y,Y	Y,Y,N,Y	Y,N,Y,N
15	Y,Y,N,Y	Y,Y,N,Y	Y,Y,N,Y	Y,Y,N,Y	Y,Y,N,Y	Y,N,N,N
16	Y,Y,N,Y	Y,Y,Y,Y	Y,Y,Y,Y	Y,Y,Y,N	Y,Y,Y,N	Y,N,N,N
17	Y,Y,N,Y	Y,Y,Y,Y	Y,Y,Y,Y	Y,Y,Y,N	Y,Y,Y,N	N,Y,N,N
18	Y,Y,N,Y	Y,Y,Y,N	Y,Y,Y,N	Y,Y,Y,N	Y,Y,Y,N	Y,N,Y,N
19	Y,Y,N,Y	Y,Y,Y,N	Y,Y,Y,N	Y,Y,Y,Y	Y,N,Y,Y	Y,N,Y,N
20	Y,Y,Y,Y	Y,Y,Y,N	Y,Y,Y,N	Y,N,Y,Y	Y,N,Y,Y	N,N,Y,N
21	Y,Y,Y,Y	Y,Y,Y,N	Y,Y,Y,N	Y,N,Y,Y	Y,N,Y,Y	N,N,N,N
22	Y,Y,Y,Y	Y,Y,Y,N	Y,Y,Y,Y	Y,N,Y,Y	Y,N,Y,Y	N,N,N,N
23	Y,Y,Y,Y	Y,Y,Y,N	Y,Y,Y,N	Y,N,Y,Y	Y,N,Y,Y	N,N,N,N
24	Y,Y,N,Y	Y,Y,Y,N	Y,Y,Y,N	Y,Y,Y,Y	Y,N,Y,Y	N,N,N,N
25	Y,Y,Y,Y	Y,Y,Y,Y	Y,Y,Y,N	Y,Y,Y,Y	Y,Y,Y,Y	N,Y,N,Y
26	Y,Y,Y,Y	Y,Y,Y,Y	Y,Y,Y,Y	Y,Y,Y,N	Y,Y,Y,N	N,N,N,N
27	N,Y,Y,Y	Y,Y,Y,Y	Y,Y,Y,Y	Y,Y,Y,N	Y,Y,Y,N	Y,N,N,N
28	N,Y,N,Y	Y,Y,Y,Y	Y,Y,Y,Y	Y,Y,Y,N	Y,Y,Y,N	N,N,N,N
29	N,Y,N,Y	Y,Y,Y,Y	Y,Y,Y,Y	Y,Y,Y,N	Y,Y,Y,N	N,N,N,N
30	N,Y,N,Y	Y,Y,Y,Y	Y,Y,Y,Y	Y,Y,Y,Y	Y,Y,Y,Y	N,N,N,N

The RAND and RECRS perturbations do not reflect the spectral evolution of error energy as well as LGPH (Fig. 3.12a,b,c). Except for the spurious precipitation at 1h, LGPH already approximately reflects the error energy maximum of ~32-128 km

wavelength by 6h (Fig. 3.12c). However, RAND and RECRS still do not even reflect the error energy maximum of 64 km wavelength at 12h (Fig. 3.12a,b). By 24h, all methods reflect the error energy maximum of 32 km wavelength (Fig. 3.12). At later lead times, LGPH generally has a broader spectrum, with more energy on the larger scales, than RAND and RECRS (Fig. 3.12a,b,c). Except for the very early lead times where RECRS and LGPH_RECRS are dominated by the spurious precipitation, there are not substantial differences in perturbation energy spectra between RAND and RECRS or between LGPH and LGPH_RECRS.

3.4 Summary and conclusions

The purpose of this study is to understand the multi-scale characteristics of the evolution of different sources of perturbations on convection-allowing precipitation forecasts on two case studies and for 34 forecasts on average, for the purpose of understanding the optimal SSEF design. In particular, three main goals are addressed. First, the impact of small scale IC perturbations (RAND and RECRS) is compared to the larger scale IC and physics perturbations (LGPH and LG) that are currently used in the CAPS Spring Experiment SSEF. Second, two methods of generating small scale IC perturbations (RAND and RECRS) are compared to each other. Third, LGPH is compared to a method of combining the small and large scale IC perturbations (LG_RECRS) and combining multi-scale IC and physics perturbations (LGPH_RECRS).

It is found that the relative impacts of the different types of perturbation are case-dependent. On the 10 May case the evolution of the precipitation systems in the

background forecast are driven primarily by a synoptic scale disturbance. After the first few hours, the 10 May forecasts containing large scale IC perturbations, with or without physics perturbations, have more perturbation energy than the small scale IC-only perturbations, RAND and RECRS, on medium and large scales while the small scale perturbation energy is similar for all methods. As a result, the perturbation energy spectra are generally broader for LG and LGPH than RAND and RECRS. On this case the RECRS method creates more forecast perturbation energy than RAND at most lead times for the medium scales and for many lead times after ~20h for the small scales. LGPH_RECRS and LG_RECRS do not increase the perturbation energy relative to LGPH and LG, respectively, on this case. In contrast, the 20 May case has ongoing convection in the background forecast at the initial time that grows upscale into an MCS. The 20 May forecasts are generally less sensitive to the scale of IC perturbations, with LG and LGPH not showing a clear increase of perturbation energy, relative to RAND and RECRS, on any scale. The perturbation energy spectra are also less sensitive to the perturbation method on 20 May than on 10 May. There is less forecast energy for LGPH than for RAND and RECRS on small scales at ~20-27h due to the physics scheme differences. On 20 May, RECRS shows increased perturbation energy, relative to RAND, for only the first ~12-15h on small and medium scales. Unlike the 10 May case, the 20 May case shows greater perturbation energy at ~20-26h for LGPH_RECRS and LG_RECRS than LGPH and LG, respectively.

One of the main differences in perturbation evolution between the two cases is the greater sensitivity to the small scale IC perturbations, relative to the larger scale IC and physics perturbations, on the 20 May case. This is consistent with past case studies

suggesting that lower predictability generally results from the release of deep moist convective instability (e.g., Hohenegger et al. 2006). However, Zhang et al. (2006) found less sensitivity of the mesoscales to small scale random IC perturbations for a warm season heavy precipitation event than a large scale winter cyclone event. This contrasts with the results in the present study. Reasons for this difference may include the direct consideration of precipitation forecasts, instead of wind and temperature differences as in Zhang et al. (2006), as well as differences in the forcing mechanisms of the precipitation systems. For example, our 20 May case is characterized by upscale growth of convection due to internal storm dynamics rather than the large scale moisture transport interacting with topography in Zhang et al. (2006).

The perturbations are evaluated over a large number of forecasts to better understand their systematic behavior, independent of the many factors of individual cases that can affect the predictability. Averaged over 34 forecasts, there is a diurnal cycle of forecast error and perturbation energy on the small and medium scales. Compared to RAND and RECRS, the forecast sensitivity is dominated by LGPH and LGPH_RECRS perturbations on large and medium scales. However, on medium scales RAND and RECRS alone can generate at least half as much forecast perturbation energy as LGPH throughout the forecast period. This sensitivity of the medium forecast scales to small scale IC perturbations is more similar to the 20 May case than the 10 May case. This similarity is consistent with the expectation that during the late spring and early summer season convective episodes are often dominated by localized and/or diurnal forcings, such as those on the 20 May case, rather than the dominant large scale forcing like the 10 May case (Stensrud and Fritsch 1993). Perturbation energy for

LGPH and LGPH_RECRS is systematically reduced on small scales during the diurnal convective maximum due to the different biases of the physics schemes. The most prominent difference between RAND and RECRS is an increase of medium scale perturbation energy at all times for RECRS. RECRS also shows greatly increased energy at 1-2h due to spurious precipitation. Refinement of the RECRS method would therefore be necessary before inclusion in an ensemble forecast system. On average, LGPH_RECRS does not create significantly more perturbation energy than LGPH on any scale after the first few hours which are dominated by the spurious precipitation.

The dominant impact of large scale IC and physics perturbations suggests that the current CAPS ensemble configuration, sampling only large scale IC and physics errors, already samples the primary forecast sensitivity. The comparable, although lesser, impact of only small scale IC perturbations on medium scales also implies a process of upscale growth of the initially small errors that can substantially contribute to the medium scale forecast sensitivity. However, the method of generating multi-scale IC perturbations represented by LGPH_RECRS does not show a systematic increase in medium scale perturbation energy, relative to LGPH. The three most likely reasons for this lack of impact are that (1) better methods of combining multiple scales of IC perturbation need to be developed, (2) there is only an advantage of including small scales in the IC perturbations under certain conditions such as situations of rapid upscale error propagation (e.g., the May 20 Case), or (3) the downscale energy cascade of the large scale IC perturbations implicitly accounts for small scale errors that are not explicitly sampled.

More work is needed to understand how to realistically and efficiently sample, and optimally combine, all scales of uncertainty, from synoptic to convective, into IC/LBC perturbations, along with physics perturbations, for SSEFs. The methods of defining the small scale IC perturbations in this chapter are not flow-dependent, may not reflect the actual analysis errors, and can result in unbalanced initial fields that are detrimental to short term forecasts. For example, the RAND perturbations exhibit no initial spatial structure and result in less growth than the RECRS perturbations. The RECRS perturbations are defined to have a fixed, uniform spatial structure and amplitude but create spurious precipitation at early lead times. The differences between RAND and RECRS, especially on the medium forecast scales, show the importance of the spatial structure of small scale IC perturbations. Flow-dependent methods should be developed to better sample the small scale error structure in the ICs. The following chapters investigate the use of ensemble based data assimilation to provide flow-dependent multi-scale IC perturbations for SSEFs. In addition to IC/LBC perturbation methods, different physics perturbations may also yield different results. Investigation of physics perturbation methods such as using different physics schemes and different parameters within a fixed scheme is left for future study. While this study focuses primarily on the spatial scales of forecast perturbation, the questions of which variables should be perturbed and what the covariance should be among the perturbed variables for SSEF design remains an open question. Ensemble-based data assimilation may also be useful to address such questions.

Chapter 4: OSSE study of multi-scale initial condition perturbation methods. Part 1: Case study of MCS upscale growth case

4.1 Introduction

This part of the study has three main goals aimed at better understanding the optimal SSEF IC perturbation design. The first goal is to better understand the advantages of flow-dependent multi-scale IC perturbations for ensemble forecasts of mid-latitude convection, compared to IC perturbations downscaled from a coarser resolution ensemble. The different IC perturbation methods include not only differences in resolution, but also differences on commonly resolved scales as a result of being generated on different model grids with different DA methods. The second goal of this study is therefore to understand the impacts of the mesoscale component (i.e., on commonly resolved scales) of the differences between the IC perturbation methods. The third goal is to better understand the impacts of the small scale IC perturbations which are only resolved in the multi-scale method.

Convection-permitting forecasts provide information that is useful for users interested in applications ranging from ~1h predictions of individual storms and severe weather events (e.g., Stensrud et al. 2009; Yussouf et al. 2013) to mesoscale quantitative precipitation forecasting (e.g., Clark et al. 2009, 2012; Duc et al. 2013). The impacts of the IC perturbations are therefore evaluated in terms of hourly accumulated precipitation forecasts out to 9 h in mesoscale (i.e., 48 km radius) neighborhoods as well as short term (2 h) reflectivity forecasts in storm-scale neighborhoods ranging from 0-48 km.

While model and physics diversity are also an important part of the ensemble design (e.g., Clark et al. 2008), this study focuses on the IC perturbation design. Therefore, perfect model Observation System Simulation Experiments (OSSEs) are used to isolate the IC error from the model and physics errors. In this chapter, a case study of cellular convection growing upscale into a long-lived mesoscale convective system (MCS) is used to qualitatively understand the impacts of the IC perturbation design. In the following chapter, systematic evaluation of eleven diverse cases of mid-latitude convection is used to draw more robust conclusions and relate them to the optimal design of SSEF IC perturbations. Impacts of model error are also considered in Chapter 7. In this chapter, Section 4.2 describes the IC perturbation methods and verification methods. Results are presented in Section 4.3 while Section 4.4 includes a summary and conclusions.

4.2 Methods

4.2a OSSE design

In an OSSE, a model simulation referred to as the nature run represents the "true" atmosphere, the state and dynamics of which are perfectly known. In this study, the nature run is initialized from the NCEP GFS analysis at 00 UTC 19 May and run at 4 km grid spacing over the outer domain in Fig. 2.1. Observations of wind, temperature, water vapor, sea level pressure, radar radial velocity and radar reflectivity are then simulated by sampling the nature run at observation locations representative of the actual observation networks (e.g., Fig. 2.1), with representative observation error characteristics. The simulated observations are then assimilated into the experiment

forecasts in order to try to recover the "true" state of the Nature run, using the GSI-based multi-scale DA system.

An advantage of the OSSE framework is that the truth is perfectly known at the model grid points. For this study of IC perturbations, the OSSE framework has the additional advantage of eliminating model and physics uncertainty as a source of forecast error by using identical model configurations for the nature run and experiment forecasts (i.e., a "perfect model" OSSE). The outer domain analyses do contain model error arising from the coarser resolution and convection parameterization. However, such errors only enter the convection-permitting forecasts through the ICs provided by the inner domain DA and the LBCs from the outer domain.

The actual evolution of the 20 May case study, including upscale growth of initially cellular convection into a long-lived MCS in central OK, has been described in Chapter 2. The nature run for this case also shows similar upscale growth of convection into a long-lived MCS, as seen in the observation contours in Fig. 4.4. This case is chosen for an initial investigation into multi-scale IC perturbation methods because of the multiple scales of motion influencing such upscale growing MCSs (e.g., Perkey and Maddox 1985), the sensitivity of this case to IC errors on multiple scales in non-OSSE experiments (Chapter 3), and the similarity between the nature run and actual evolution for this case. Experiment forecasts are initialized at both 0000 UTC, about half-way through the upscale growth of the MCS, and 2100 UTC, very early in the process of upscale growth. The comparison of experiments initialized at different phases of the upscale growth provides an estimate of the case dependence of the results.

4.2b IC perturbation methods

In all experiments, the ensemble forecasts have the same mean analysis, provided by the ensemble mean analysis of the multi-scale GSI-based DA system. The only difference among the experiments is the IC perturbations added to the ensemble mean to generate the initial ensemble. The first goal of this part of the study is to understand the impact of the multi-scale IC perturbations generated on the inner domain (hereinafter, MULTI) in comparison to larger scale IC perturbations downscaled from the outer domain (hereinafter, LARGE). The MULTI IC perturbations are obtained by directly using the inner domain multi-scale analyses to initialize the ensemble forecasts. The LARGE IC perturbations are obtained by adding the difference between each outer domain ensemble member and the outer domain ensemble mean (both interpolated to the inner domain using the WRF ndown utility) to the inner domain ensemble mean. The second and third goals of this part of the study are to understand the impacts of (a) the differences between MULTI and LARGE on commonly resolved scales and (b) the smaller scale IC perturbations in MULTI. A third ensemble, MULTI48, is therefore constructed by filtering[1] wavelengths less than 48 km from each MULTI perturbation before adding it back to the inner domain ensemble mean. Since MULTI48 is the same as MULTI except for the absence of perturbations on scales not resolved by LARGE, comparison of MULTI48 with LARGE allows (a) to be investigated while comparison of MULTI48 with MULTI allows (b) to be investigated. For simplicity, wavelengths less than 48 km are therefore referred to as "small scale" IC perturbations in this study while the larger scales are referred to as "mesoscale" IC perturbations in this study.

[1] The filtering consists of truncation of wavelengths below 48km in the two-dimensional Discrete Cosine Transform (Denis et al. 2002) of the IC perturbation field.

Although there is not such a sharp cut-off in the scales resolved by the LARGE IC perturbations, the difference in perturbation energy between MULTI and LARGE is particularly pronounced at wavelengths smaller than ~50 km (Fig. 4.1a,f), motivating the choice of 48 km to separate small scales and mesoscales in the IC perturbations.

4.2.c Verification methods

SSEFs have proven useful for users interested in convective precipitation forecast applications on space and time scales ranging from very short term warn-on-forecast applications (e.g., Stensrud et al. 2009; Yussouf et al. 2013) to mesoscale quantitative precipitation forecasting (e.g., Clark et al. 2009, 2012; Duc et al. 2013). Forecasts on such different scales may show different sensitivities to the multi-scale IC perturbation methods. In order to provide a robust understanding of the impacts of IC perturbation methods, the convection forecasts are here evaluated in terms of both instantaneous reflectivity during the first two forecast hours and mesoscale hourly accumulated precipitation out to nine hours. Reflectivity results are shown using model level 12 (~750mb). Reflectivity at model level 5 (~900mb) was also evaluated and showed very similar results (not shown).

The forecasts are objectively verified using the Brier Skill Score (BSS; Brier 1950; Murphy 1973; Wilks 2006) of Neighborhood Ensemble Probability (NEP; Theis et al. 2005; Schwartz et al. 2010). The NEP is the percentage of grid points from all ensemble member forecasts within a search radius that exceed the threshold being forecast. The use of the NEP reduces the sensitivity to errors on scales smaller than the search radius (Roberts and Lean 2008). A radius of 48 km is chosen for the mesoscale

hourly accumulated precipitation forecasts in order to eliminate the impact of smaller scale and less predictable details (Johnson and Wang 2012). The reflectivity forecasts are evaluated across a range of different spatial scales (i.e., radii less than 48 km) and verification thresholds (Stratman et al. 2013). The BSS provides a simple way to verify the ensemble probabilistic forecasts that is sensitive to both the reliability and resolution of the forecasts (Murphy 1973). In addition to the objective verification, subjective verification is also conducted to qualitatively understand the physical processes behind the objective skill metrics for this case study.

Verification of mesoscale precipitation and the non-precipitation variables is conducted over the verification domain plotted in Fig 4.4. The reflectivity verification is conducted in a smaller verification domain focused on the MCS of interest (e.g., Fig. 4.7).

4.3 Results

4.3a Non-precipitation variables

Since the non-precipitation variables are the directly perturbed IC variables, results for wind, temperature and water vapor are first considered. One-dimensional de-trended Fourier spectra for these variables are calculated along east-west grid lines, and averaged over all possible such grid lines (Skamarock 2004). The spectra of the ensemble mean error and the ensemble member perturbations, averaged over the 40 ensemble members, are compared for the u-component of wind at model level 5 (~900 mb) in Fig. 4.1. Results for this variable are similar to wind at model level 12 (~750 mb) as well as temperature and water vapor (not shown). The ensemble mean error

spectra are very similar for the MULTI and LARGE ensembles at the lead times shown in Fig. 4.1, with the green and blue dashed lines nearly on top of each other. At the initial time, the LARGE ensemble perturbations are markedly under-dispersive, compared to the ensemble mean error, at scales less than ~50 km for both the 0000 and 2100 UTC cases (Fig. 4.1a,f). The lack of small scale spread is a result of the coarser resolution of the outer domain ensemble used to generate the LARGE perturbations.

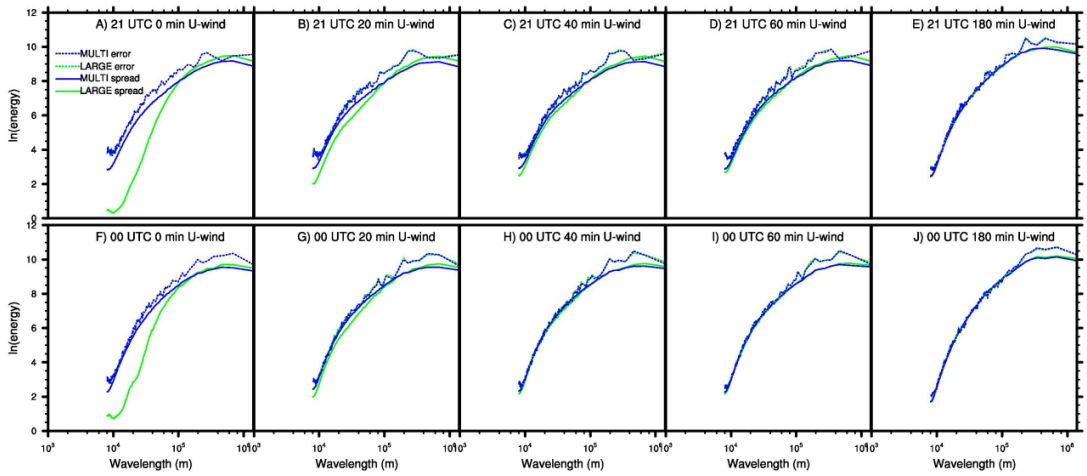

Figure 4.1: Fourier spectra decomposition of ensemble perturbations (ensemble member minus ensemble mean, averaged over all members; solid) and ensemble mean error (dashed) for the u component of wind at model level 5 (~900 mb) for the 2100 UTC case at (a) the analysis time, (b) 20-minute forecast time, (c) 40-minute forecast time, (d) 60-minute forecast time, (e) 180-minute forecast time and (f)-(j) as in (a)-(e) except for the 0000 UTC case.

Although the small scales are initially very under-dispersive for LARGE, downscale energy propagation results in rapid perturbation growth on such scales, consistent with the results of Durran and Gingrich (2014). The small scale energy for LARGE catches up to that for MULTI within about an hour, depending on the case (Fig. 4.1). This confirms that explicitly including small scale IC perturbations has little

impact on the ensemble spread of the directly perturbed variables on such scales for lead times beyond ~1 h. However, it is not clear what impacts the small scale IC perturbations during the first hour have on the convective precipitation forecasts both during and after the first hour and on larger scales.

The ensemble spread (i.e., standard deviation) of wind, temperature and moisture at model levels 5 and 12 (~900 and 750 mb, respectively) are also evaluated and compared to the ensemble mean root mean square error (RMSE) in Fig. 4.2 for the 0000 UTC case. At early lead times, MULTI has less spread than LARGE for most variables (Fig. 4.2), consistent with the largest scales of Fig. 4.1 which dominate the total spread. An exception to this trend at early lead times is level 12 moisture (Fig. 4.2f), discussed further in Section 4.3b(2)(ii). MULTI has more spread than LARGE for most variables by the end of the 9h forecast period, although all of the ensembles are still under-dispersive at this time for most variables. Since MULTI generally has less spread than LARGE initially, this indicates greater perturbation growth during the forecast period in the MULTI ensemble than the LARGE ensemble for this case.

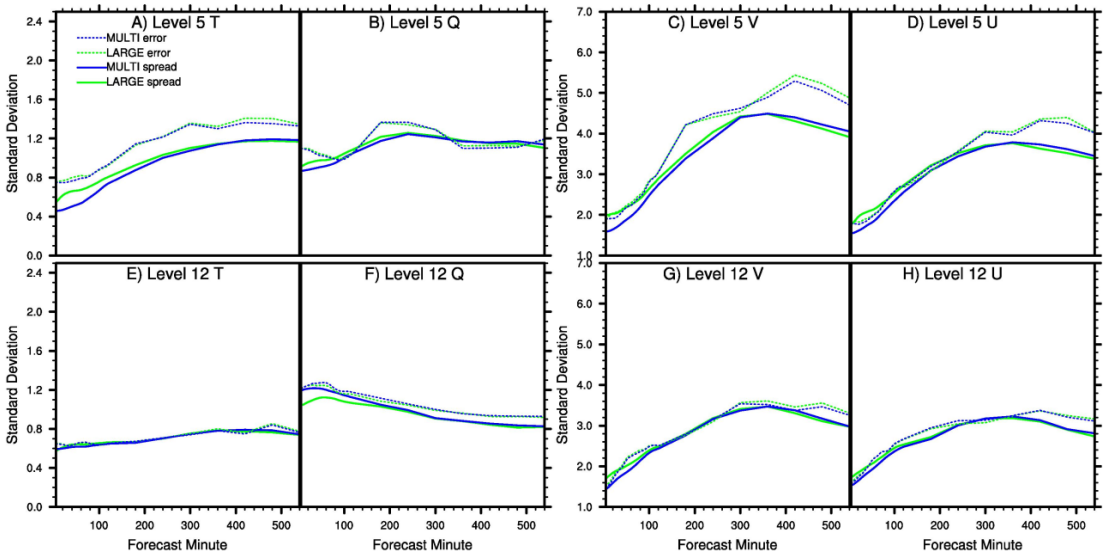

Figure 4.2: Ensemble spread (i.e., standard deviation) and ensemble mean RMSE as a function of forecast lead time for the 0000 UTC case at model level 5 (~900 mb) for (a) temperature (K), (b) water vapor mixing ratio (g kg^{-1}), (c) v wind component (m s^{-1}), (d) u wind component (m s^{-1}), and (e)-(h) as in (a)-(d) except at model level 12 (~750mb).

4.3b Convective precipitation forecasts

A wavelength of 48 km is used to distinguish the impacts of "small scale" and "mesoscale" differences in the IC perturbations. This approximately corresponds to the scale below which the LARGE IC perturbations have very little energy compared to the MULTI IC perturbations in Fig. 4.1. The impacts on the convective precipitation forecasts on different spatial scales are also distinguished using different forecast variables (hourly accumulated precipitation and instantaneous reflectivity) and different neighborhood radii (48 km for precipitation and 0-48 km for reflectivity).

The following sub-sections evaluate the differences between MULTI and LARGE, MULTI48 and LARGE, and MULTI and MULTI48, consistent with the three goals of this study. The results in this section emphasize the experiments with forecasts

initialized at 0000 UTC in order to minimize redundancy. Results from the 2100 UTC case are also noted where additional information is added by the 2100 UTC case.

4.3b(1) OVERALL IMPACT OF IC PERTURBATION METHOD

4.3b(1)(i) Mesoscale hourly accumulated precipitation

The first goal of this part of the study is to understand the differences in forecast performance between IC perturbations generated on the inner domain and those downscaled from the outer domain (i.e., MULTI vs. LARGE). The relative performance of the MULTI and LARGE ensembles for hourly accumulated precipitation depends on the forecast initialization time (0000 or 2100 UTC; Fig. 4.3a,b,c or Fig. 4.3d,e,f, respectively). The differences are often smaller than the sampling uncertainty of the verification statistic[2] (Fig. 4.3), emphasizing the need for a more systematic comparison over many diverse cases in Chapter 5. At most lead times MULTI is more skillful than LARGE for the 0000 UTC case (Fig. 4.3a,b,c) while LARGE is more skillful than MULTI for the 2100 UTC case (Fig. 4.3d,e,f).

[2] 90% confidence intervals of the LARGE BSS in Fig. 4.3 are calculated using 5000 bootstrap resamples with replacement of the 40 ensemble members for each BSS value. The purpose is to estimate the uncertainty in the verification statistic resulting from sampling errors due to the finite ensemble size.

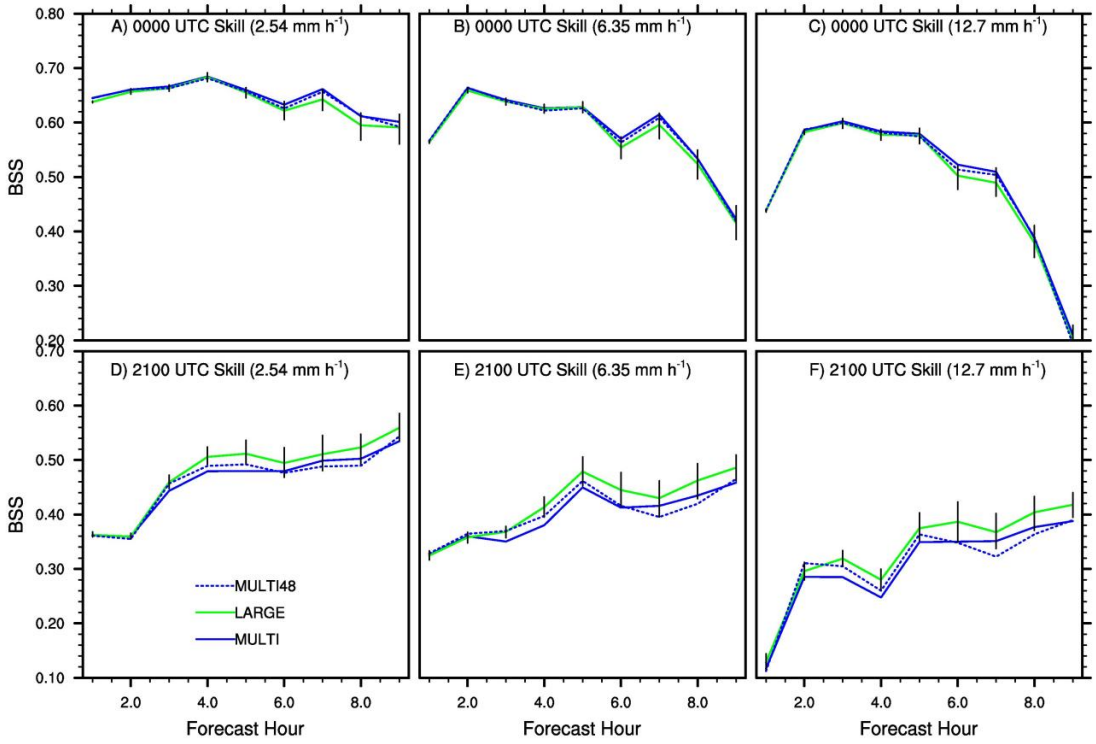

Figure 4.3: Brier Skill Score (BSS) of the Neighborhood Ensemble Probability (NEP) forecasts initialized at (a)-(c) 0000 UTC and (d)-(f) 2100 UTC for hourly accumulated precipitation thresholds of (a),(d) 2.54 mm h^{-1}, (b),(e) 6.35 mm h^{-1} and (c),(f) 12.7 mm h^{-1}. 90% confidence intervals for the LARGE ensemble are calculated as described in footnote 2.

Despite the case dependence of the objective metrics, subjective evaluation of the two cases provides physical understanding of the causes of such differences in forecast skill. For both cases the LARGE forecast is subjectively reasonable (e.g., Fig. 4.4 for the 0000 UTC case). However, there are subtle errors such as a slight westward displacement of the axis of maximum NEP, relative to the observed MCS, at the southern end of the MCS at later lead times (Fig. 4.4). The LARGE ensemble also predicts some spurious cells in the MCS cold pool resulting in non-zero probability northwest of the observed MCS during the first ~2 h (Fig. 4.4a,b). The figures

discussed below are plotted as differences from the LARGE (or MULTI48) NEP in order to emphasize subtle forecast differences.

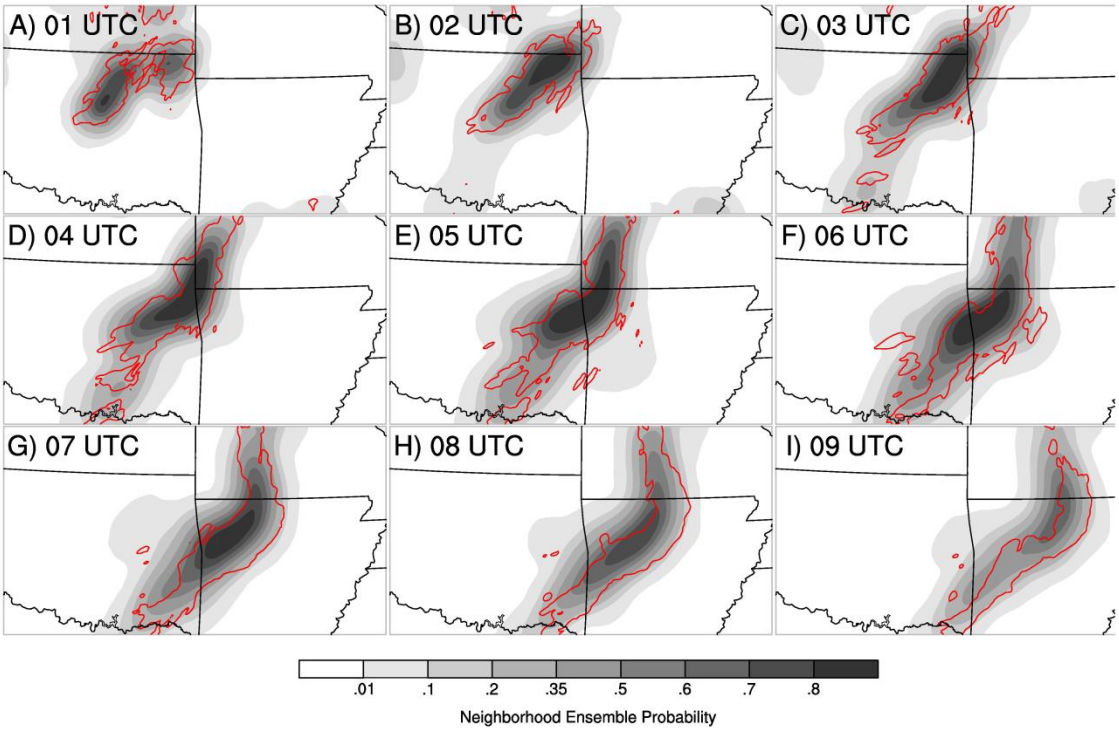

Figure 4.4: Forecast NEP (shaded) and observation contour (red line) for the LARGE ensemble forecast of hourly accumulated precipitation exceeding 6.35 mm h^{-1}, initialized at 0000 UTC 20 May 2010.

The generally greater skill for MULTI than LARGE for the 0000 UTC case (Fig. 4.3a,b,c) is consistent with subjective evaluation (Fig. 4.5a-i). Initially, MULTI shows reduced probability, compared to LARGE, in the cold pool region northwest of the MCS and increased probability farther east. The MULTI advantage of reducing the NEP in the cold pool region persists for ~3-4 h (Fig. 4.5a-d). Starting at ~0500 UTC, MULTI has higher probability along the eastern edge of the MCS and another area of reduced probability west of the southern end of the MCS (Fig. 4.5e-i). The westward

displacement of the maximum NEP at later lead times for LARGE at the southern end of the MCS is therefore partly corrected in the MULTI forecast.

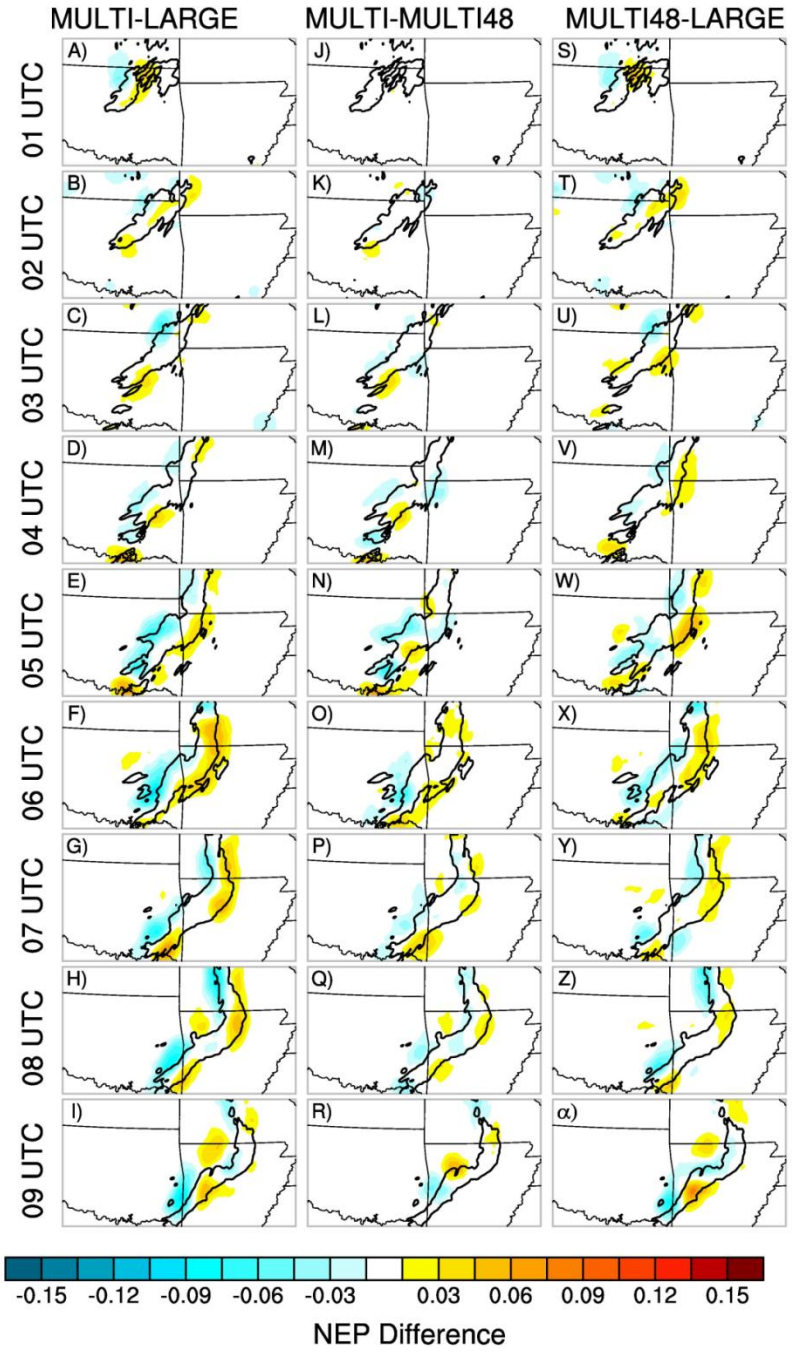

Figure 4.5: Difference in NEP between (left column) MULTI and LARGE, (center column) MULTI and MULTI48 and (center column) MULTI48 and LARGE, for hourly accumulated precipitation forecasts initialized at 0000 UTC 20 May 2010, for the 6.35 mm h^{-1} threshold.

Although the relative skill of the ensembles between the 0000 and 2100 UTC cases is quite different (Fig. 4.3), subjectively there are also common features between the two cases. For example, there is also a reduction of probability for MULTI, compared to LARGE, west and north of the incipient MCS during the first couple of hours of the 2100 UTC forecast (not shown). However, the skill is dominated by spurious storms to the south of the observed MCS in several members for the 2100 UTC case. These storms result in a new MCS to the southeast of the initial MCS that dominates and reduces the intensity of the initial MCS in those members (not shown).

4.3b(1)(ii) Storm scale reflectivity

The impacts of the IC perturbation methods on convective precipitation forecasts are also evaluated in terms of the reflectivity forecasts which contain smaller scale detail than hourly accumulated precipitation forecasts. The reflectivity forecasts, verified on smaller scales than the accumulated precipitation forecasts, are not considered beyond the 2 h forecast range because of the intrinsic lack of predictability at longer lead times for storm scale features (Cintineo and Stensrud 2013). Figure 4.6 shows the difference in BSS between the MULTI and LARGE reflectivity forecasts at 5 minute intervals for the 0000 UTC case. MULTI is generally the more skillful ensemble where the BSS difference exceeds a magnitude of 0.01 (Color shading in Fig. 4.6). The MULTI advantage is most pronounced at ~30-40 minutes, while slight LARGE advantages begin to appear during the last ~20 minutes.

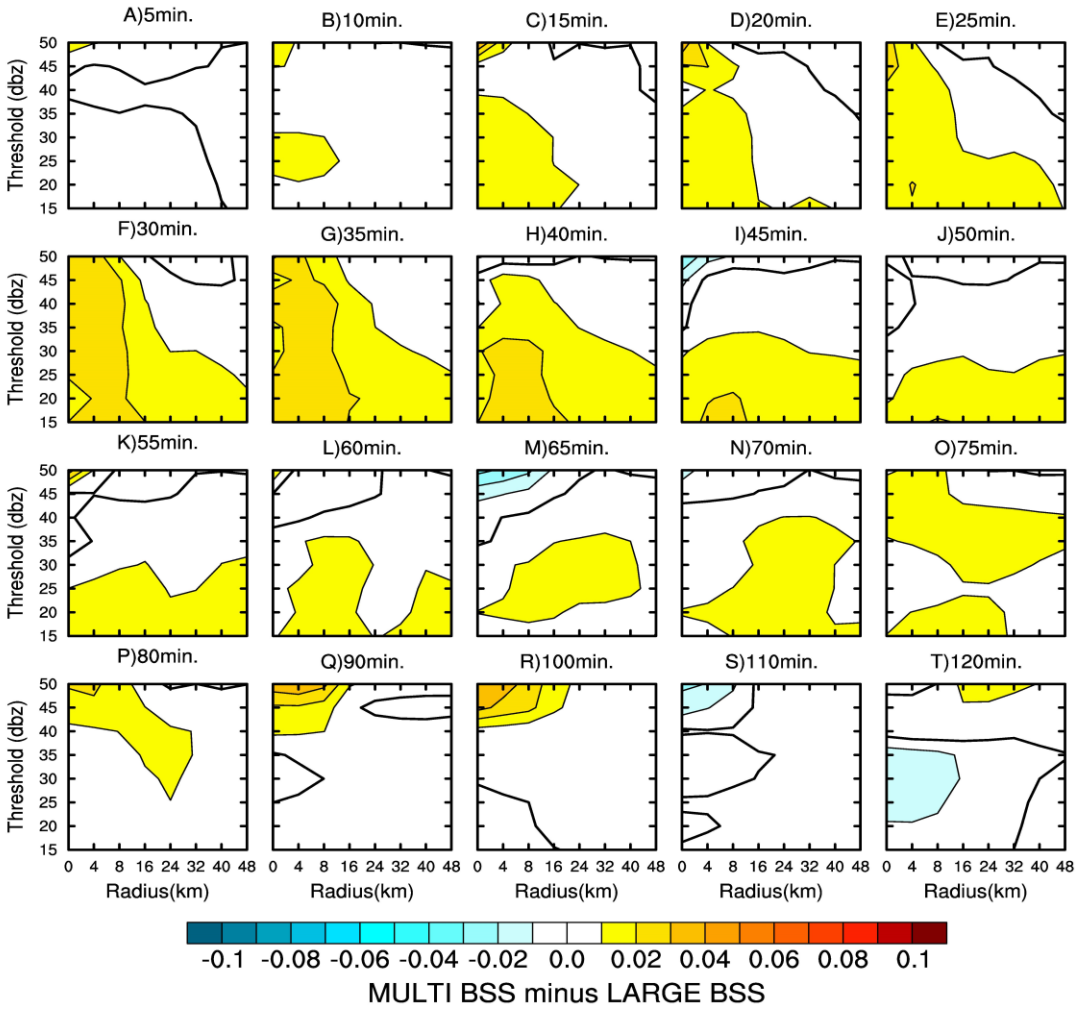

Figure 4.6: Difference in BSS for reflectivity at model level 12 between the MULTI and LARGE ensemble at five minute intervals during the first 80 minutes and at ten minute intervals between 80 and 120 minutes. The vertical axis on each panel is the reflectivity threshold (dBZ) and the horizontal axis is the neighborhood radius (km).

Subjectively, there are two competing factors that qualitatively explain the differences between MULTI and LARGE forecast skill, as illustrated with the representative 30 dBZ threshold forecasts for the 0000 UTC case (Fig. 4.7a-h). First, MULTI provides a sharper forecast of the MCS with greater resolution than LARGE, since MULTI has lower probability outside of the observed MCS and higher probability within the observed MCS. In particular, there is reduced MULTI probability outside

and to the west of the observed MCS and a corresponding increase in MULTI probability inside the northern end of the observed MCS contour (Fig. 4.7a-g). This difference is most pronounced at ~35-45 minutes, consistent with the greatest MULTI skill advantage in Fig. 4.6. Second, MULTI enhances the forecast probability outside of the observation contour at the southern and eastern edges of the observed MCS, negatively impacting the forecast skill. This difference becomes more pronounced at the later lead times, explaining the decreasing MULTI skill advantage and eventual slight LARGE skill advantage in Fig. 4.6. The causes of these qualitative differences are discussed further in the following sub-sections.

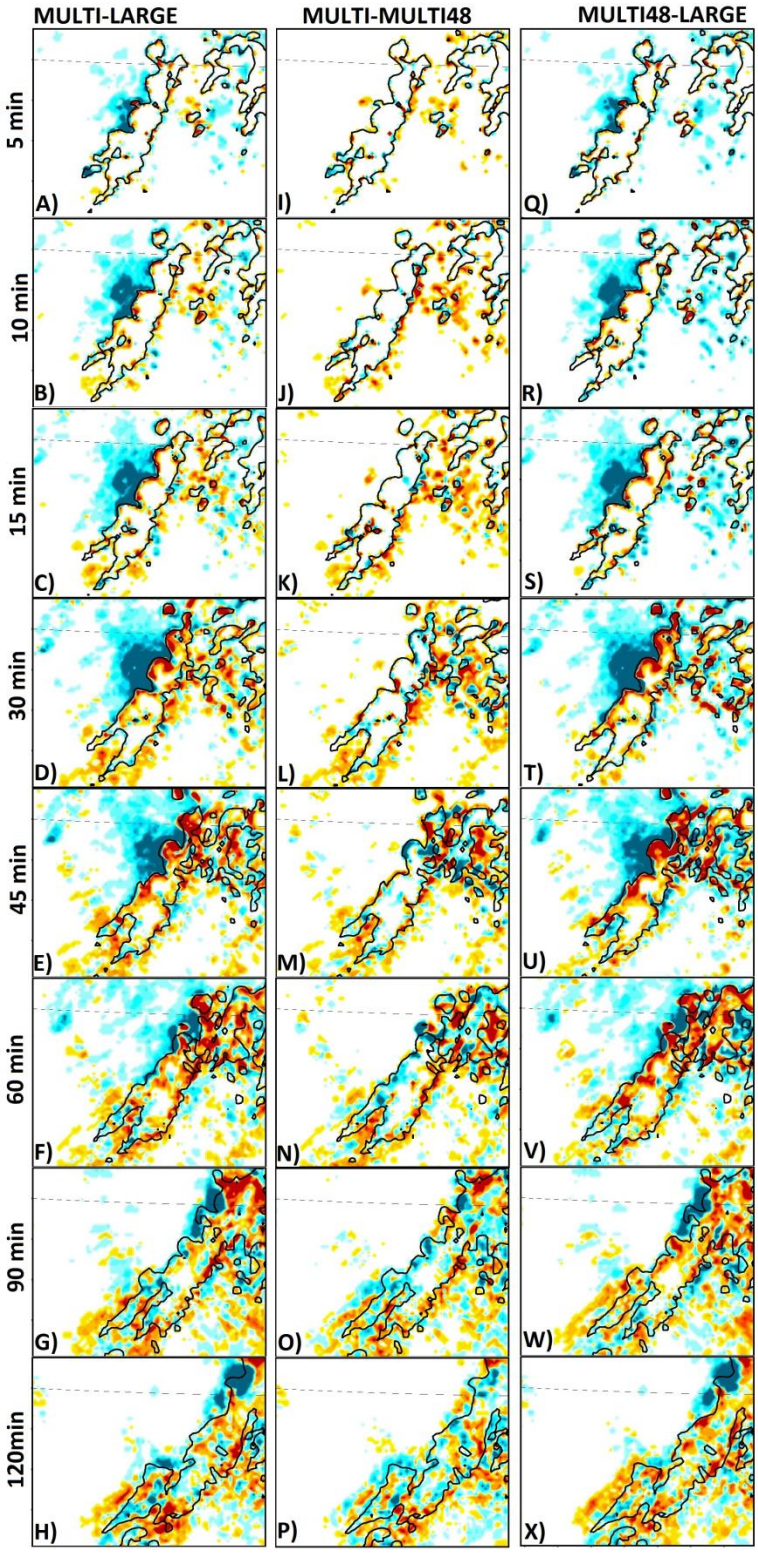

Figure 4.7: As in Fig. 4.5, except for forecasts of reflectivity exceeding 35 dBZ in the verification domain focused on the MCS of interest for forecasts initialized at 0000 UTC.

4.3b(2) IMPACT OF MESOSCALE COMPONENT OF IC PERTURBATION METHOD

4.3b(2)(i) Mesoscale hourly accumulated precipitation

The second goal of this part of the study is to understand the impacts of the different methods of generating the mesoscale component of the IC perturbations (i.e., MULTI48 vs. LARGE). For the 0000 UTC case, the mesoscale precipitation forecast differences between MULTI and LARGE are primarily determined by the differences between MULTI48 and LARGE (i.e., subjective similarity between left and right columns of Fig. 4.5 and between blue lines in Fig. 4.3a-c). In particular, the reduction of spurious precipitation behind the MCS for MULTI, and the subsequent differences from LARGE in the northern and eastern parts of the MCS at later times are also present in the differences between MULTI48 and LARGE (Fig. 4.5s-α). The differences from LARGE in the southern part of the MCS are also more strongly impacted by the mesoscale IC perturbation differences than the small scale IC perturbations since the left and right columns of Fig. 4.5 are more similar in this area than the left and center columns at most lead times. For the 2100 UTC case, the precipitation forecast differences between MULTI and LARGE are also determined mainly by the differences between MULTI48 and LARGE (e.g., Fig. 4.3d-f). Since the mesoscale differences in the IC perturbation methods have similar qualitative impacts on both precipitation and reflectivity forecasts, the qualitative explanation of these forecast differences are explained in the following sub-section.

4.3b(2)(ii) Storm scale reflectivity

For the reflectivity forecasts, the differences in skill are also dominated by the differences between MULTI48 and LARGE, with the exception of the 0-4 km radius neighborhoods during the first hour (Fig. 4.8). The difference from LARGE is even more pronounced for MULTI48 than MULTI at many times and extends to higher thresholds during the first ~75 minutes (Fig. 4.8). During the first ~75 minutes, the MULTI48 advantages are subjectively consistent with a similar (to MULTI) reduction in spurious probability in the cold pool and correspondingly sharper MCS forecast (Fig. 4.7q-x). The MULTI48 advantages during the first ~75 minutes are generally more pronounced than for MULTI because the enhanced probabilities outside of the southern and eastern edges of the MCS are less pronounced for MULTI48 than MULTI (Fig. 4.7; left and right columns). Reduced spurious probability for MULTI48 in the cold pool area at early lead times, compared to LARGE, is also present in the 2100 UTC case (not shown).

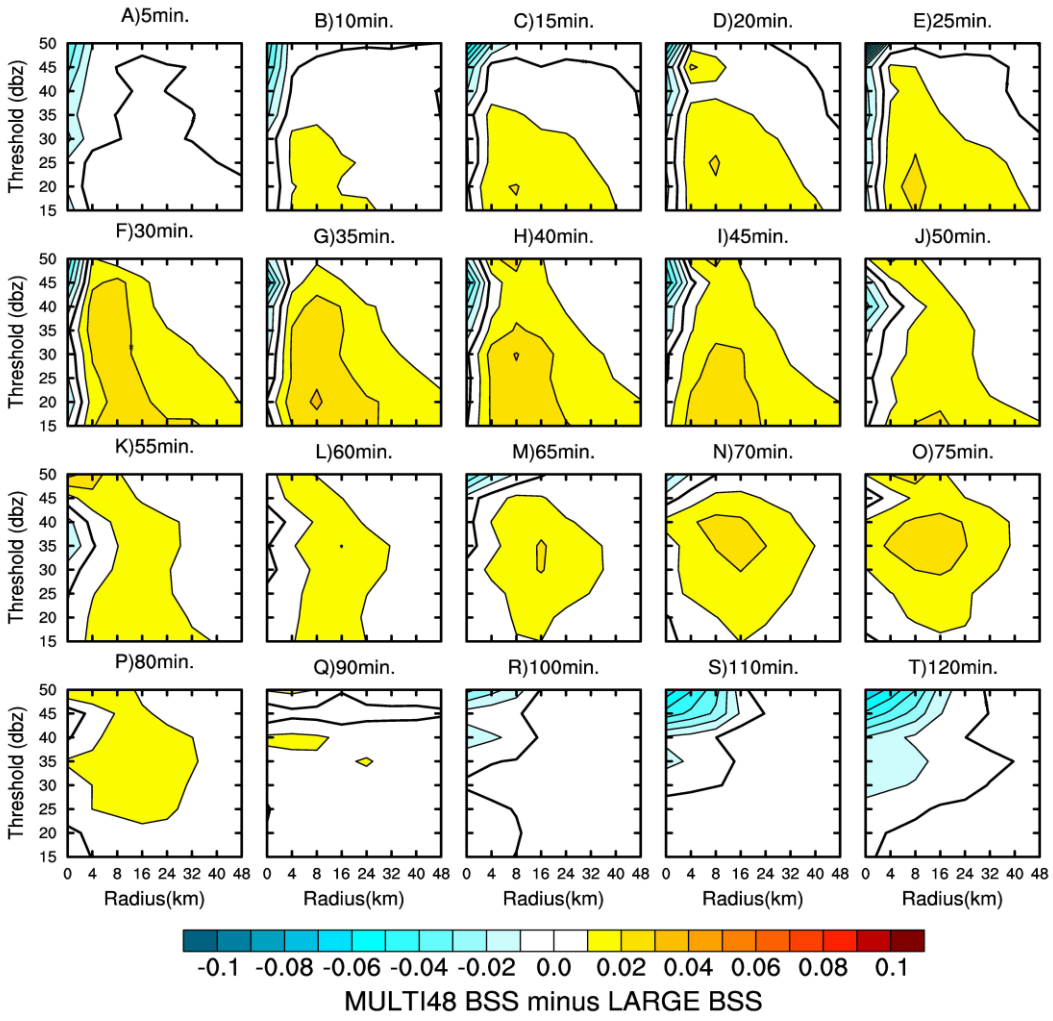

Figure 4.8: As in Fig. 4.6, except for the difference in BSS between MULTI48 and LARGE.

The MULTI48 advantages over LARGE, for both the reflectivity and precipitation forecasts, are attributed to greater consistency of the mesoscale IC perturbations with the analysis errors in the vicinity of the analyzed MCS for MULTI48. For example, member 6 from the LARGE ensemble (hereinafter LARGE_006) shows several spurious cells west of the main MCS in the cold pool region at early lead times (Fig. 4.9a). The corresponding member 6 of the MULTI48 ensemble (hereinafter MULTI48_006) does not show these spurious cells (Fig. 4.9e).

The spurious cells result from strong convergence and moisture perturbations in the vicinity of the MCS cold pool for LARGE_006 that are not present for MULTI48_006 (Fig. 4.9b,c,d,f,g,h). Such perturbations may be consistent with the poorly resolved and poorly assimilated cold pools in the outer domain analysis. However, they are inconsistent with the actual errors of the inner domain analysis of this feature after radar DA (Fig. 4.9i,j,k). Therefore the improved consistency between the mesoscale IC perturbations and analysis errors near the MCS for MULTI48, compared to LARGE, explains the reduction in spurious probability in the cold pool region at early lead times. The excessively large magnitude mesoscale perturbations in and near the initial MCS for LARGE also result in the less sharp probabilistic forecast of the MCS for LARGE. The smaller magnitude mesoscale IC perturbations for MULTI48 are consistent with the initially lower ensemble spread of non-precipitation variables for MULTI, compared to LARGE (Fig. 4.2). Subjectively similar results are also seen in the 2100 UTC case (not shown), although the objective skill is more strongly impacted by another difference between the MULTI48 and LARGE IC perturbations, discussed below.

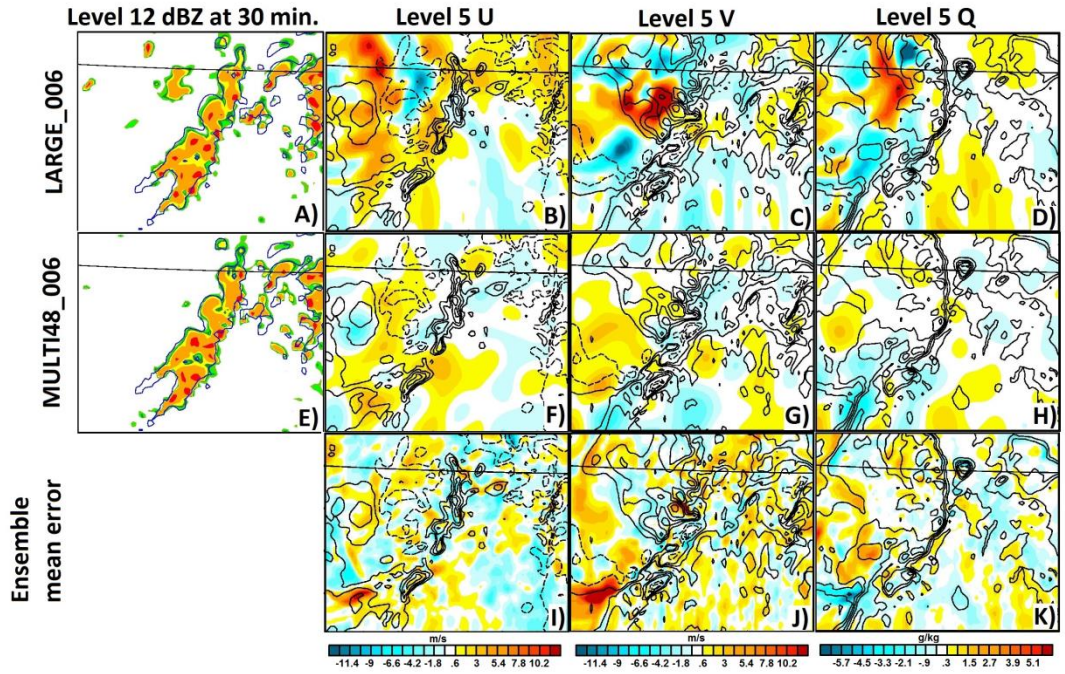

Figure 4.9: Comparison of 0000 UTC case initial perturbations of member 006 from the LARGE and MULTI48 ensembles with the corresponding ensemble mean error. Panels (a) and (c) show the 30-minute reflectivity forecast at model level 12 for LARGE_006 and MULTI48_006, respectively. Panels (b), (c) and (d) show the LARGE_006 perturbation from the ensemble mean for the u component of wind, v component of wind and water vapor, respectively, at model level 5. (f)-(h) are as in (b)-(d) except for the MULTI48_006 perturbation. Panels (i)-(k) show the corresponding ensemble mean error (ensemble mean minus truth). Black contour overlays are the ensemble mean fields with contour interval of 5 m s^{-1} for wind (negative values dashed) and 2 g kg^{-1} for water vapor.

Another difference between the MULTI48 and LARGE IC perturbations in both cases is that positive mid-level mesoscale moisture perturbations appear in some MULTI48 members but not the corresponding LARGE members. These perturbations impact the convective development in such areas, especially for the 2100 UTC case (e.g., Fig. 4.10 for the representative member 23). Although both LARGE_023 and MULTI48_023 develop spurious cells in southern Oklahoma, they develop earlier in MULTI48_023 (Fig. 4.10d) than LARGE_023 (Fig. 4.10a) and become more intense

and numerous in MULTI48_023 (Fig. 4.10e). The most prominent subjective difference between the MULTI48_023 and LARGE_023 ICs is in the level 12 (~750mb) moisture variable near and immediately upstream (i.e., southwest) of where the spurious cells develop (Fig. 4.10c,f). Although both members initially (i.e., before radar DA) have the same perturbation, the cumulative impact of the radar DA, and the interaction with the smaller scales of motion that are resolved during the radar DA, is to moisten the mid-levels (Fig. 4.10f). This leads to more robust development of the spurious cells since there is less dry air entrainment to impede cell development in MULTI48_023. Similar mid-level moisture perturbations are also subjectively seen for the 0000 UTC case (not shown) and objectively evident in the enhanced initial MULTI spread for the mid-level moisture variable (Fig. 4.2f).

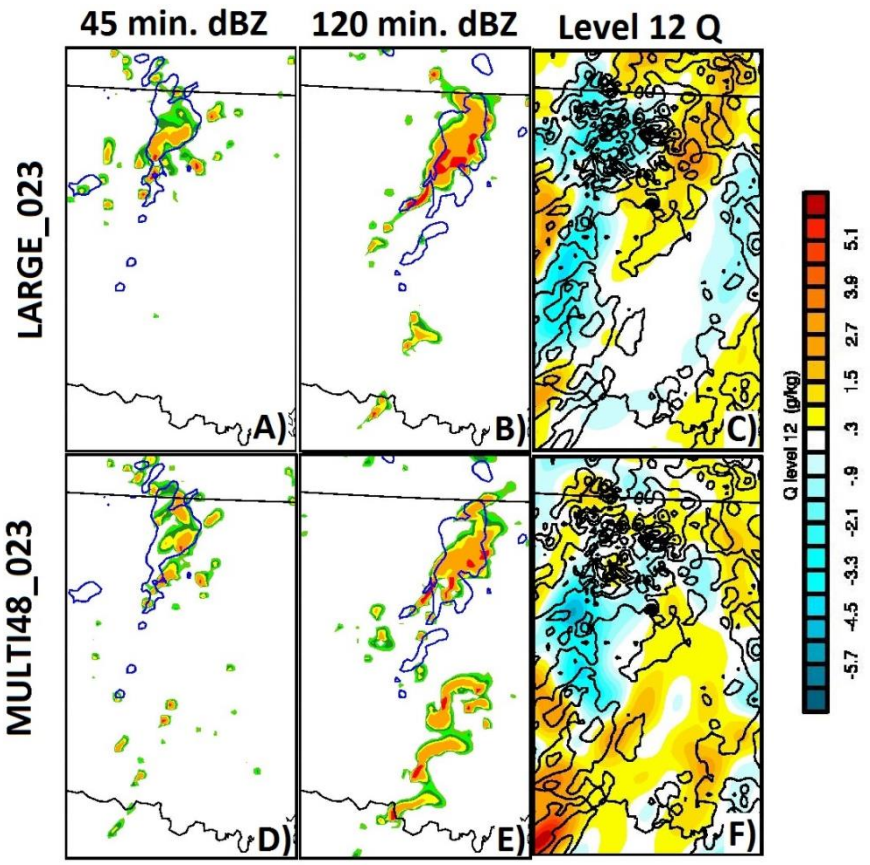

Figure 4.10: Panels (a) and (d) show level 12 reflectivity forecasts at 45 minutes from the 2100 UTC case for members LARGE_023 and MULTI48_023, respectively. Panels (b) and (e) are as in (a) and (d) except at the 120 minute forecast time. Panels (c) and (f) are as in Fig. 4.9d and 4.9h, respectively, except for the 2100 UTC case for members LARGE_023 and MULTI48_023 and for model level 12, instead of level 5.

The above differences between the MULTI48 and LARGE IC perturbations are seen in both the 2100 and 0000 UTC cases. The greater consistency between the IC perturbations and analysis errors near the developing MCS for MULTI48 dominates in the 0000 UTC case, leading to more skillful forecasts than for LARGE for both mesoscale precipitation and storm-scale reflectivity at many lead times and neighborhood radii. However, enhanced spurious convection resulting from unrealistic mesoscale mid-level moisture perturbations in MULTI48 dominates in the 2100 UTC

case, leading to generally lower skill for MULTI48 than LARGE, especially for the mesoscale precipitation forecasts. Thus, while the qualitative impacts of the differences in IC perturbation methods can be understood from this case study, the evaluation of additional cases in Chapter 5 is needed to determine their systematic impact on forecast skill.

4.3b(3) IMPACT OF SMALL SCALE COMPONENT OF IC PERTURBATION METHOD

4.3b(3)(i) Mesoscale hourly accumulated precipitation

The third goal of this part of the study is to understand the impacts of the small scale component of the multi-scale IC perturbations (i.e., MULTI vs. MULTI48). Although the mesoscale IC perturbations have a dominant impact on the precipitation forecast skill, the small scale IC perturbations also contribute to the overall NEP difference for some locations and lead times of the 0000 UTC forecast. For example, the small scale IC perturbations increase the probability of precipitation where a storm is observed along the Oklahoma-Texas border at ~03-05 UTC (Fig. 4.5l-n). This leads to a corresponding increase in probability along the southeast edge of the MCS at ~06-08 UTC (Fig. 4.5o-q). The small scale IC perturbations also contribute to the decrease in forecast probability to the west of the southern half of the MCS at later lead times, especially at ~04-07 UTC (Fig. 4.5m-p). The impact of the small scale IC perturbations in this area is nearly as large, and at some times and places larger than, the impact of the differences in mesoscale IC perturbations. Therefore, while the mesoscale component of the IC perturbations dominates the ensemble forecast skill, the small scale IC

perturbations are not entirely unimportant for the mesoscale hourly accumulated precipitation forecasts. The overall impact on skill of the small scale IC perturbations is neutral or positive since the MULTI skill is generally similar or slightly higher than the MULTI48 skill for the 0000 UTC case (Fig. 4.3a,b,c). For the 2100 UTC case the overall impact on skill of the small scale IC perturbations depends on the forecast lead time (Fig. 4.3d,e,f).

Subjective evaluation of the differences between individual members of the MULTI and MULTI48 ensembles shows that the small scale IC perturbations can directly affect the development of new convection during the early forecast hours (e.g., Fig. 4.11). Given the smaller spatial scale of newly developing convection, it is not surprising that it is particularly sensitive to the small scale IC perturbations. Such convection can then grow upscale during the forecast period, influencing the mesoscale precipitation forecast at later lead times. The continued development of new convection during the early forecast period thus provides a plausible mechanism for the small scale IC perturbations to impact the mesoscale precipitation forecasts at later lead times. The small scale perturbation energy that rapidly develops through downscale energy propagation (i.e., Fig. 4.2) may not have as much impact on such newly developing convection. An example of this mechanism is demonstrated by ensemble member 18 in Fig. 4.11. While the MULTI48 member does not forecast convection along the Texas-Oklahoma border at 0400 UTC (Fig. 4.11a), the corresponding MULTI member does forecast such a convective cell (Fig. 4.11d). The location of the cell in MULTI is slightly west of and weaker than the observed convection at 0400 UTC. However, the upscale growth of the cell results in the southern end of the MCS being farther southeast

and closer to the observed MCS by 0800 UTC for MULTI (Fig. 4.11f) than MULTI48 (Fig. 4.11c). Whether this leads to systematic forecast advantages will be evaluated in Chapter 5.

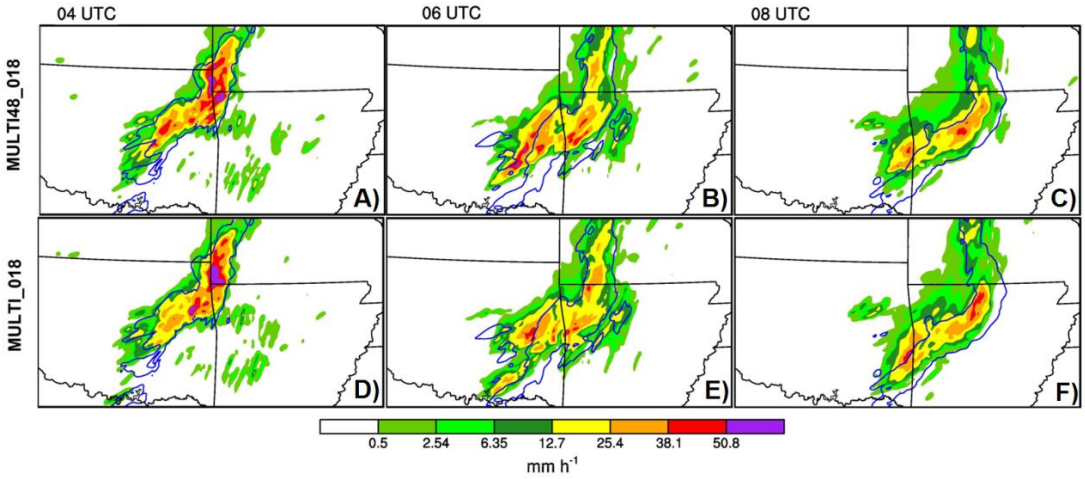

Figure 4.11: Forecasts of hourly accumulated precipitation initialized at 0000 UTC 20 May and valid at (a),(d) 0400 UTC, (b),(e) 0600 UTC and (c),(f) 0800 UTC for member 18 of the (a)-(c) MULTI and (d)-(f) MULTI48 ensemble. Truth contour at the 6.35 mm h^{-1} level is overlaid in blue.

4.3b(3)(ii) Storm scale reflectivity

The main impacts of the small scale IC perturbations for the reflectivity forecasts are on the smallest forecast scales (i.e., no neighborhood radius) during the first hour for the 0000 UTC case (Fig. 4.12). Subjectively, there are two clear impacts of the small scale IC perturbations. First, the MULTI48 NEP forecasts at short lead times show small scale features of the MCS with strong probability gradients that do not necessarily line up with the observation contour (e.g., Fig. 4.13b; blue circles). The small scale IC perturbations in MULTI smooth out the NEP gradient in such cases, making the probabilistic forecasts more consistent with the uncertainty of such features (e.g., Fig. 4.13a; blue circles). This explains the better BSS for MULTI than MULTI48

for zero or small neighborhood radii during the first ~1 h (Fig. 4.12). Second, the small scale IC perturbations increase the probability surrounding the areas of observed precipitation (Fig. 4.7i-p). This is a result of large areas of weak convection resulting from the small scale IC perturbations and is most pronounced at lower reflectivity thresholds.

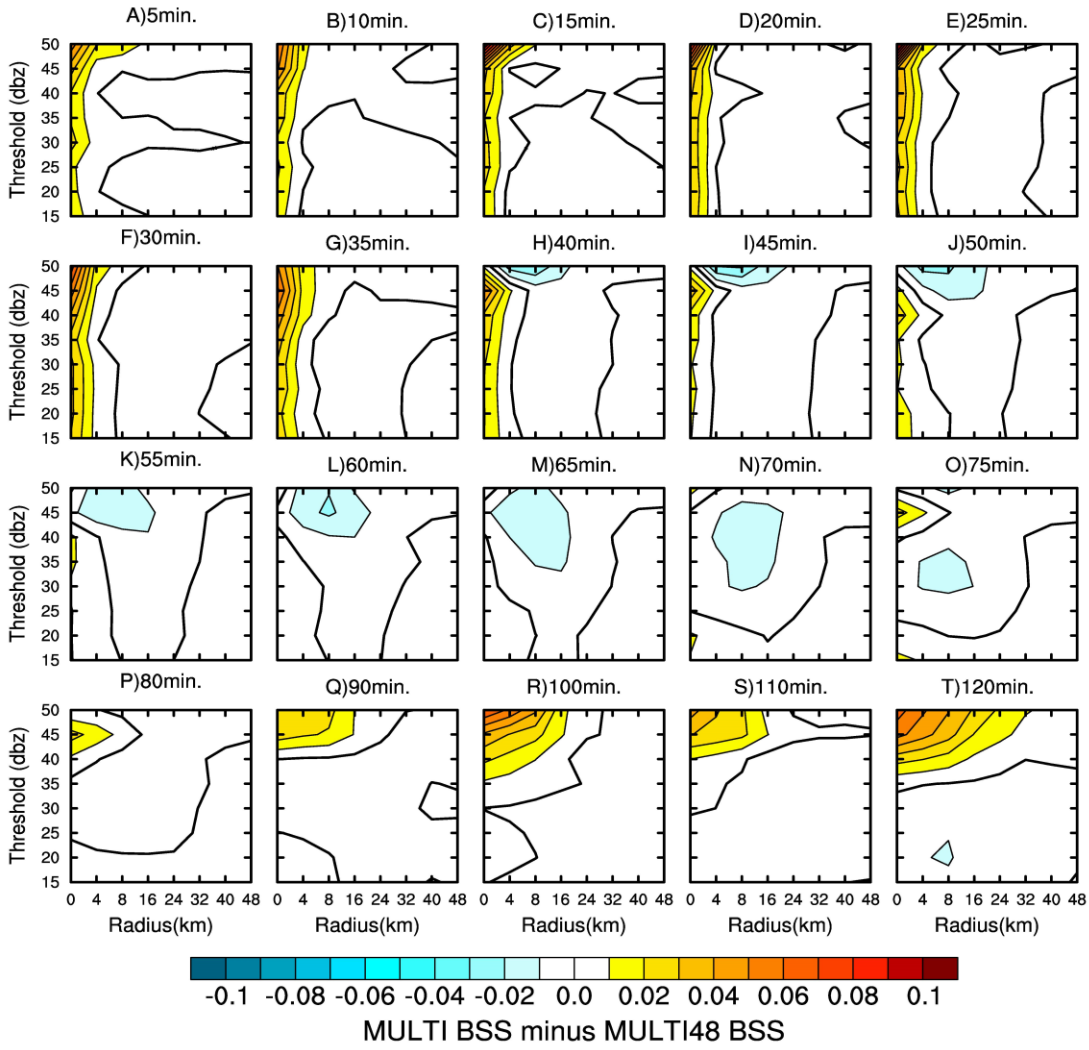

Figure 4.12: As in Fig. 4.6, except for the difference in BSS between MULTI and MULTI48.

Fig. 4.12 also shows more pronounced skill differences at later lead times and high thresholds. Unlike the skill differences during the first hour, there is not a clear

subjective explanation for such differences. The skill differences at later times and high thresholds are likely a result of large sampling variability due to the very small area of observed reflectivity exceeding such high thresholds (e.g., 50 dBZ; Fig 4.14). They are also not consistent throughout the forecast period. For example, at 60 minutes many of the locations with lower NEP for MULTI than MULTI48 occur within the observation contours (Fig. 4.14a), while at 90 and 120 minutes there are several locations with higher NEP for MULTI than MULTI48 within the observation contours (Fig. 4.14b,c). There is little continuity in which features are better predicted between the 90 and 120 minute lead times, even though the difference in skill happens to be of the same sign (Fig. 4.14b,c). Additional cases are particularly necessary for such rare thresholds.

For the 2100 UTC case the positive impact at early lead times quickly extends up to radii of 4-8 km, especially for higher thresholds, instead of remaining at the grid scale, and lasts through ~90 minutes (not shown). This again shows a need for additional cases to quantify the impacts of small scale IC perturbations on forecast skill.

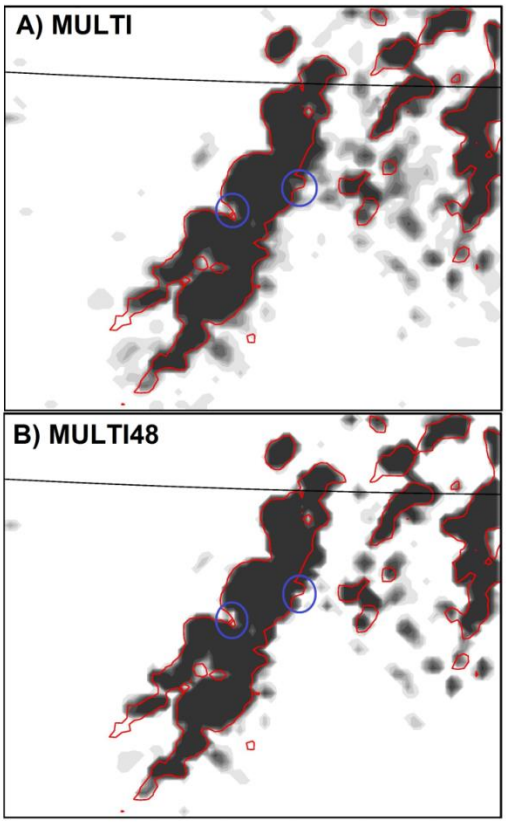

Figure 4.13: NEP forecast of reflectivity exceeding 30 dBZ at the 15 minute lead time (shaded) and observation contour for the 0000 UTC case for (a) MULTI and (b) MULTI48.

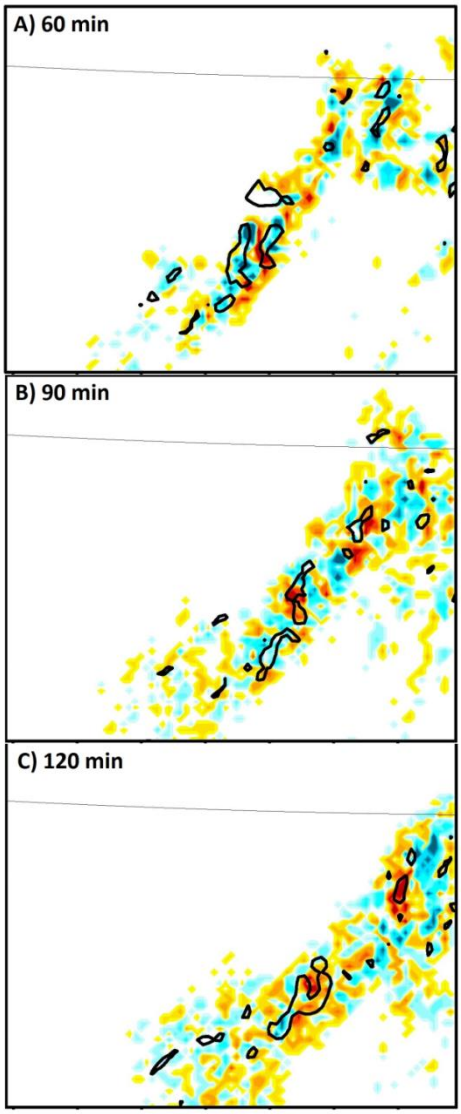

Figure 4.14: Difference in NEP between the MULTI and MULTI48 forecasts of reflectivity exceeding 50 dBZ for the 0000 UTC case at (a) 60 minute lead time, (b) 90 minute lead time and (c) 120 minute lead time.

4.4 Summary and Conclusions

Given the multi-scale nature of convective precipitation forecasts, optimal SSEF design requires an understanding of how flow-dependent multi-scale IC perturbation methods impact convective precipitation forecast skill. This study contributes to such understanding by considering the following three questions. First, what are the impacts

on ensemble forecast skill of generating IC perturbations with a multi-scale ensemble DA system (MULTI), compared to downscaling larger scale IC perturbations from a coarser domain (LARGE)? Second, what role does the mesoscale component (i.e., resolved by both MULTI and LARGE) of the IC perturbation differences have in determining the ensemble forecast skill? Third, what role do the small scale (i.e., only resolved with MULTI) IC perturbations have in the differences between MULTI and LARGE ensemble forecast skill? The impacts of the IC perturbations are evaluated in terms of 2 h reflectivity forecasts over a range of neighborhood radii less than 48 km and in terms of 9 h mesoscale (i.e., 48 km neighborhood radius) hourly accumulated precipitation forecasts. A perfect-model OSSE framework is used to isolate the impacts of IC error and the corresponding IC perturbations. In this chapter, an upscale growing MCS case study is used to address the above three questions both objectively and qualitatively.

The impact of the different IC perturbation methods on the spread of the directly perturbed non-precipitation variables is first evaluated. The LARGE IC perturbations are much more under-dispersive than MULTI on scales less than ~50 km. However, as expected from the results of Durran and Gingrich (2014), the downscale cascade of perturbation energy results in similar perturbation spectra between MULTI and LARGE within ~1 h. The total spread of non-precipitation variables is dominated by the larger scales which initially show less spread for MULTI than LARGE for all variables except for level 12 (~750 mb) moisture.

In addition to the spread of the directly perturbed non-precipitation variables, the skill of the ensemble forecasts of convective precipitation on different time and space

scales is also evaluated. Comparison of the MULTI and LARGE ensembles addresses the question of how the forecast skill is affected by the differences between multi-scale IC perturbations generated with multi-scale ensemble DA versus coarser resolution downscaled IC perturbations. For the 0000 UTC case, MULTI is generally more skillful than LARGE for both storm-scale reflectivity and mesoscale hourly accumulated precipitation forecasts. For the 2100 UTC case, MULTI is generally less skillful than LARGE, especially for the hourly accumulated precipitation forecasts. The case dependence of the results emphasizes the need for systematic evaluation of many diverse cases, which is the goal of Chapter 5. However, the qualitative analysis of the case study provides physical understanding of how the differences in forecast skill result from the IC perturbation differences, aiding interpretation of the systematic results in Chapter 5.

Comparison of MULTI48 and LARGE addresses the question how the differences in the mesoscale component of the IC perturbations affect the forecast skill. With the exception of the first hour of reflectivity forecasts using neighborhood radii of 0-8 km, the differences in ensemble forecast skill are explained mainly by two differences in the mesoscale IC perturbations. First, the MULTI48 perturbations have less amplitude and are more consistent with the analysis error in the vicinity of the analyzed MCS and corresponding cold pool. For the 0000 UTC case, this leads to subjectively and objectively improved forecasts of both reflectivity and hourly accumulated precipitation. Second, the MULTI48 perturbations contain some mesoscale areas of enhanced mid-level moisture away from the main MCS. For the 2100 UTC case, this degrades the forecast skill by further enhancing the effects of

spurious precipitation in some members. Evaluation of additional cases in Chapter 5 is therefore necessary to systematically determine the relative importance and systematic impact on skill of these two main differences between the mesoscale IC perturbations in MULTI48 and LARGE.

The mesoscale increase in mid-level moisture in some members, resulting from radar DA, is hypothesized to be due to incomplete suppression of the effect of spurious convection during DA. Although missing or <5dBZ reflectivity is assimilated as a no-precipitation observation (Aksoy et al. 2009), the effects of the spurious cells that are suppressed by such observations may still be accumulating during the DA period. In particular, the presence of non-precipitating clouds and the detrainment of moisture into the mid-levels during each forecast step is hypothesized to be insufficiently corrected by the assimilation of the no-precipitation observations. Methods to more effectively suppress all effects of spurious convection during radar DA may lead to corresponding improvements in the analysis ensemble for the purpose of SSEFs of convective precipitation. Assimilation of satellite-retrieved cloud water path may provide one such method (e.g., Jones et al. 2014).

Comparison of MULTI and MULTI48 addresses the question of how the presence of small scale IC perturbations in MULTI affects the forecast skill. The impact of the small scale IC perturbations is most pronounced for reflectivity forecasts at neighborhood radii of 0-8 km for about 1 h. However, the duration, magnitude and maximum spatial scale of this advantage varies between the 2100 UTC and 0000 UTC cases. The more pronounced impact of the small scale IC perturbations for the 2100 UTC case may be due to the forecast being initialized earlier in the upscale growth

process, resulting in greater sensitivity to the small scale perturbations and their upscale growth. The hourly accumulated precipitation forecasts also show some impacts of the small scale IC perturbations. While such impacts are generally localized in time and space, they show that the flow-dependent small scale IC perturbations are not necessarily unimportant, even for mesoscale hourly accumulated precipitation forecasts after the first few forecast hours. The impact of the small scale IC perturbations on the mesoscale precipitation forecasts may result mainly from their impact on new convection that develops during the early forecast period. The new cells originate from small scale features, explaining their sensitivity to the small scale IC perturbations. At later times, such cells also influence the mesoscale convective systems, explaining the upscale growth of this impact onto the mesoscale hourly accumulated precipitation forecasts throughout the 9 h forecast period. Determining whether the small scale IC perturbations are systematically advantageous for the hourly accumulated precipitation forecasts requires more cases and will be considered in Chapter 5.

Chapter 5: OSSE study of multi-scale initial condition perturbation methods. Part 2: Systematic results and impact of model error

This chapter provides a systematic investigation of the impact of multi-scale IC perturbations generated on the convection-permitting grid using ensemble-based DA, compared to mesoscale perturbations downscaled from a coarser grid. This chapter is a continuation of the perfect model OSSE case studies in the previous chapter. As in the previous chapter, there are three main goals of the study. First, the performance of ensembles initialized with coarser resolution downscaled perturbations (LARGE) is compared to that of ensembles initialized with multi-scale IC perturbations generated using the GSI-based multi-scale ensemble DA system at the full model resolution (MULTI). The IC perturbations generated with these different methods contain both different resolvable spatial scales and differences on commonly resolved scales. The second goal is therefore to understand the impact of the differences in the IC perturbation method on the commonly-resolved mesoscales. Third, the impact and importance of the small scale (i.e., not resolved by LARGE) IC perturbations is evaluated within the perfect-model OSSE framework.

In Chapter 4, a case study of upscale growth of cellular convection into a long-lived MCS was used to qualitatively understand the impacts of the IC perturbation methods. Forecasts were evaluated in terms of both hourly accumulated precipitation in mesoscale (i.e., 48 km radius) neighborhoods and short term instantaneous reflectivity in storm-scale neighborhoods ranging from 0-48 km. The small scale IC perturbations, resolved only in MULTI, contributed to forecast advantages for ~1 h for neighborhood

radii of 0-8 km for storm-scale reflectivity. In some instances, the small scale IC perturbations also influenced the longer lead time mesoscale hourly accumulated precipitation forecasts. For reflectivity forecasts at times >1 h and/or radii >8 km, and for mesoscale precipitation forecasts, the differences in forecast skill were explained mainly by differences in the mesoscale component of the IC perturbations. Two differences between the mesoscale IC perturbations were noted. First, MULTI IC perturbations were more consistent with the analysis uncertainty than LARGE IC perturbations near the analyzed MCS. Second, spurious convection during radar DA led to unrealistic mesoscale perturbations in mid-level moisture for MULTI away from the observed MCS. For forecasts initialized at 0000 UTC, the first difference was most important, leading to more skillful MULTI forecasts than LARGE. However, for forecasts initialized at 2100 UTC the second difference played a greater role, leading to generally less skillful MULTI forecasts than LARGE. Therefore, it was not clear if the advantages of MULTI would lead to systematically more skillful forecasts than LARGE when evaluated over many cases.

This chapter extends the OSSE study to 11 diverse cases of mid-latitude convection in the central United States. The purpose of this chapter is to assess the robustness of the results in Chapter 4 using systematic evaluation of forecasts under different synoptic scale environments, including many convective systems with different levels of mesoscale organization.

The remainder of this chapter is organized as follows. Section 5.2 provides a brief review of the selected cases and methods of determining statistical significance of

the systematic results. The OSSE results are then presented in Section 5.3. Section 5.4 contains a summary and conclusions.

5.2 Methods

5.2a Review of selected cases

Eleven diverse cases are selected to determine the robustness of the results in Chapter 4. The same 10 cases used in Chapter 2 are also adopted for this OSSE study, with the addition of the 2100 UTC 19 May forecast case from Chapter 4 for a total of 11 cases. Like the real data cases (Chapter 2), the nature run simulations include a variety of forcing mechanisms and levels of convective organization ranging from disorganized cellular convection to supercells to long-lived mesoscale convective systems (MCSs). An advantage of the OSSE framework is that the nature run provides the exact truth values for verification of all forecast variables on the same grid as the forecast variables. For the hourly accumulated precipitation forecasts, rectangular verification domains for each case are chosen to include the areas of active convection at all lead times while excluding large areas where convection is neither observed nor forecast. For the 2 h lead time reflectivity forecasts, smaller rectangular verification domains are used to encompass each subjectively identified mesoscale area of organized convection during the first two forecast hours. Some of the forecast cases contain multiple areas of mesoscale organized convection, resulting in a total of 18 unique verification domains for the reflectivity verification. Since different MCSs on the same case occur within the same larger scale environment, such MCSs are not treated as independent samples for statistical significance testing.

5.2b Statistical significance tests

The statistical significance of differences in Brier Skill Score (BSS) is determined using permutation resampling of the 11 (10 for real-data) cases (Hamill 1999). For reflectivity forecasts with multiple MCSs on the same day, results are first aggregated for that day and treated as a single sample since the different MCSs may not be statistically independent. For the hourly accumulated precipitation forecasts, statistical significance is plotted at the 80% confidence level. The relatively low confidence level is chosen because an 11-sample data set is rather small to expect very high levels of confidence. While this choice does leave a 20% chance of a "significant" result occurring due to random chance, it allows the more robust results to be distinguished from the less robust results. For the reflectivity forecasts, statistical significance is plotted at the 90% confidence level because the impacts of the IC perturbation methods on reflectivity forecast skill are more consistent from case to case, allowing for greater levels of statistical significance to be established.

5.3 Impact of IC perturbation method on forecast skill

5.3a Non-precipitation forecasts

The ensemble average spectra of the ensemble perturbations (i.e., spread) and the spectra of ensemble mean error for the directly perturbed non-precipitation variables are first evaluated using the u-component of wind at level 5 (~900 mb; Fig. 5.1), which is also representative of other levels and other non-precipitation variables. As in Chapter 4, one-dimensional detrended Fourier spectra are calculated along east-west grid lines then averaged over all possible such grid lines (Skamarock 2004). Consistent

with Durran and Gingrich (2014) and the case study in Chapter 4, initial under-dispersion at scales less than ~50 km in LARGE quickly catches up to MULTI as a result of downscale energy propagation (Fig. 5.1). On average, the MULTI and LARGE spread at small scales are nearly indistinguishable after ~40-60 minutes (Fig. 5.1). The spectra are calculated over the entire convection-permitting forecast domain in order to include the full spectrum of spatial scales. Therefore, it should also be noted that more substantial localized differences in small scale ensemble spread may be present after 1 h but not obvious in the spectra averaged over the entire domain and multiple cases.

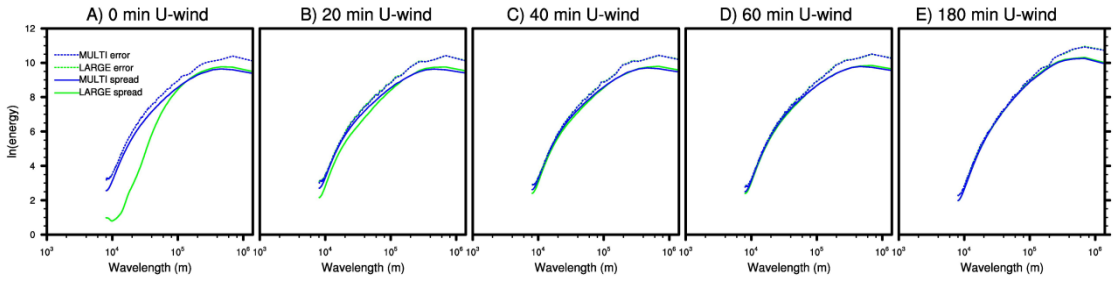

Figure 5.1: Fourier spectra decomposition of ensemble perturbations (ensemble member minus ensemble mean, averaged over all members; solid) and ensemble mean error (dashed) for the u component of wind at model level 5 (~900 mb) averaged over all 11 cases at (a) the analysis time, (b) 20-minute forecast time, (c) 40-minute forecast time, (d) 60-minute forecast time, and (e) 180-minute forecast time.

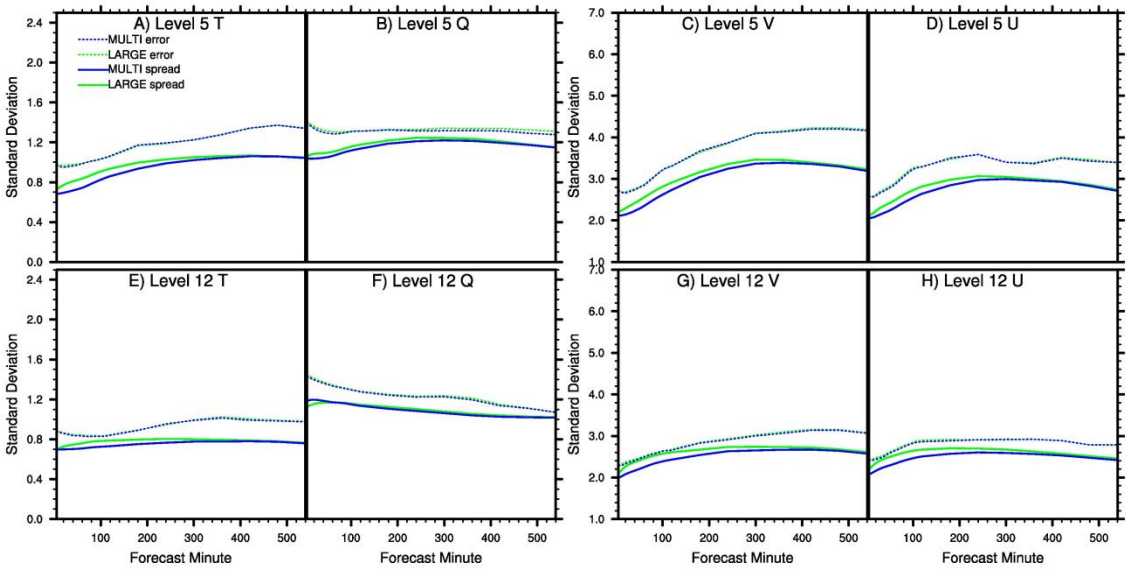

Figure 5.2: Ensemble spread (i.e., standard deviation) and ensemble mean RMSE as a function of forecast lead time averaged over all 11 cases at model level 5 (~900 mb) for (a) temperature (K), (b) water vapor mixing ratio (g kg^{-1}), (c) v wind component (m s^{-1}), (d) u wind component (m s^{-1}), and (e)-(h) as in (a)-(d) except at model level 12 (~750mb).

The systematic comparison of total ensemble spread and ensemble mean error (Fig. 5.2) also shows similarities to the case study in Chapter 4. All variables except level 12 (~750 mb) water vapor are initially more under-dispersive for MULTI than LARGE (Fig. 5.2). The decreased initial spread for MULTI is mainly due to the differences on scales larger than 100 km (Fig. 5.1). The greater initial MULTI spread, compared to LARGE, for level 12 water vapor suggests that the impacts of spurious convection during the radar DA period, discussed in Chapter 4, are a systematic feature of this DA system. However, as will be shown in Section 5.3b, the negative impact of this feature on forecast skill for the 2100 UTC 20 May case in Chapter 4 is not representative of the systematic results.

Most variables and lead times show much less ensemble spread than ensemble mean error for both MULTI and LARGE (Fig. 5.2). This systematic under-dispersion

cannot be attributed to insufficient sampling of model errors in the ensemble design because of the perfect-model OSSE framework. It also likely is not attributable to the LBC perturbations, generated on the outer domain which does contain model error, since it is present from the beginning of the forecasts. However, the LBC perturbations may contribute to limiting the spread growth after ~5-6 hours (Fig. 5.2).

5.3b Convective precipitation forecasts

Since 50 km is the approximate scale below which LARGE is particularly under-dispersive compared to MULTI (Fig. 5.1), the 48 km wavelength is used to distinguish "small scale" and "mesoscale" IC perturbations. The small scale IC perturbations are only resolved in MULTI while MULTI, MULTI48 and LARGE all resolve the mesoscale IC perturbations. The following sub-sections evaluate the convective precipitation ensemble forecast differences between MULTI and LARGE, MULTI48 and LARGE, and MULTI and MULTI48, consistent with the three goals of this study. As in Chapter 4, the convective precipitation forecasts are evaluated separately for mesoscale (48 km neighborhood radius and out to 9 h) hourly accumulated precipitation forecasts and storm scale (neighborhood radii between 0 and 48 km and out to 2 h) instantaneous reflectivity forecasts in order to account for the potentially different perspectives of different SSEF users.

5.3b(1) OVERALL IMPACT OF IC PERTURBATION METHOD

5.3b(1)(i) Mesoscale hourly accumulated precipitation

The first goal of this study is to understand the systematic impacts of the different methods of generating IC perturbations (i.e., downscaling vs. multi-scale DA) by comparing MULTI and LARGE. For the mesoscale precipitation forecasts, MULTI is more skillful than LARGE during the first hour and, for the 2.54 and 6.35 mm h^{-1} thresholds, after 4 h (Fig. 5.3). MULTI is slightly less skillful than LARGE at 8-9 h for the 12.7 mm h^{-1} threshold but this difference is not statistically significant (Fig. 5.3). The MULTI skill advantages are significant at the 1 h lead time for all thresholds, the 7 h lead for the 6.35 mm h^{-1} threshold and the 6 and 8 h lead times for the 2.54 mm h^{-1} threshold. The 7 and 9 h lead times at the 2.54 mm h^{-1} thresholds are almost significant, with p-values of 0.2024 and 0.236, respectively (not shown). Therefore the 0000 UTC case from Chapter 4 is representative of the systematic results in that the significant differences in mesoscale precipitation forecast skill all favor MULTI over LARGE.

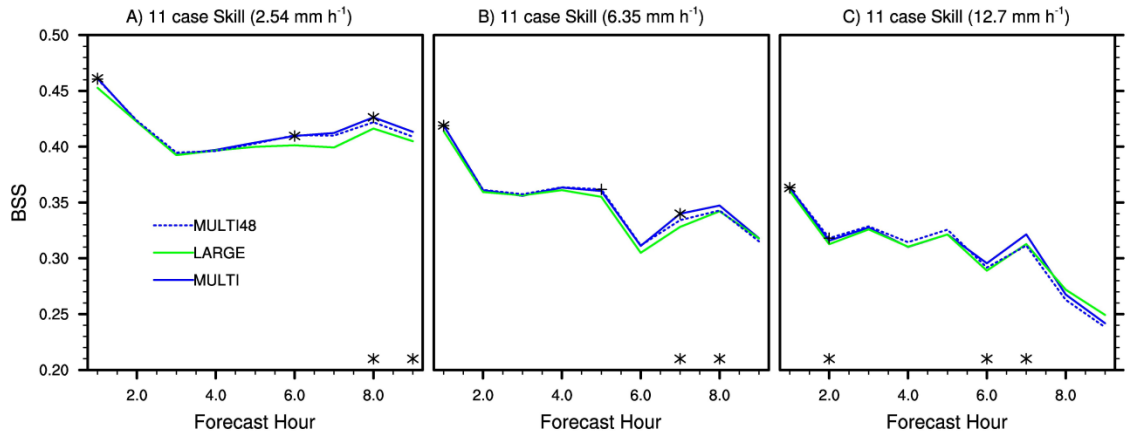

Figure 5.3: Brier Skill Score (BSS) of the Neighborhood Ensemble Probability (NEP) forecasts over all 11 cases of hourly accumulated precipitation for thresholds of (a) 2.54 mm h^{-1}, (b) 6.35 mm h^{-1} and (c) 12.7 mm h^{-1}. Statistical significance is plotted at the 80% confidence level, with significant differences between MULTI and LARGE, MULTI48 and LARGE, or MULTI and MULTI48 indicated by asterisks on the MULTI line, plus signs on the MULTI48 line, or asterisks along the horizontal axis, respectively.

5.3b(1)(ii) Storm-scale reflectivity

The impacts of the different IC perturbation methods are also evaluated on forecasts over smaller space and time scales using instantaneous reflectivity at 5 minute intervals and neighborhood radii of 0-48 km (Fig. 5.4). For the storm-scale reflectivity forecasts, MULTI is again more skillful than LARGE where there are statistically significant differences (Fig. 5.4). The statistically significant MULTI advantages last for about 65 minutes at the lower thresholds (e.g., 20-25 dBZ; Fig. 5.4) and about 45 minutes at the higher thresholds (e.g., 40 dBZ; Fig. 5.4). As in Chapter 4, the MULTI advantage is most pronounced at ~35 minute lead time. Therefore the advantages for MULTI in the 0000 UTC case of Chapter 4 are also representative of the systematic results in terms of storm-scale reflectivity forecasts. Since the differences between MULTI and LARGE include both smaller scale perturbations in MULTI and different

methods of generating the mesoscale perturbations, the impacts of these two factors are distinguished in the following sub-sections.

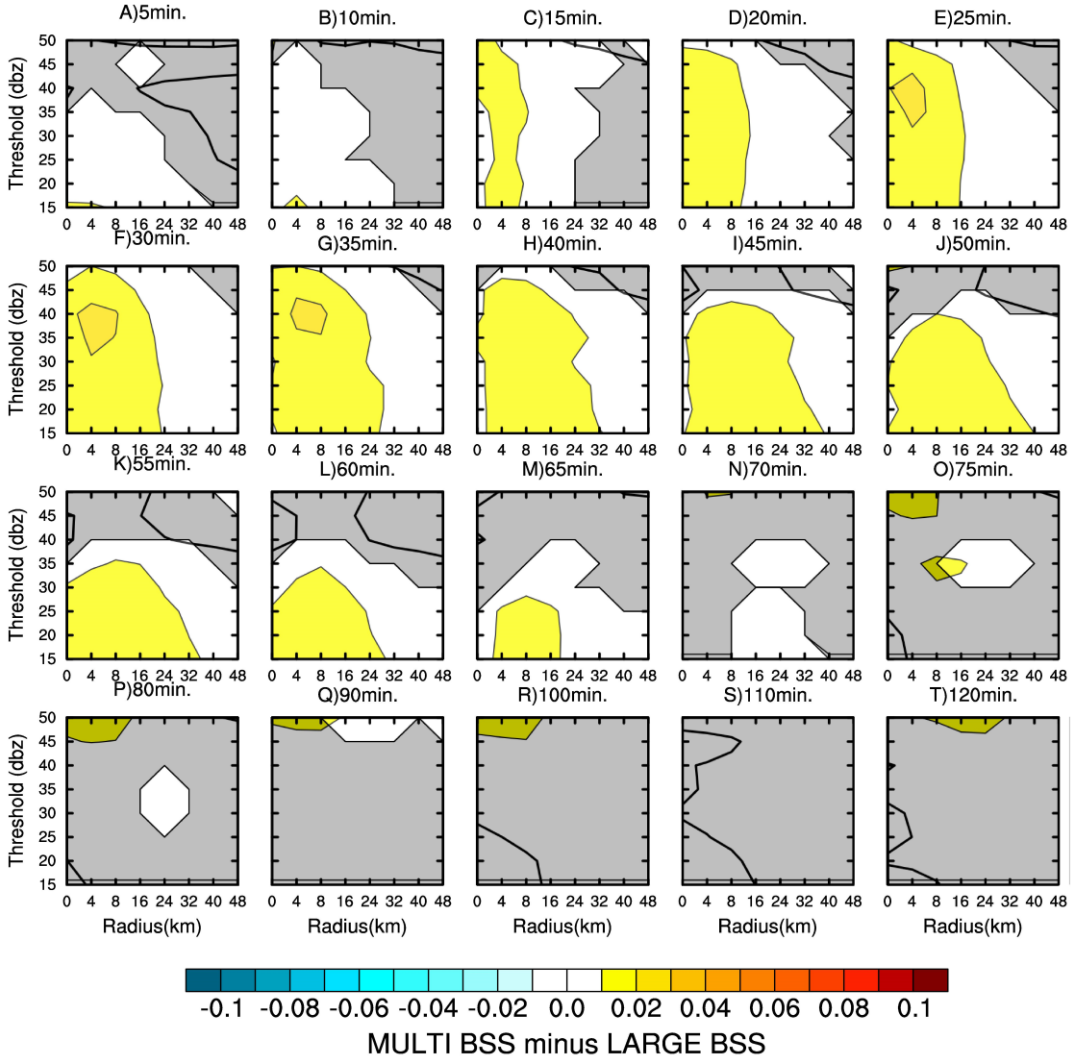

Figure 5.4: Difference in BSS between the MULTI and LARGE ensembles, averaged over all 18 MCS cases, for reflectivity at model level 12 at five minute intervals during the first 80 minutes and at ten minute intervals between 80 and 120 minutes. The vertical axis on each panel is the reflectivity threshold (dBZ) and the horizontal axis is the neighborhood radius (km). Values that are not statistically significant at the 90% level are covered by shading. The unshaded values are statistically significant at the 90% level.

5.3b(2) IMPACT OF MESOSCALE COMPONENT OF IC PERTURBATION METHOD

5.3b(2)(i) Mesoscale hourly accumulated precipitation

The second goal of this study is to understand the systematic impacts of the differences between MULTI and LARGE on commonly resolved scales (i.e., mesoscales), as opposed to the smaller scales resolved only by MULTI. MULTI48 is therefore compared to LARGE in order to focus only on the mesoscale IC perturbations (Fig. 5.3). The differences in mesoscale precipitation forecast skill between MULTI48 and LARGE are similar in many ways to the differences between MULTI and LARGE. Specifically, MULTI48 is more skillful than LARGE after ~4h, except for the last few hours at the 12.7 mm h^{-1} threshold and the last couple of hours at the 6.35 mm h^{-1} threshold. The impact of the mesoscale IC perturbations is similar in magnitude to the impact of the small scale IC perturbations for the 12.7 mm h^{-1} threshold. The similarity between MULTI and MULTI48 skill at the 2.54 and 6.35 mm h^{-1} thresholds is consistent with the dominance of the mesoscale IC perturbations for mesoscale precipitation forecasts in the 20 May case study (Chapter 4). However, unlike the differences between MULTI and LARGE, the differences between MULTI48 and LARGE are generally not statistically significant, except at the 1h lead time (Fig. 5.3). This shows that the small scale IC perturbations, omitted from MULTI48, also play an important role as further discussed in section 5.3b(3).

One of the main qualitative differences between MULTI48 and LARGE in the Chapter 4 case study was a sharper MCS forecast at early lead times resulting from smaller magnitude mesoscale perturbations near the developing MCS that were more

consistent with the analysis uncertainty for MULTI48. Subjective examination reveals a similar impact in the other cases as well (not shown). The case study also revealed unrealistically large mid-level moisture perturbations for MULTI48 away from the observed MCS which can negatively impact the precipitation forecast skill in at least some cases (e.g., the 2100 UTC case in Chapter 4). Fig. 5.3 shows that such mid-level moisture perturbations do not lead to systematic negative impacts on precipitation forecast skill. Instead the greater consistency between mesoscale IC perturbations and IC uncertainty near the analyzed convective systems leads to generally greater precipitation forecast skill for MULTI48 and MULTI than LARGE for most times and thresholds. This shows that the method of generating flow dependent mesoscale IC perturbations directly on the convection-permitting grid with cycled multi-scale DA is more optimal for mesoscale precipitation forecasting than downscaled mesoscale IC perturbations from a coarser ensemble.

The differences in mesoscale IC perturbations result in skill advantages for MULTI48 and MULTI, compared to LARGE, at both early (1 h) and later (~5-9 h) lead times (Fig. 5.3a,b). However, the skill is generally statistically indistinguishable among the forecasts at ~2-4 h lead times (Fig. 5.3). One possible explanation is that this is a result of the diurnal cycle of convective precipitation. Many of the cases show more convection over larger areas during the evening hours (i.e., first ~4 h) than the overnight hours when only the better organized systems tend to be maintained (after ~0300-0400 UTC; not shown). It may be that the advantage of MULTI48 is greatest for the more organized long-lived MCSs, allowing the advantage to be objectively more pronounced after ~4 h when most of the precipitation is associated with such systems.

5.3b(2)(ii) Storm-scale reflectivity

The impact of the mesoscale component of the differences in IC perturbations on the reflectivity forecasts is also similar to the 0000 UTC case of Chapter 4. For neighborhood radii greater than ~4 km, the advantage over LARGE is slightly more pronounced, and lasts longer, for MULTI48 than MULTI (Fig. 5.5 vs. Fig. 5.4). Statistically significant MULTI48 advantages for some radii and thresholds persist throughout the 2 h forecast period, although the differences become very small by the end of the period (Fig. 5.5). Like the hourly accumulated precipitation forecasts, the advantage of MULTI48 is also due to the mesoscale IC perturbations being systematically more consistent with the analysis error near ongoing convective systems. This leads to sharper convective scale forecasts as demonstrated in Chapter 4 and confirmed with subjective evaluation of the individual cases (not shown). Generating the mesoscale IC perturbations directly on the convection-permitting grid is therefore advantageous for storm-scale reflectivity forecasts, in addition to the mesoscale precipitation forecasts.

The impact on ensemble spread and accuracy of the smaller magnitude mesoscale perturbations in MULTI48 than LARGE is quantified with the dispersion and error Fractions Skill Score (dFFS and eFSS, respectively; Dey et al. 2014). As described in greater detail in Dey et al (2014), the Fractions Score (FS) is the mean square difference between the forecast and observed neighborhood probability (NP) field. The observed NP is calculated the same way as the forecast NP, instead of using a binary verification field as in the other NEP skill scores presented herein. The

Fractions Skill Score (FSS) is then calculated as 1-FS/FS_{ref} where FS_{ref} is the FS that would be obtained if there were no overlap between the forecast and observed NP fields (i.e., the sum of the mean square forecast and observed NP fields). The dFSS is calculated as the average FSS between all possible member-member pairs as a measure of ensemble spread. Smaller values of dFSS indicate greater spread. The eFSS is calculated as the average FSS between all member-observation pairs and is a measure of the deterministic forecast accuracy of the ensemble members. Smaller values of eFSS indicate greater error. An advantage of this method is that it can be calculated over a range of radii to understand the scale dependence of the ensemble characteristics.

Fig. 5.6 shows that after ~20 minutes the dFSS of reflectivity forecasts is systematically larger for MULTI48 than LARGE, indicating less ensemble spread for MULTI48 (Fig. 5.6c). This is a result of the smaller magnitude mesoscale IC perturbations in MULTI48. The larger spread for MULTI48 than LARGE during the first ~20 minutes is likely due to the fact that hydrometeor IC perturbations are not present in the LARGE ensemble. It therefore takes some time for the directly perturbed variables to generate reflectivity spread. After ~15-30 minutes, depending on spatial scale, the smaller spread for MULTI48 also corresponds to larger eFSS values, indicating less error for MULTI48 than LARGE (Fig. 5.6f). Therefore, Fig. 5.6 shows that the MULTI48 members are systematically both closer to each other and closer to the observations than the LARGE members, consistent with the generally more skillful forecasts for MULTI48 (Fig. 5.5).

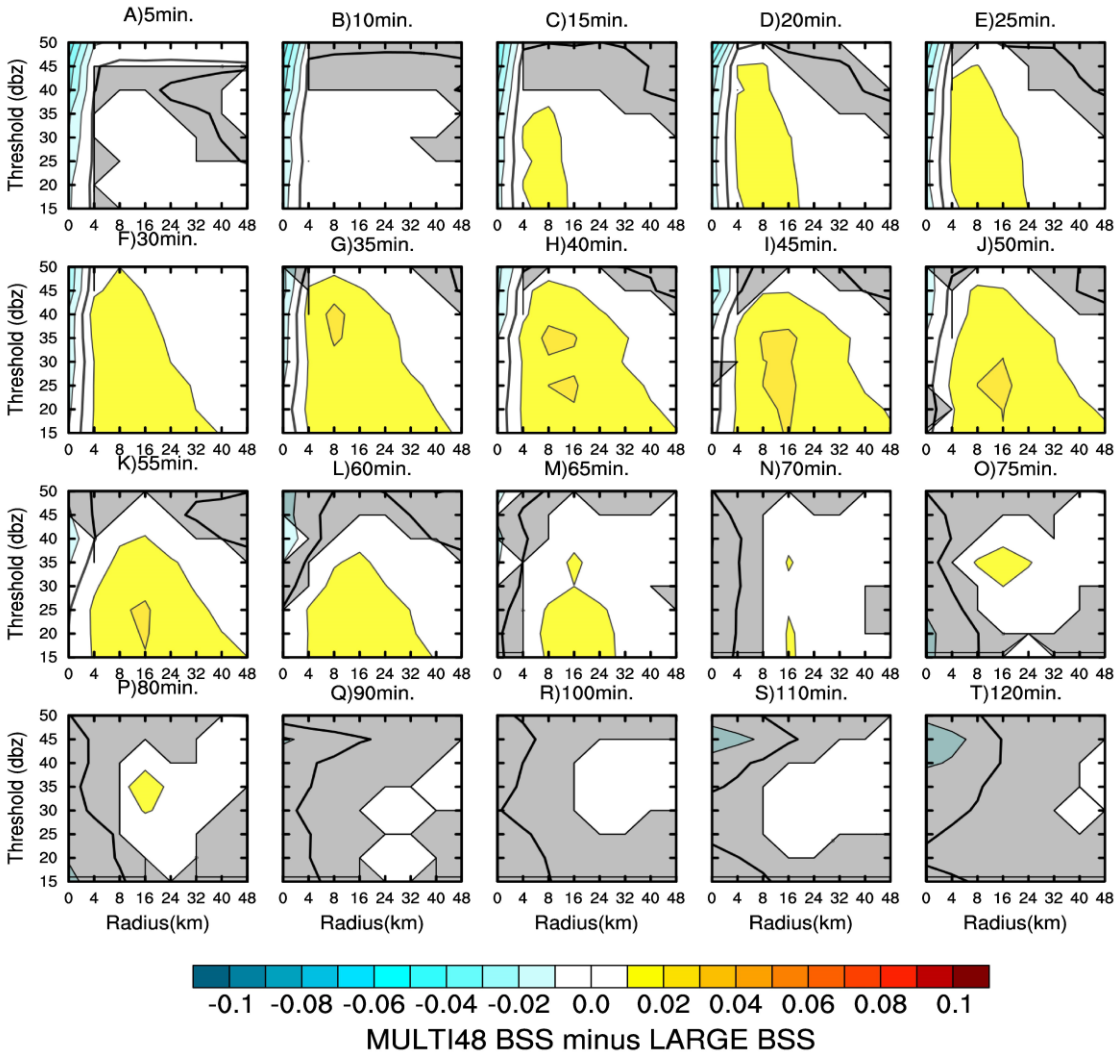

Figure 5.5: As in Fig. 5.4, except for the BSS difference between MULTI48 and LARGE.

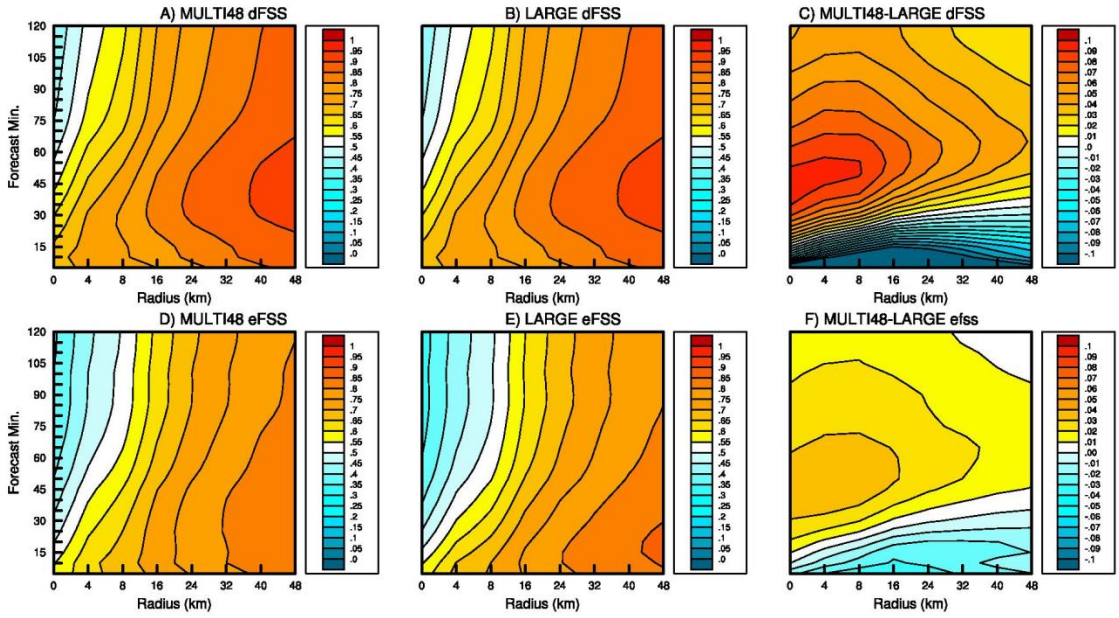

Figure 5.6: Dispersion Fractions Skill Score (dFSS) for the (a) MULTI48 ensemble, (b) LARGE ensemble, (c) difference between the MULTI48 and LARGE ensembles and error Fractions Skill Score (eFSS) for the (d) MULTI48 ensemble, (e) LARGE ensemble and (f) difference between the MULTI48 and LARGE ensembles.

5.3b(3) IMPACT OF SMALL SCALE COMPONENT OF IC PERTURBATION METHOD

5.3b(3)(i) Mesoscale hourly accumulated precipitation

The third goal of this study is to understand the systematic impacts of the small scale IC perturbations which are resolved by MULTI but not LARGE. MULTI is therefore compared to MULTI48 which does not contain such small scale IC perturbations. For the mesoscale precipitation forecasts, localized impacts of the small scale IC perturbations were noted in Chapter 4. The upscale impacts of the small scale IC perturbations on mesoscale hourly accumulated precipitation forecasts out to 9 h were explained by their impact on the development of new convection from small scale

features during the early forecast period. However, it was not clear from a single case study if such impacts would systematically improve the mesoscale precipitation forecasts. The systematic verification reveals that at forecast hours 2-5, MULTI is slightly less skillful than MULTI48 at all thresholds, although the difference is only significant at the 2 h lead time for the 12.7 mm h^{-1} threshold (Fig. 5.3). This negative impact of the small scale IC perturbations may be related to an initial enhancement of disorganized weak convection surrounding the observed convective systems (Chapter 4). Starting at ~6 h there is a more pronounced and positive impact of the small scale IC perturbations. MULTI is more skillful than MULTI48 at these later lead times for all thresholds, with statistical significance at 8-9 h, 7-8 h and 6-7 h at 2.54, 6.35 and 12.7 mm h^{-1} thresholds, respectively (Fig. 5.3). Thus, there are statistically significant systematic advantages of the upscale growth of the flow-dependent small scale IC perturbations on mesoscale precipitation forecast skill out to 9 h. This shows that it is important to explicitly include such perturbations in the IC perturbation design, rather than rely on the downscale propagation of perturbation energy indicated by Fig. 5.1 and Durran and Gingrich (2014).

5.3b(3)(ii) Storm-scale reflectivity

As in the 0000 UTC case study of Chapter 4, the small scale IC perturbations in MULTI also systematically improve the reflectivity forecasts on small forecast scales (i.e., no neighborhood radius; Fig. 5.7). The significant advantage of the small scale IC perturbations lasts for about 1 h and is most pronounced during the first ~30 minutes (Fig. 5.7). In other words, the advantage of the small scale IC perturbations for storm-

scale reflectivity forecasts occurs on similar space and time scales as the directly perturbed non-precipitation variables (Fig. 5.1). As shown in Chapter 4, this advantage corresponds to subjectively smoother probability gradients where grid scale details of the observation contour at a particular threshold are not well forecast. This contrasts with the mesoscale precipitation forecasts which are improved by upscale growth of the explicitly added flow-dependent small scale IC perturbations. For storm-scale reflectivity forecasts, the impact of the small scale IC perturbations is limited to the very small scales and the time period before downscale propagation generates sufficient perturbation energy on such scales (Fig. 5.1).

The impact on ensemble spread and accuracy of the small scale IC perturbations is also quantified with the dFSS and eFSS (Fig. 5.8). Compared to MULTI48, MULTI initially has greater spread at the grid scale which grows to slightly larger scales during the first ~45-60 minutes (Fig. 5.8c). This difference in spread, resulting from the small scale IC perturbations in MULTI, remains maximized in neighborhoods of 0-8km, consistent with the impact on ensemble forecast skill occurring on such scales (Fig. 5.7). Although greater spread for MULTI corresponds to more error of the ensemble members (Fig. 5.8f), the NEP skill is greater for MULTI than MULTI48 at similar times and scales (Fig. 5.7). The impact of the small scale IC perturbations on both ensemble spread and accuracy begins to diminish after ~60 minutes (Fig. 5.8c,f), also consistent with Fig. 5.7.

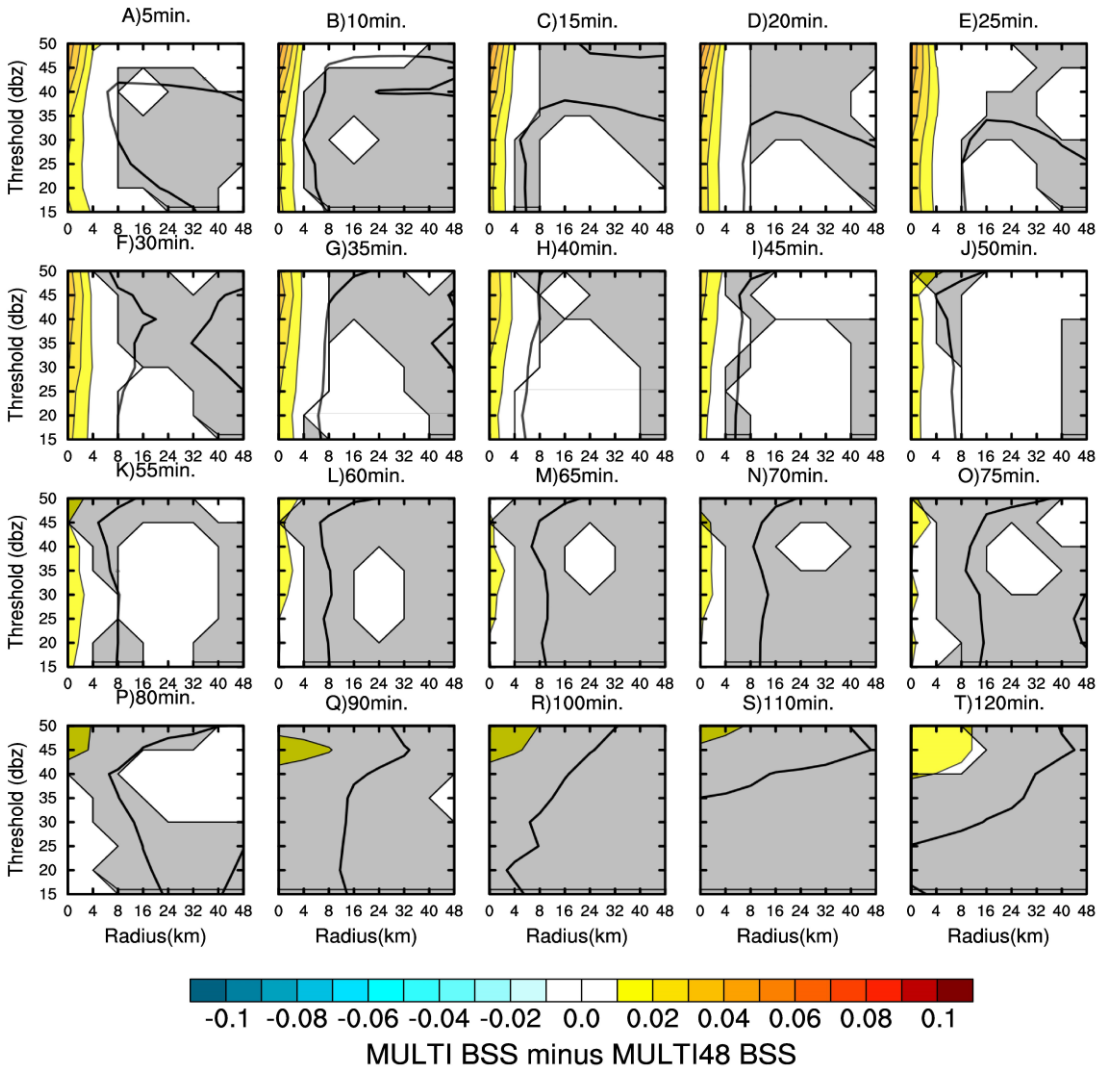

Figure 5.7: As in Fig. 5.4, except for the BSS difference between MULTI and MULTI48.

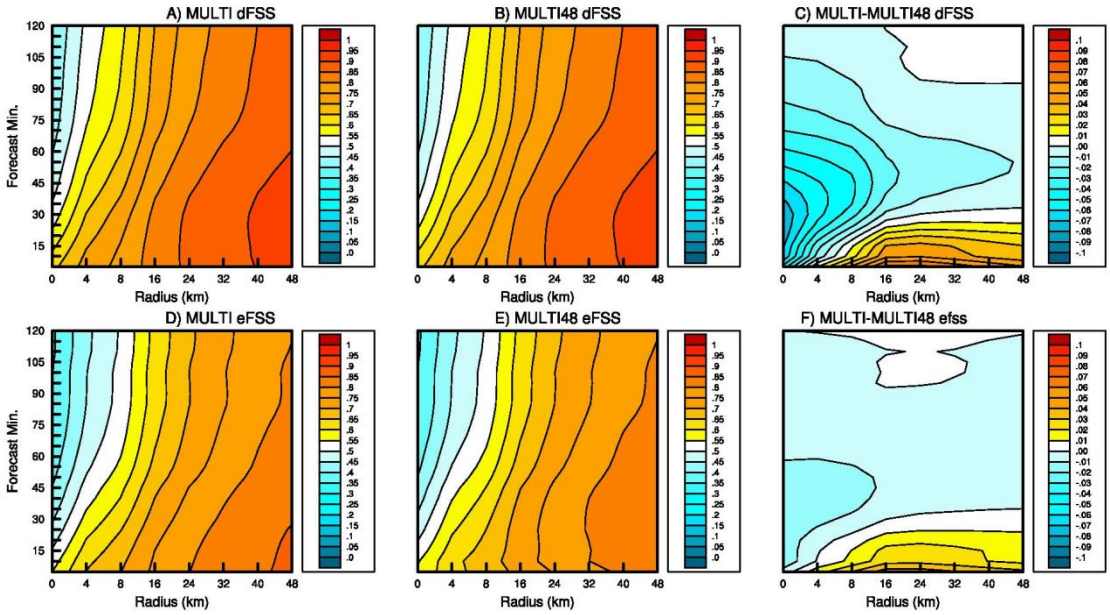

Figure 5.8: As in Fig. 5.6, except for the MULTI and MULTI48 ensembles.

5.4 Summary and Conclusions

In this chapter the systematic impacts of different IC perturbation methods on the skill of SSEFs are investigated. Three aspects of the IC perturbations are investigated. First, the overall impacts on forecast skill of IC perturbations generated with a multi-scale data assimilation (DA) system (MULTI) are compared to IC perturbations downscaled from a larger scale ensemble (LARGE). Second, the impacts of the differences between the IC perturbation methods on the commonly resolved mesoscales are isolated from the impacts of the convective scale IC perturbations that are only resolved by MULTI. This is accomplished by comparing the MULTI perturbations with small scales (i.e., <48 km) filtered out (MULTI48) to LARGE. Third, the impacts of these convective scale perturbations are investigated by comparing MULTI to MULTI48. Since different users may be interested in different time and space scales of the convection forecasts, the forecasts are evaluated in terms of both

hourly accumulated precipitation in mesoscale (i.e., 48 km radius) neighborhoods out to 9 h and storm scale (i.e., neighborhood radii of 0-48 km) reflectivity forecasts out to 2 h.

The MCS upscale growth case study in Chapter 4 showed that the small scale IC perturbations in MULTI result in more skillful reflectivity forecasts during the first forecast hour for neighborhood radii of ~0-8 km, depending on the forecast initialization time. The small scale IC perturbations were also shown to have the potential to impact the mesoscale precipitation forecasts through upscale growth of new convection developing from small scale features during the early forecast period. Two qualitative differences between the mesoscale IC perturbations in MULTI and LARGE were also identified in Chapter 4. First, smaller magnitude mesoscale IC perturbations in MULTI than LARGE near the analyzed MCS were more consistent with the analysis uncertainty and therefore advantageous for forecast skill. Second, enhanced mesoscale mid-level moisture perturbations in MULTI made the environment away from the observed MCS more susceptible to spurious convection and were therefore disadvantageous for forecast skill.

Like the Chapter 4 case study, the directly perturbed non-precipitation variables are initially very under-dispersive on scales < ~50 km for the LARGE ensemble during the first ~40-60 minutes of the forecast. Also consistent with Chapter 4, there is initially less spread of the non-precipitation variables, except for mid-level moisture, for MULTI than LARGE. This is also consistent with the smaller magnitude mesoscale IC perturbations near analyzed convective systems for MULTI. The larger magnitude of the initial mid-level moisture perturbations for MULTI suggests that the insufficient

suppression of spurious convection during radar DA is a systematic feature of the multi-scale DA system. Assimilation of satellite-retrieved cloud water path (Jones et al. 2014) is a possible way to mitigate this feature in the future. Considered over all scales, both MULTI and LARGE are under-dispersive for the non-precipitation variables throughout the 9 h forecast period. Since the OSSEs contain no model error, this suggests that both MULTI and LARGE IC perturbation methods represent sub-optimal sampling of the IC and forecast errors. One possible remedy for the under-dispersion of non-precipitation forecast variables is to further inflate the IC perturbations after the DA period. However, it is not clear whether this will also improve the ensemble precipitation forecasts. Experimenting with the details of how to best implement such a method is left for future work.

The perfect-model OSSE comparison of MULTI and LARGE for forecasts of mesoscale hourly accumulated precipitation reveals statistically significant skill advantages for MULTI at the 1 h lead time for all thresholds and at several lead times after 4 h for the 2.54 and 6.35 mm h^{-1} thresholds. The storm scale reflectivity forecasts are more skillful for MULTI than LARGE for about 45 (at higher thresholds) to 65 (at lower thresholds) minutes. These results show that the 0000 UTC case of Chapter 4 is more representative of the systematic results than the 2100 UTC case. On average, MULTI therefore represents a more optimal method of generating IC perturbations for SSEFs of mid-latitude convection than LARGE. The comparisons of MULTI48 with LARGE and MULTI with MULTI48 provide further understanding of the reasons for the MULTI advantages over LARGE.

The comparison of MULTI48 with LARGE is generally similar to the comparison of MULTI and LARGE for mesoscale precipitation forecasts at 2.54 and 6.35 mm h^{-1} thresholds and storm-scale reflectivity forecasts with neighborhood radii >4 km. This shows that the smaller magnitude mesoscale IC perturbations in MULTI near convective systems, which are more consistent with the multi-scale analysis uncertainty, systematically account for most of the skill differences between MULTI and LARGE. However, the advantages of MULTI48 over LARGE for mesoscale precipitation at ~5-9 h lead times are less statistically significant than for MULTI. Furthermore, at the 12.7 mm h^{-1} threshold, the mesoscale and small scale IC perturbations have impacts on skill of similar magnitude, showing that the small scale IC perturbations also play an important role. The small scales (i.e., 0-4 km neighborhoods) of the reflectivity forecasts during the first hour are also an exception to the dominance of the mesoscale IC perturbations.

The comparison of MULTI with MULTI48 shows the systematic importance for both short lead time (~1 h) reflectivity forecasts on small scales (i.e., 0-4 km radii) and mesoscale precipitation forecasts out to 9 h. Since Fig. 5.1 and Durran and Gingrich (2014) both suggest that small scale perturbations rapidly develop as a result of downscale energy propagation, the appearance of mesoscale forecast advantages for MULTI, compared to MULTI48, at much later lead times is particularly noteworthy. The systematic difference is likely explained by the impact of the small scale IC perturbations on new convection developing during the early forecast hours and growing upscale during the forecast period, as shown qualitatively in Chapter 4. This shows that optimal multi-scale perturbation design for SSEFs requires appropriate flow

dependent small scale IC perturbations to be explicit added, rather than relying on downscale propagation from the mesoscale IC perturbations.

Chapter 6: Importance of consistency between initial and lateral boundary condition perturbations

6.1 Introduction

The previous two chapters evaluated the direct impacts of differences in IC perturbation methods on ensemble forecast skill. However, it is also of interest to consider the potential impacts of inconsistencies between the IC and LBC perturbations (Caron 2013; Wang et al. 2014).

Caron (2013) introduced a method to alleviate the mismatch between IC and LBC perturbations in limited area models with different perturbation methods for the ICs and LBCs. This was accomplished by blending multi-scale IC perturbations generated on an inner domain with the largest scales of the IC perturbations generated on an outer domain that correspond to the LBC perturbations. The resulting IC and LBC perturbations are consistent with each other on large scales. It was shown that this blending method reduces spurious sea level pressure (SLP) variance that originates at the LBCs and propagates into the interior of the domain. The reduction of spurious SLP variance led to improved forecasts of precipitation and other variables (Caron 2013; Wang et al. 2014). A relatively simple blending method is therefore also applied in this chapter to evaluate the importance of maintaining consistency between the IC and LBC perturbations in the SSEF system used in this study.

Section 6.2 introduces the blending method used to enforce consistency between the IC and LBC perturbations. The impacts on ensemble forecast performance for the

20 May case study, and systematically over all 11 OSSE cases, are presented in Section 6.3. Section 6.4 contains a brief summary of the results.

6.2 Blending method

The blending method of Caron (2013) consists of retaining the high resolution multi-scale IC perturbations at relatively small scales and coarser resolution IC perturbations at large scales, with a gradual transition in the blending zone between the small and large scales. The blending zones for this study are 48-96km (BLEND48) and 192-384km (BLEND192). BLEND192 retains more of the MULTI perturbations while BLEND48 retains more of the LARGE perturbations. The goal is to determine if the advantages of the MULTI IC perturbations can be even further improved by increasing the consistency of the larger scales of such perturbations with the LBC perturbations.

The blending of the LARGE and MULTI IC perturbations is obtained by first applying the two-dimensional discrete cosine transform (DCT; Denis et al. 2002) to both sets of IC perturbations. The DCT is used because it avoids problems related to the non-periodic domain (Denis et al. 2002). For MULTI, the transformed field is fully retained for wavelengths less than λ_L and set to zero for wavelengths greater than λ_U, where λ_L and λ_U are the lower and upper bounds of the blending zone, respectively. The MULTI component at wavelength, λ, between λ_L and λ_U is multiplied by a factor of $1-(\lambda-\lambda_L)/(\lambda_U-\lambda_L)$, resulting in a smooth transition between the retained and truncated spatial scales. For LARGE, the components are multiplied by a factor of 1.0, 0.0 and $(\lambda-\lambda_L)/(\lambda_U-\lambda_L)$ for $\lambda>\lambda_U$, $\lambda<\lambda_L$, and $\lambda_L<\lambda<\lambda_U$, respectively. The MULTI and LARGE transformed fields are then added together and converted back to spatial (rather than

spectral) components using the inverse discrete cosine transform (Denis et al. 2002), resulting in the BLEND IC perturbations. The BLEND IC perturbation spectra follow the MULTI spectra at small scales and the LARGE spectra at large scales, with a gradual transition in the blending zone (not shown).

6.3 Impact of blending on ensemble performance

In this section, the BLEND48 and BLEND192 ensemble forecasts are compared to the MULTI ensemble to determine if further advantages of MULTI can be obtained by enforcing consistency between the IC and LBC perturbations on large scales. Impacts on the storm-scale reflectivity forecasts are minimal (not shown). Therefore the results presented in this chapter focus on the mesoscale hourly accumulated precipitation forecasts. Results from the 20 May case study of Chapter 4 are presented first in Section 6.3a, followed by systematic results in Section 6.3b.

6.3a 20 May 2010 OSSE case study

For the 20 May case study the impact of the IC/LBC inconsistency in the MULTI ensemble is reflected in a temporary increase in SLP variance that lasts for about 1 h in both the 0000 and 2100 UTC cases (Fig. 6.1). While there is little impact on the ensemble mean RMSE for SLP (not shown), the average RMSE of individual ensemble members does show a temporary increase for the MULTI ensemble (also not shown). There is no apparent impact on the other non-precipitation variables (e.g., Fig. 4.2).

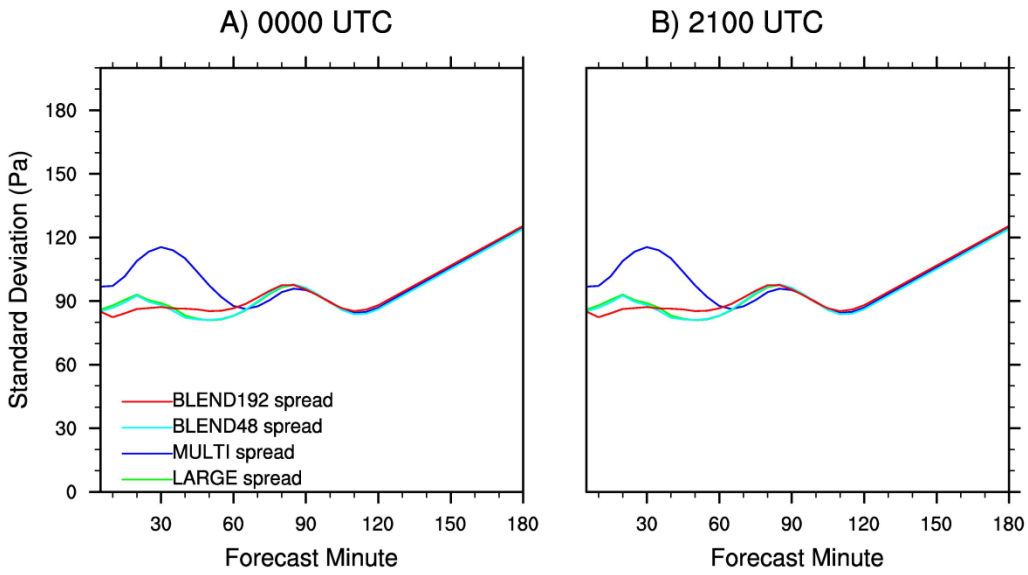

Figure 6.1: Average ensemble standard deviation of sea level pressure, calculated in the same verification domain used for other non-precipitation variables, for (a) the 0000 UTC 20 May case and (b) the 2100 UTC 20 May case.

When compared to the MULTI forecasts, the BLEND192 forecast skill shows very few differences from MULTI for early lead time mesoscale precipitation forecasts (Fig. 6.2) and the short-term reflectivity forecasts (not shown). At later lead times the blending ensembles are generally similar or more skillful than MULTI, indicating an advantage of the greater consistency between the ICs and LBCs for both the 0000 and 2100 UTC cases (Fig. 6.2). While the difference between BLEND192 and BLEND48 depends on the forecast lead time and threshold for the 2100 UTC case (Fig. 5.16e,f), BLEND192 is consistently the most skillful for the 0000 UTC case (Fig. 5.16a,b,c). A blending zone between 384 and 768 km was also evaluated but did not perform quite as well as BLEND192, likely due to retaining the inconsistent (with LBCs) IC perturbations from MULTI on too large of a scale (not shown).

In summary, the inconsistencies between the IC and LBC perturbations in MULTI have little impact on the non-precipitation variables, except for a temporary

increase in ensemble spread of SLP. This may suggest that the IC/LBC inconsistencies are not particularly problematic in this ensemble system. However, a slight increase in mesoscale precipitation forecast skill in BLEND48 and BLEND192 compared to MULTI may be a reflection of a negative impact of the IC/LBC inconsistency on the convective precipitation forecasts for this case.

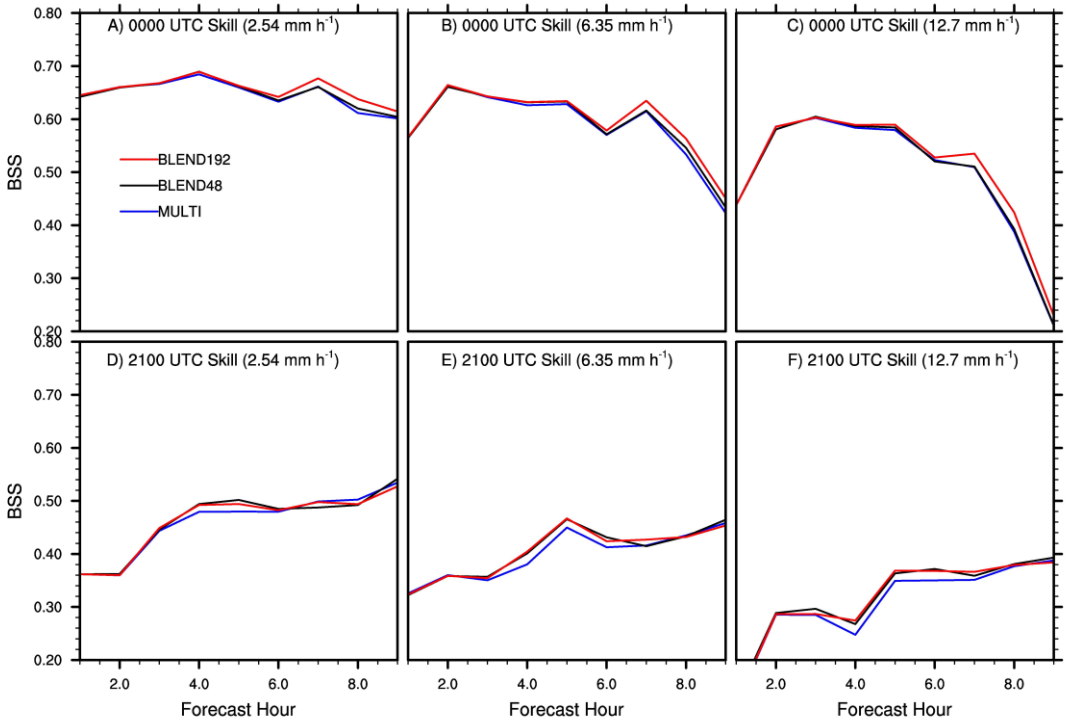

Figure 6.2: As in Fig. 4.3, except for the MULTI, BLEND and BLEND192 ensembles.

6.3b Systematic results from 11 OSSE cases

The systematic impacts of the IC/LBC perturbation inconsistency on non-precipitation variables are very similar to the 20 May case study and therefore not shown. For the mesoscale hourly accumulated precipitation forecasts there is not a clear systematic advantage of either BLEND48 or BLEND192, compared to MULTI (Fig. 6.3). After ~5 h, BLEND48 is significantly less skillful than MULTI (Fig. 6.3)

due to the similarity between the BLEND48 and LARGE IC perturbations. BLEND192 generally shows systematic differences from MULTI that are either very small, statistically insignificant, or both (Fig. 6.3). The inconsistencies between the IC and LBC perturbations in the MULTI ensemble thus do not systematically have a substantial negative impact on the ensemble forecast skill for either precipitation or non-precipitation variables.

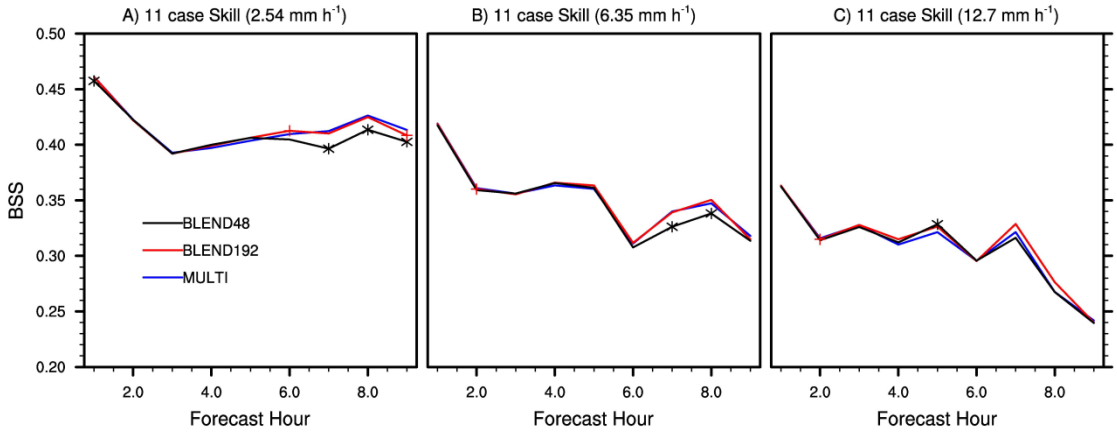

Figure 6.3: As in Fig. 5.3, except for the MULTI, BLEND48 and BLEND192 ensembles. Significant differences from the MULTI line, at the 80% confidence level, are indicated by markers of the same color as the line that is significantly different from MULTI.

6.4 Summary

The importance of the consistency between the IC and LBC perturbations is briefly evaluated in this chapter. Previous studies have demonstrated spurious pressure waves originating at the lateral boundaries resulting from the use of independent perturbation methods with different resolutions for the IC and LBC perturbations (Caron 2013; Wang et al. 2014). In these initial studies, precipitation forecast skill was improved through the use of a blending method to enforce consistency between the IC

and LBC perturbations on the largest scales. A similar method is applied in this chapter for the SSEF system used in this study.

For the 20 May case study, the inconsistency between the IC and LBC perturbations in MULTI leads to rapid growth of SLP spread and error for ~30-60 minutes. However, there are not corresponding impacts on the other non-precipitation variables. For this case, blending the MULTI perturbations with the largest scales of the LARGE IC perturbations (i.e., BLEND192) results in greater IC/LBC consistency and more skillful convective precipitation forecasts, consistent with Caron (2013) and Wang et al. (2014). However, considered over all 11 cases there is not a significant advantage of applying a blending method to enforce greater consistency between the IC and LBC perturbations on large scales for either precipitation or non-precipitation variables.

There are two possible explanations for the difference between this result and the results of Caron (2013). First, in this study the same cycled ensemble DA method is used to generate the inner and outer domain analyses used for IC perturbations. This may lead to less inconsistency between the ICs from the inner domain and the LBCs from the outer domain than using LBCs from a completely different ensemble system. Second, the difference in resolution between the IC and LBC perturbations in MULTI is only a factor of 3 in the present study, compared to a factor of 16 in Caron (2013).

Chapter 7: Impact of IC perturbation methods in real-data experiments

7.1 Introduction

A perfect-model OSSE framework was used in Chapters 4 and 5 to eliminate forecast errors due to model and physics errors and isolate the impacts of sampling the forecast errors resulting from analysis errors with different IC perturbation methods. In a more realistic setting there would also be model and physics errors that should be sampled in the optimal ensemble design. Therefore it is also of interest to extend the study to corresponding real data cases (i.e., abandoning the perfect-model OSSE framework). This allows the implications for optimal design of SSEF IC perturbations in more realistic operational settings to be assessed. Model and physics errors may play a more prominent role relative to IC errors in such settings (e.g., Stensrud et al. 2000; Clark et al. 2008; Johnson et al. 2011b).

The same IC perturbation experiments are repeated for the same ten cases using real observations to test the hypothesis that more optimally sampling the IC perturbations in a real data (i.e., not OSSE) setting will also lead to improved ensemble forecast performance. The main difference between the real-data and OSSE cases is that only the real-data forecasts contain model and physics errors.

The real-data experiments are conducted by performing the same IC perturbation experiments using the analysis ensembles from the real-data cases in Chapter 2. Since real-data analyses were not generated for the 2100 UTC 19 May case, only 10 cases are used for the real-data results. For the real-data cases, the precipitation

forecasts are verified against radar derived quantitative precipitation estimates from the National Severe Storms Laboratory (NSSL) Q2 product (Zhang, J. et al. 2011). In contrast to the single model level evaluation of reflectivity in the OSSEs, composite reflectivity (i.e., maximum reflectivity in the vertical column) is verified against the NSSL composite reflectivity mosaic (Zhang, J. et al. 2011). This provides a direct comparison between the forecast and observed variables. The results are also very similar when model level 5 (~900 mb) reflectivity is verified against the NSSL Hybrid Scan Reflectivity product (i.e., reflectivity at the lowest available scan level). The NSSL composite reflectivity mosaic on a uniform 0.01° (~1 km) grid is interpolated to the WRF grid using bilinear interpolation. The same verification domains as the OSSEs are used for the real-data hourly accumulated precipitation verification. For the real-data reflectivity verification, the smaller verification domains are relocated and resized to better capture the locations of the actual mesoscale precipitation systems which are different than the Nature run locations in many cases. There are three fewer mesoscale precipitation systems in the real-data experiments, resulting in a total of 15 unique verification domains for the real-data reflectivity verification. Multiple verification domains on the same case are not considered independent samples for the statistical significance tests.

7.2 Results

7.2a Mesoscale hourly accumulated precipitation

Unlike the OSSE results for mesoscale precipitation forecasts, MULTI and MULTI48 generally result in less skillful forecasts than LARGE in the real data

experiments (Fig. 7.1). The difference between MULTI or MULTI48 (or both) and LARGE is statistically significant at many lead times, especially at the 6.35 and 12.7 mm h^{-1} thresholds (Fig. 7.1). The contrast with the OSSE results is attributed to the presence of model error in the real-data experiments. Although the real-data forecasts contain model error, all ensemble members still use the same model configuration. Therefore, even if the IC perturbations perfectly sample the IC uncertainty, the ensemble forecasts are expected to be under-dispersive as a result of the un-represented model errors. As demonstrated in Chapters 4 and 5, the LARGE OSSE ensembles have larger magnitude mesoscale IC perturbations which degrade the ensemble forecast skill compared to MULTI. Greater precipitation forecast spread for LARGE than MULTI is also seen in the real-data experiments, as reflected in the lower correspondence ratio (CR) for LARGE for the representative 6.35 mm h^{-1} threshold (Stensrud and Wandishin 2000) in Fig. 7.2. The CR is the number of grid points at which N members (N=20 in Fig. 7.2) forecast the threshold to be exceeded, divided by the number of grid points where any member forecasts the threshold to be exceeded. Smaller values of CR indicate greater spread since a CR of 1.0 indicates no diversity of the forecasts and a CR of 0.0 indicates no overlap of the forecasts. Unlike the OSSEs, the greater LARGE spread compensates for un-represented model errors in the real-data experiments, leading to better ensemble forecast skill than MULTI. Optimal IC perturbation methods for SSEFs may therefore only show their full advantage as part of a holistic approach to ensemble design that also includes model and physics diversity.

Although LARGE is more skillful than MULTI and MULTI48 for mesoscale precipitation forecasts in the real-data experiments, there may still be an advantage of

the small scale IC perturbations resolved only by MULTI (Fig. 7.1). Small but statistically significant positive impacts of the small scale IC perturbations (i.e., MULTI vs MULTI48; significance markers along bottom axis of Fig. 7.1) are seen at 3 lead times each for the 6.35 and 12.7 mm h^{-1} thresholds. However, like the OSSE results, the differences between LARGE and MULTI skill are mostly explained by the differences between LARGE and MULTI48.

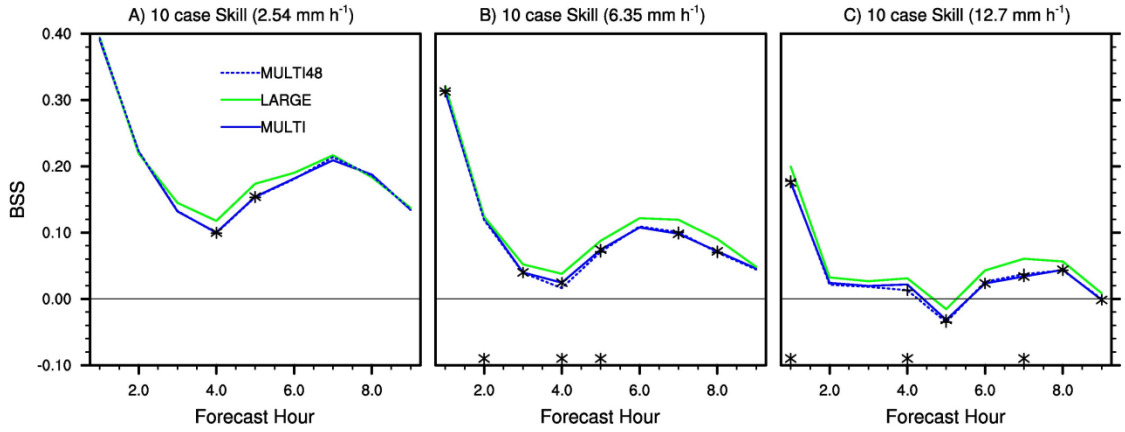

Figure 7.1: As in Fig. 5.3, except for the 10 real-data cases.

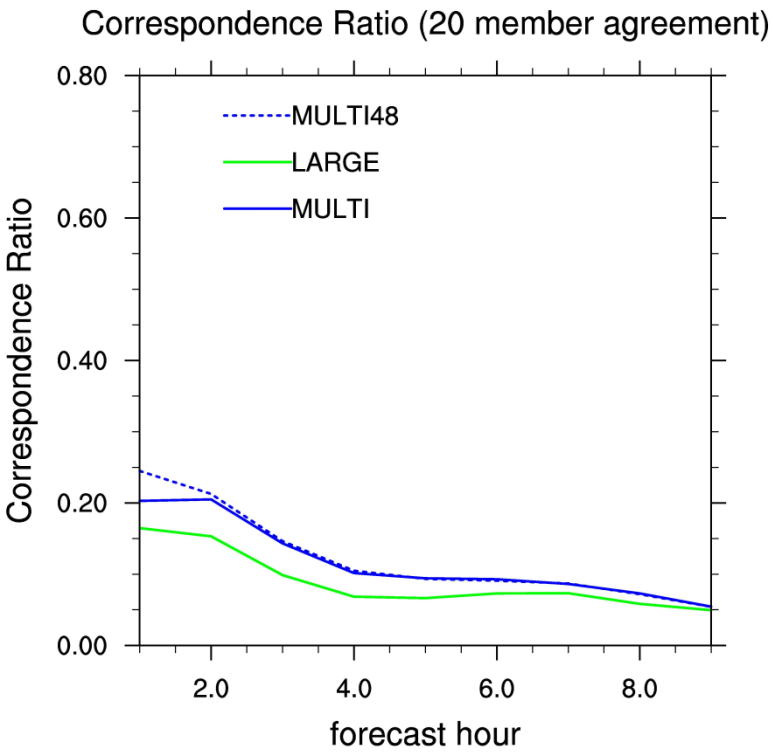

Figure 7.2: Correspondence ratio with 20 member agreement (see text for explanation) averaged over all 10 real-data cases of hourly accumulated precipitation at the 6.35 mm h^{-1} threshold.

7.2b Storm-scale reflectivity

The storm-scale reflectivity forecasts for the real-data experiments show mixed results with MULTI significantly more skillful than LARGE during the first ~20 minutes. After this time, a LARGE advantage begins to appear for the smallest neighborhood radii (Fig. 7.3). The MULTI advantage becomes increasingly limited to larger radii and smaller thresholds until only the LARGE advantage is present at the end of the 2 h forecast period (Fig. 7.3). Subjective evaluation reveals that, like the OSSE cases, the MULTI forecasts of the MCSs are much sharper than the LARGE forecasts. For example, the 20 May real-data forecast shows lower NEP for MULTI where the LARGE NEP is relatively low and higher NEP for MULTI where the LARGE NEP is

relatively high (e.g., Fig. 7.4a-f). However, the real-data cases show more pronounced differences between the forecast and observed MCSs than the OSSE cases, which is likely a result of model errors (e.g., Fig. 7.4a-c). Since such model errors grow faster on smaller scales, the sharper MULTI forecast is advantageous only for very short lead times on small scales and slightly longer lead times on larger scales (Fig. 7.3). Otherwise, the less-sharp LARGE forecast is more skillful (Fig. 7.3).

The MULTI advantage during the first 20 minutes suggests that the sharper MULTI forecast is consistent with the forecast uncertainty associated with the IC errors. The rapid loss of this skill advantage during the forecast period suggests that it is the growth of spread during the forecast period that is lacking in MULTI. Since the perfect-model OSSEs do not show the same result (Fig. 5.4), the insufficient spread growth in MULTI is attributed to the unrepresented model and physics errors. This further emphasizes that effective model and physics perturbations are essential to obtain the full advantage of optimal IC perturbation methods.

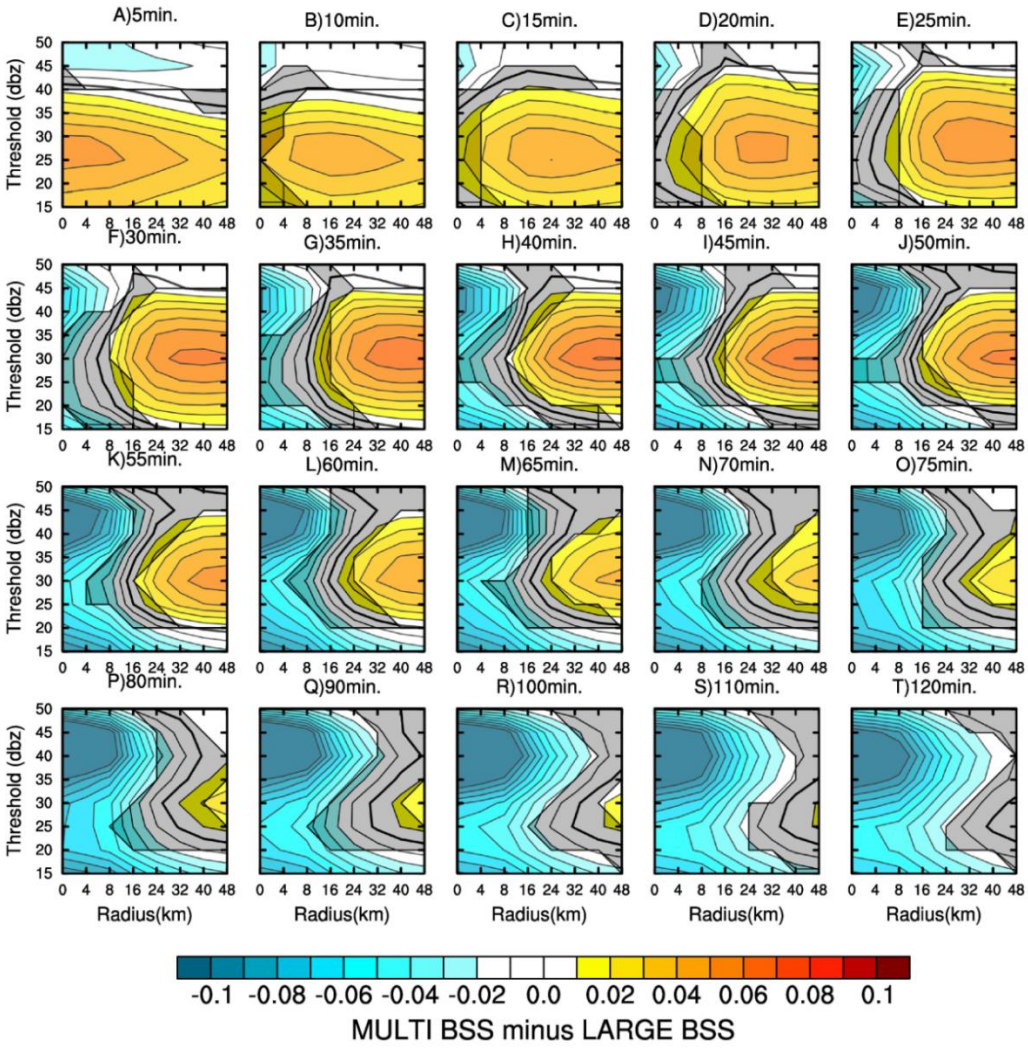

Figure 7.3: As in Fig. 5.4, except for the real-data cases.

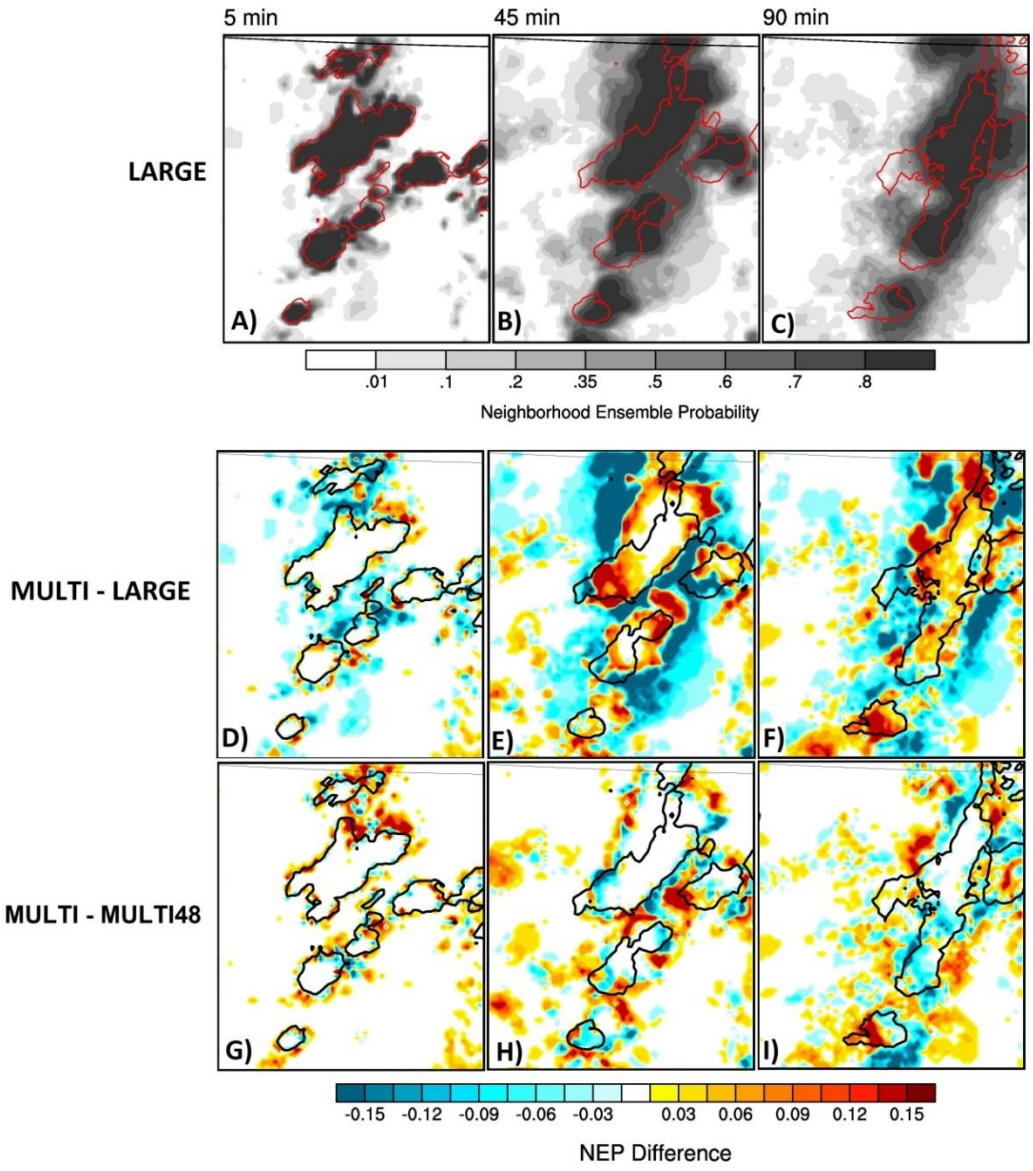

Figure 7.4: Example of real-data reflectivity forecasts, initialized at 0000 UTC 20 May 2010, for the 30 dBZ threshold. The NEP of the LARGE ensemble at 5, 45, and 90 minute lead times is plotted in panel (a), (b) and (c), respectively, with the observation contour overlaid in red. Panels (d), (e) and (f) show the difference between the MULTI and LARGE NEP at the same times, with the observation contour overlaid in black. Panels (g)-(i) are as in (d)-(f), except for the difference between MULTI48 and MULTI NEP.

Like the mesoscale precipitation forecasts, the storm-scale reflectivity forecasts show an advantage of the small scale IC perturbations in the real-data experiments. In

the real-data experiments, the small scale IC perturbations lead to significant skill advantages for small neighborhood radii at early lead times, growing upscale into significant advantages at even larger radii with increasing lead time (Fig. 7.5). In contrast to the OSSEs, the skill advantages resulting from the small scale IC perturbations persist throughout the 2 h forecast period (Fig. 7.5). This may also be due to the more pronounced differences between the forecast and observed MCSs that result from model error. In Chapter 4 it was noted that the main advantage for reflectivity forecasts of the small scale IC perturbations in the OSSEs is a smoothing of the NEP forecasts where strong NEP gradients do not correspond to the small scale features in the observation contour. Subjective evaluation of the real-data cases reveals even more smoothing of the NEP forecasts on even larger scales. For example, the 20 May real-data forecast shows higher NEP for MULTI where MULTI48 NEP is relatively low and lower NEP for MULTI where MULTI48 NEP is relatively high (e.g., Fig. 7.4g-i). This shows that the flow-dependent small scale IC perturbations are even more important for the storm scale reflectivity forecasts in an under-dispersive SSEF system that does not adequately sample the model and physics errors.

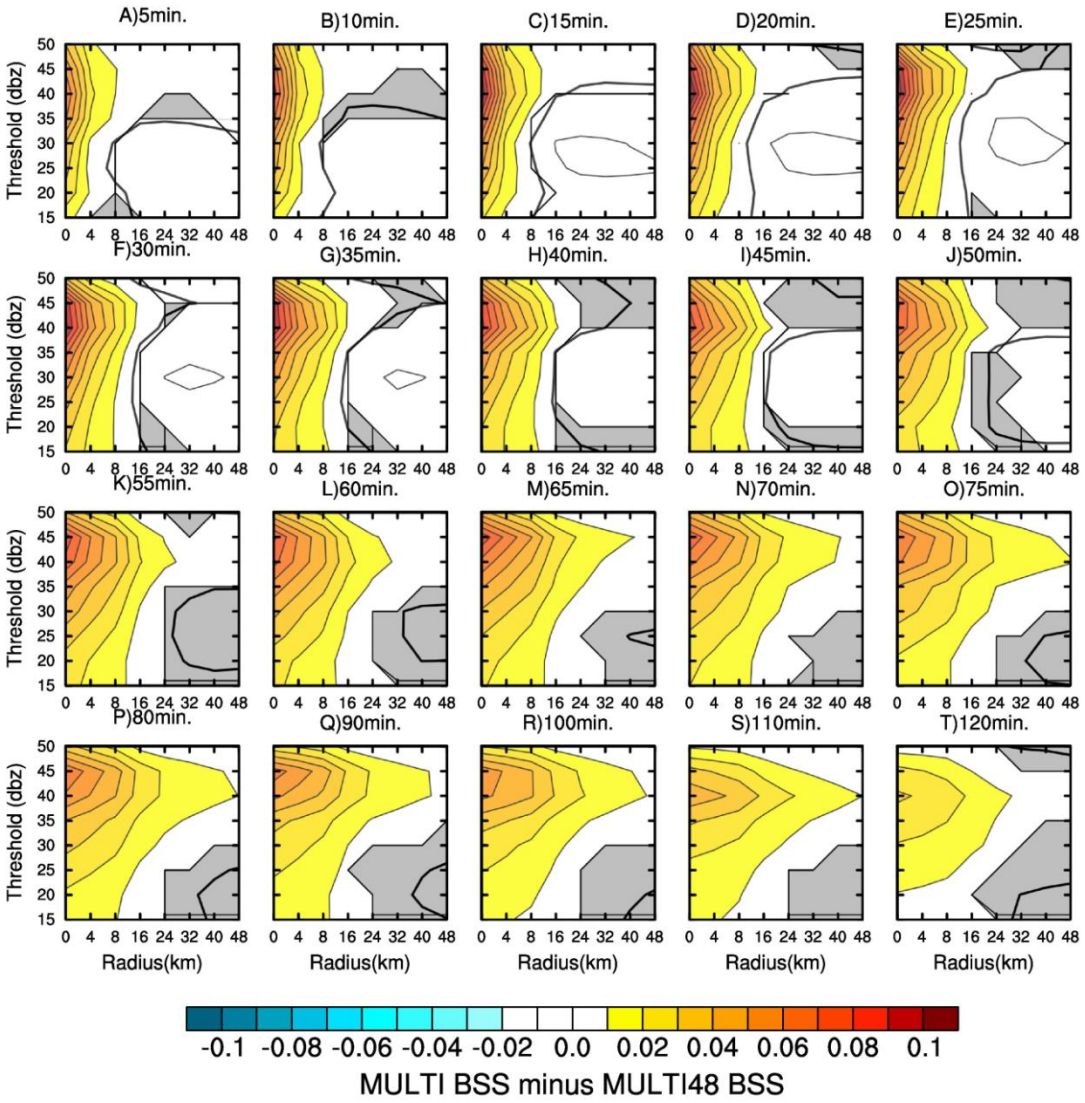

Figure 7.5: As in Fig. 5.7, except for the real-data cases.

7.3 Summary

Whereas the perfect-model OSSEs of Chapters 4 and 5 allow the impact of IC uncertainty and IC perturbations to be isolated, the real-data experiments in the present chapter address the importance of model and physics uncertainty. The impacts of the IC perturbation method in the real-data experiments are similar in some ways to the OSSE results. The LARGE IC perturbations generally have larger magnitude near ongoing convection than the MULTI IC perturbations. However, in contrast to the

OSSE results the LARGE skill is greater than the MULTI skill in the real-data experiments. The difference is hypothesized to be a result of unrepresented model errors in the real-data experiments. While the LARGE mesoscale IC perturbations are too large in magnitude for the actual IC uncertainty they can compensate for the unrepresented model errors, improving the forecast skill. Like the OSSEs, the small scale component of the MULTI IC perturbations has a small positive impact on forecast skill even for mesoscale hourly accumulated precipitation forecasts beyond the first few hours of forecast lead time. For reflectivity forecasts, the small scale IC perturbations are even more important for the real-data experiments with unrepresented model errors than for the perfect-model OSSEs.

The real-data experiments show that the impact on ensemble forecast performance of the IC perturbation method should not be considered in isolation from other aspects of the ensemble design. In particular, if model and physics errors are not adequately sampled then unrealistic IC perturbations downscaled from a coarser ensemble may outperform multi-scale IC perturbations that are more consistent with the analysis uncertainty. This is similar to the concept of compensating biases in numerical modeling, whereby improving just one aspect of the model may actually degrade the overall performance (e.g., Neggers and Siebesma 2013). Therefore, model and physics diversity in a SSEF system should also be carefully configured and optimized in order to obtain the full advantages of optimal IC perturbation methods. Systematic controlled experiments on the optimal sampling of model and physics error in SSEFs for mid-latitude convection forecasting are suggested for future work.

Chapter 8: Summary and Conclusions

Despite widespread use of storm-scale ensemble forecast (SSEF) systems in research settings and limited but increasing interest in operational settings, there has been relatively little systematic study of how to optimally design such systems. The design of SSEFs is distinguished from the design of coarser global and mesoscale ensembles in that SSEFs resolve a much broader range of spatial scales, including the rapid non-linear error growth associated with moist convection. Initial studies have investigated aspects of SSEF design such as the optimal ensemble size (Clark et al. 2011) and the impacts of different sources of ensemble diversity on forecast spread (Clark et al. 2010; Johnson et al. 2011b). This dissertation builds on such early studies by investigating the impacts of different multi-scale data assimilation (DA) and initial condition (IC) perturbation methods, with the goal of moving closer to an understanding of the optimal design of SSEF systems.

8.1 Data assimilation methods

An accurate IC analysis is a pre-requisite for an accurate forecast. Such an analysis is typically provided by a DA system. Since convection-permitting forecasts can be strongly influenced by features ranging from synoptic to convective scales, the DA system should also be able to accurately analyze features on very different scales. In the first part of this study, the GSI-based EnKF and 3DVar DA system is extended to assimilate radar reflectivity using the WRF model with WSM6 microphysics for multi-scale DA where both the model and observations resolve synoptic to convective scale features. The EnKF and 3DVar components of the GSI-based system are then

systematically compared to each other in the multi-scale context to understand how the differences between the two techniques affect the analysis quality on multiple scales and the subsequent precipitation forecast skill.

The multi-scale analyses from GSI-based EnKF systematically lead to significantly more skillful forecasts than the GSI-based 3DVar analyses. The more skillful EnKF-initialized forecasts result from more accurate analyses of both the mesoscale environment and the convective scale features. A case study of an upscale growing MCS is used to better understand the causes of such skill differences. The more accurate mesoscale environment with EnKF is attributed to the flow dependent shape and spatial scale of the ensemble-based background error covariance. The more accurate convective scale analysis for EnKF is attributed to the presence of coherent cross-variable correlations in the ensemble-based background error covariance for reflectivity DA. The forecast skill improvement resulting from the radar DA lasts for ~5h using EnKF but only ~1h using 3DVar, further showing the advantage of the GSI-based EnKF over 3DVar. The more skillful forecasts initialized from EnKF analyses are primarily due to the more accurate convective scale analysis at early lead times and increasingly due to the more accurate mesoscale environment at later lead times. This emphasizes the need for the multi-scale approach to DA for convective precipitation forecasting.

While the GSI-based EnKF generally outperforms the GSI-based 3DVar, 3DVar does have some useful features for radar DA such as a more rapid spin up of convective features and the ability to add reflectivity where no reflectivity is present in the first guess forecast. These differences occur because realistic flow-dependent error

covariance structures take some time to spin up in EnKF and the weight given to observation information in EnKF depends on the first guess ensemble variance which is zero if no members forecast any reflectivity. Further study is needed to develop more effective methods of defining the static background error covariance for 3DVar and to combine the advantages of EnKF and 3DVar using hybrid ensemble-variational methods.

8.2 Initial condition perturbation methods

The GSI-based DA system can provide both the ensemble mean analysis and the IC perturbations for a SSEF system. In the second part of this study, a series of experiments with increasing complexity are conducted to better understand the impacts on SSEF performance of different methods of generating the IC perturbations. First, the forecast sensitivity to simple random homogeneous IC perturbation methods is systematically evaluated on different spatial scales. Perfect model Observation System Simulation Experiments (OSSEs) are then conducted using a case study to qualitatively understand the impacts on forecast skill of more realistic flow-dependent multi-scale IC perturbations. Finally, the OSSE case study is extended to 11 diverse cases to obtain systematically robust objective results. The study is also extended to consider the IC/LBC perturbation consistency and real-data (non-OSSE) experiments to better understand the implications for practical applications which contain model error.

Convective precipitation forecasts on medium to large scales are more sensitive to large scale IC and physics perturbations than to random homogeneous small scale IC perturbations. The forecasts on small scales are similarly sensitive to all of the

considered IC perturbation methods. On medium forecast scales, the small scale IC perturbations generate at least half as much perturbation energy as the larger scale IC and physics perturbations. This shows the potential importance of small scale IC perturbations for convective precipitation forecasting. However, the forecasts are not significantly more sensitive to small scale IC perturbations together with large scale IC and physics perturbations than to the large scale IC and physics perturbations alone. This may be a result of the lack of flow dependence in the simple random homogeneous small scale IC perturbations.

The GSI-based EnKF multi-scale analyses provide an opportunity to evaluate more complex flow-dependent multi-scale IC perturbations. A perfect model OSSE framework is used to neglect model errors and focus only on sampling the multi-scale analysis uncertainty using multi-scale IC perturbations. The multi-scale IC perturbations (i.e., MULTI) contain smaller scale features than are resolved by IC perturbations downscaled from a coarser resolution mesoscale ensemble (i.e., LARGE). There are also differences between the MULTI and LARGE IC perturbations on the commonly resolved mesoscales as a result of the different methods of generating the perturbations.

A case study of an upscale growing MCS provides qualitative understanding of how the differences in IC perturbation methods can affect the ensemble forecast skill. The small scale IC perturbations, resolved only in MULTI, lead to more skillful forecasts of storm-scale reflectivity in 0-4 km neighborhoods at lead times up to about one hour. This is a direct result of the greater ensemble diversity for features on such scales. The small scale IC perturbations are also shown to be capable of upscale

impacts on the hourly accumulated precipitation forecasts out to 9 h lead times. This is a result of the impact of the small scale IC perturbations on convection during the early forecast period that newly develops from small scale features and grows upscale during the forecast period. However, for the single case study a clear skill advantage resulting from such upscale impacts is not seen. Forecasts of hourly accumulated precipitation in mesoscale (48 km) neighborhoods and storm-scale reflectivity in 8-48 km neighborhoods have skill dominated by the mesoscale differences between the MULTI and LARGE IC perturbations. The MULTI IC perturbations have a mesoscale component that is more consistent with the analysis errors than the LARGE IC perturbations near the analyzed MCS, leading to forecast advantages for MULTI. However, mesoscale mid-level moisture anomalies appear in the MULTI IC perturbations away from the observed MCS, apparently due to insufficient suppression of spurious convection during the radar DA period. For forecasts initialized at 0000 UTC, the more reasonable mesoscale IC perturbations near the developing MCS dominate and lead to an overall skill advantage for MULTI. For forecasts initialized at 2100 UTC, the mesoscale mid-level moisture perturbations enhance spurious convection in MULTI, leading to an overall forecast disadvantage for MULTI.

Given the case dependence of some of the OSSE results, the OSSE case study is extended to 11 diverse cases to obtain more robust systematic conclusions. The 0000 UTC case study is more representative of the systematic results than the 2100 UTC case study since MULTI generally shows forecast advantages over LARGE. Hourly accumulated precipitation forecasts are systematically more skillful for MULTI than LARGE at the 1 h and ~5-9 h lead times, with the statistically insignificant exception of

the 12.7 mm h^{-1} threshold at 8-9 h. The skill advantage primarily results from the greater consistency near analyzed convective systems between the IC uncertainty and the mesoscale component of the MULTI IC perturbations, compared to LARGE. The smaller scales of IC perturbation that are resolved by MULTI also grow upscale into significant hourly accumulated precipitation forecast advantages after the 5 h lead time. Storm scale reflectivity forecasts in neighborhoods larger than 4 km are significantly more skillful for MULTI than LARGE during the first ~1 h as a result of the better mesoscale IC perturbations for MULTI. The small scale IC perturbations in MULTI also lead to significant skill advantages for the reflectivity forecasts in 0-4 km neighborhoods during the first hour. During this time, sufficient small scale spread has not yet been generated by downscale energy propagation from the larger scale perturbations. These results show that the method of generating multi-scale IC perturbations directly on the convection-permitting grid using multi-scale ensemble-based DA is indeed more optimal than downscaling larger scale perturbations for the purpose of SSEFs of mid-latitude convection. While most of the advantage comes from the method of generating the mesoscale component of the IC perturbations, which is more consistent with the analysis uncertainty, the small scale IC perturbations also contribute to statistically significant forecast advantages.

The OSSE cases are also used to evaluate the importance of the consistency between IC and LBC perturbations. The use of independent perturbations of different resolution for the IC and LBC perturbations results in spurious pressure waves at early lead times. Caron (2013) demonstrated a method of improving the consistency between IC and LBC perturbations by blending multi-scale IC perturbations with the largest

scales of coarser perturbations that are consistent with LBC perturbations. This method can lead to improved precipitation forecast skill (Caron 2013; Wang et al. 2014). However, in the SSEF system used in this study the inconsistency between the IC and LBC perturbations is not sufficiently problematic that a similar blending method leads to systematically improved precipitation forecast skill. This may be a result of using the same multi-scale DA system to generate the mesoscale and multi-scale perturbations used for the LBCs and ICs, respectively, and the relatively small difference in resolution between the inner and outer domain.

The OSSE results show that the MULTI IC perturbations more optimally sample the forecast uncertainty resulting from IC errors than the LARGE IC perturbations. However, in real-data scenarios model and physics errors also contribute to the forecast uncertainty. In real-data experiments it is shown that the LARGE ensemble is significantly more skillful than the MULTI ensemble. Although the mesoscale component of the LARGE IC perturbations has too much magnitude for the actual IC uncertainty, this extra ensemble spread compensates for the lack of model and physics diversity in the experiment design. Such diversity is not added in these experiments for two reasons. First, it would complicate the clean comparison between the real-data and OSSE results. Second, optimal methods of sampling the model and physics errors are also not yet known. It is concluded from the real-data experiments that additional future research on the optimal model and physics perturbation design is worthwhile and that such perturbations are needed to realize the advantage of using more optimal IC perturbation methods.

8.3 Summary and future work

In summary, the sensitivity of convective precipitation forecasts to the analysis of both storm scale features and the mesoscale environment shows the importance of a multi-scale approach to data assimilation for convection-permitting forecasts. The newly extended GSI-based EnKF is also shown to be an effective technique for such multi-scale DA, in comparison to the GSI-based 3DVar. The GSI-based EnKF has the further advantage of providing flow-dependent multi-scale IC perturbations for SSEFs. In comparison to IC perturbations downscaled from a mesoscale ensemble, the mesoscale component of the multi-scale IC perturbations is more consistent with the analysis uncertainty which leads to significantly improved forecast skill. Early studies (Durran and Gingrich 2014 and Chapter 3) suggested that the further addition of small scale IC perturbations in the multi-scale IC perturbation design may not be important. This hypothesis is rejected in the present study. The flow-dependent small scale IC perturbations generated by cycled multi-scale ensemble-based DA lead to significant forecast advantages for both storm-scale short lead time reflectivity forecasts and mesoscale hourly accumulated precipitation forecasts out to 9 h lead time.

Much future work on the optimal DA and IC perturbation methods for SSEFs of mid-latitude convection is still needed. Assimilation of cloud water path is suggested as a potential method of further reducing the negative impacts of spurious convection during radar DA. Application of hybrid ensemble-variational methods may further improve the storm-scale component of the multi-scale DA system by including the advantages of both the variational and ensemble-based frameworks (e.g., Wang et al. 2008a,b). The variational part of such a hybrid system could also be further improved

by continued research on methods to define the static background error covariance for reflectivity DA. Systematic research on how to optimally sample the model and physics errors in SSEF design is expected to both improve SSEF skill in real-data scenarios and reveal the full advantage of optimal IC perturbations.

References

Aksoy, A., D. C. Dowell, C. Snyder, 2009: A multicase comparative assessment of the ensemble Kalman filter for assimilation of radar observations. Part I: Storm-scale analyses. *Mon. Wea. Rev*., **137**, 1805–1824.

Aksoy, A., D. C. Dowell, C. Snyder, 2010: A multicase comparative assessment of the ensemble Kalman filter for assimilation of radar observations. Part II: Short-range ensemble forecasts. *Mon. Wea. Rev*., **138**, 1273–1292.

Benjamin, S. G., G. A. Grell, J. M. Brown, T. G. Smirnova, and R. Bleck, 2004: Mesoscale weather prediction with the RUC hybrid isentropic-terrain-following coordinate model. *Mon. Wea. Rev*., **132**, 473–494.

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

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

Brooks, H. E., J. W. Lee, and J. P. Craven, 2003: The spatial distribution of severe thunderstorm and tornado environments from global reanalysis data. *Atmos. Res*., **67 & 68**, 73–94.

Brousseau P., Desroziers G., Bouttier F. and Chapnik B. 2014: A posteriori diagnostics of the impact of observations on the AROME-France convective-scale data-assimilation system. *Quart. J. Roy. Meteor. Soc*., **140**, 982 – 994.

Buehner, M., 2005: Ensemble-derived stationary and flow-dependent background-error covariances: evaluation in a quasi-operational NWP setting. *Quart. J. Roy. Meteor. Soc*., **131**, 1013-1043.

Buehner, M., P. L. Houtekamer, C. Charette, H. L. Mitchell, B. He, 2010: Intercomparison of variational data assimilation and the ensemble Kalman filter for global deterministic NWP. Part I: Description and single-observation experiments. *Mon. Wea. Rev*., **138**, 1550–1566.

Buizza, R., T. N. Palmer, 1995: The singular-vector structure of the atmospheric global circulation. *J. Atmos. Sci*., **52**, 1434–1456.

Carley, J.R., 2012: Hybrid ensemble-3DVar radar data assimilation for the short-term prediction of convective storms. Ph.D. Dissertation, Department of Earth, Atmospheric, and Planetary Sciences, Purdue University, 205 pp.

Caron, J.-F., 2013: Mismatching perturbations at the lateral boundaries in limited-area ensemble forecasting: A case study. *Mon. Wea. Rev*., **141**, 356–374.

Casati, B., G. Ross, and D. B. Stephenson, 2004: A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteor. Appl*., **11**, 141-154.

Caya, A., J. Sun, C. Snyder, 2005: A Comparison between the 4DVAR and the ensemble Kalman filter techniques for radar data assimilation. *Mon. Wea. Rev*., **133**, 3081–3094.

Chang, W., K.-S. Chung, L. Fillion, S.-J. Baek, 2014: Radar data assimilation in the Canadian high-resolution ensemble Kalman filter system: Performance and verification with real summer cases. *Mon. Wea. Rev*., **142**, 2118–2138.

Cintineo, R. M. and D. J. Stensrud, 2013: On the predictability of supercell thunderstorm evolution. *J. Atmos. Sci*., **70**, 1993–2011.

Clark, A. J., W. A. Gallus Jr., and T.-C. Chen, 2007: Comparison of the diurnal precipitation cycle in convection-resolving and non-convection-resolving mesoscale models. *Mon. Wea. Rev*., **135**, 3456–3473.

Clark, A. J., W. A. Gallus Jr., 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 Jr., 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.

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.

Coniglio, M. C., J. Y. Hwang, D. J. Stensrud, 2010: Environmental factors in the upscale growth and longevity of MCSs derived from Rapid Update Cycle analyses. *Mon. Wea. Rev*., **138**, 3514-3539.

Crook, N. A., and J. Sun, 2002: Assimilating radar, surface and profiler data for the Sydney 2000 forecast demonstration project. *J. Atmos. Oceanic Technol*., **19**, 888–898.

Denis, B., J. Côté, and R. Laprise, 2002: Spectral decomposition of two-dimensional atmospheric fields on limited-area domains using the Discrete Cosine Transform (DCT). *Mon. Wea. Rev*., **130**, 1812–1829.

Dey, S. R. A., G. Leoncini, N. M. Roberts, R. S. Plant, and S. Migliorini, 2014: A spatial view of ensemble spread in convection permitting ensembles. *Mon. Wea. Rev*., **142**, 4091–4107.

Dixon, M., Z. Li, H. Lean, N. Roberts, and S.Ballard, 2009: Impact of data assimilation on forecasting convection over the United Kingdom using a high-resolution version of the Met Office Unified Model. *Mon. Wea. Rev*., **137**, 1562-1584.

Done, J. M., Craig, G. C., Gray, S. L. and Clark, P. A. (2012), Case-to-case variability of predictability of deep convection in a mesoscale model. *Quart. J. R. Meteor. Soc*., **138**: 638–648.

Dowell, D., F. Zhang, L. J. Wicker, C. Snyder, and N. A. Crook, 2004: Wind and temperature retrievals in the 17 May 1981 Arcadia, Oklahoma supercell: Ensemble Kalman filter experiments. *Mon. Wea. Rev*., **132**, 1982-2005.

Dowell, D. C., L. J. Wicker, C. Snyder, 2011: Ensemble Kalman filter assimilation of radar observations of the 8 May 2003 Oklahoma City supercell: Influences of reflectivity observations on storm-scale analyses. *Mon. Wea. Rev*., **139**, 272–294.

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

Du, J., and Coauthors, 2009: NCEP Short-Range Ensemble Forecast (SREF) system upgrade in 2009. Extended Abstracts, *19th Conf. on Numerical Weather Prediction and 23rd Conf. on Weather Analysis and Forecasting*, Omaha, NE, Amer. Meteor. Soc., 4A.4. [Available online at http://ams.confex.com/ams/23WAF19NWP/techprogram/paper_153264.htm].

Duc, L., K. Saito, H. Seko, 2013: Spatial-temporal fractions verification for high-resolution ensemble forecasts. *Tellus*, **65A**, 18171.

Durran, D. R. and M. Gingrich, 2014: Atmospheric predictability: Why butterflies are not of practical importance. *J. Atmos. Sci*., **71**, 2476–2488.

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

Ehrendorfer, M., 1997: Predicting the uncertainty of numerical weather forecasts: A review. *Meteorol. Zeitschrift*, **6**, 147-183.

Ek, M. B., K. E. Mitchell, Y. Lin, P. Grunmann, E. Rogers, G. Gayno, and V. Koren, 2003: Implementation of the upgraded Noah land-surface model in the NCEP operational mesoscale Eta model. *J. Geophys. Res*., **108**, 8851.

Elmore, K. L., D. J. Stensrud, K. C. Crawford, 2002: Explicit cloud-scale models for operational forecasts: A note of caution. *Wea. Forecasting*, **17**, 873–884.

Evensen, G., 2003: The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dyn*., **53**, 343–367.

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 the NCEP Eta model. Preprints, *15th Conf. on Numerical Weather Prediction*, San Antonio,TX, Amer. Meteor. Soc., 280-283.

Ferrier, B. S., W.Wang, and E. Colon, 2011: Evaluating cloud microphysics schemes in nested NMMB forecasts. *20th Conf. on Numerical Weather Prediction*, Seattle, WA, Amer. Meteor. Soc.

Fritsch, J. M., R. E. Carbone, 2004: Improving quantitative precipitation forecasts in the warm season: A USWRP research and development strategy. *Bull. Amer. Meteor. Soc*., **85**, 955–965.

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. Ocean. Tech*., **21**, 457-469.

Gao, J., D. J. Stensrud, 2012: Assimilation of reflectivity data in a convective-scale, cycled 3DVAR framework with hydrometeor classification. *J. Atmos. Sci*., **69**, 1054–1065.

Gaspari, G. and Cohn, S. E., 1999: Construction of correlation functions in two and three dimensions. *Q.J.R. Meteorol. Soc*., **125**: 723–757.

Grell, G. A., and D. Dévényi, 2002: A generalized approach to parameterizing convection combining ensemble and data assimilation techniques. *Geophys. Res. Lett*., **29**, 1693, doi:10.1029/2002GL015311.

Grimit, E. P., C. F. Mass, 2002: Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest. *Wea. Forecasting*, **17**, 192–205.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167.

Hamill, T. M., 2006: Ensemble based atmospheric data assimilation. Predictability of Weather and Climate, R. Hagedorn and T. N. Palmer, Eds., Cambridge Press, 124–156.

Hamill, T. M., J. S. Whitaker, D. T. Kleist, M. Fiorino, and S. G. Benjamin, 2011: Predictions of 2010's tropical cyclones using the GFS and ensemble-based data assimilation methods. *Mon. Wea. Rev*., **139**, 3243–3247.

Harnisch, F., 2014: Initial ensemble perturbations provided by convective-scale ensemble data assimilation. *World Weather Open Science Conference*, Montreal, QC.

Hohenegger, C., D. Lüthi, C. Schär, 2006: Predictability mysteries in cloud-resolving models. *Mon. Wea. Rev*., **134**, 2095–2107.

Hohenegger, C., C. Schär, 2007a: Predictability and error growth dynamics in cloud-resolving models. *J. Atmos. Sci*., **64**, 4467–4478.

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

Hohenegger, C., A. Walser, W. Langhans, C. Schär, 2008: Cloud-resolving ensemble simulations of the August 2005 Alpine flood. *Q. J. R. Meteorol. Soc*., **134**, 889-904.

Hong, S.-Y. and J.-O. J. Lim, 2006: The WRF single-moment 6-class microphysics scheme (WSM6). *J. Korean Meteorol. Soc*., vol. **42**, no. 2, 129–151.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, 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: 3DVAR and 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.

Iacono, M. J., J. S. Delamere, E. J. Mlawer, M. W. Shephard, S. A. Clough, and W. D. Collins, 2008: Radiative forcing by long-lived greenhouse gases: Calculations with the AER radiative transfer models. *J. Geophys. Res*., **113**, D13103, doi:10.1029/2008JD009944.

Janjic´, Z. I., 1994: The step-mountain eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. *Mon. Wea. Rev*., **122**, 927–945.

Janjic´, Z. I., 2001: Nonsingular implementation of the Mellor-Yamada level 2.5 scheme in the NCEP meso model. NCEP Office Note, 91 pp. [437].

Jing, Z., G. Wiener, 1993: Two-dimensional dealiasing of doppler velocities. *J. Atmos. Oceanic Technol*., **10**, 798–808.

Johnson, A., 2011: Clustering analysis, verification and calibration of a convection-allowing ensemble forecast system. School of Meteorology M.S. Thesis, University of Oklahoma, 178 pp.

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 object-oriented cluster analysis method for precipitation fields. *Mon. Wea. Rev*., **139**,3673-3693.

Johnson, A., X. Wang, M. Xue, and F. Kong, 2011b: Hierarchical cluster analysis of a convection-allowing ensemble during the Hazardous Weather Testbed 2009 Spring Experiment. Part II: Season-long ensemble clustering and implication for optimal ensemble design. *Mon. Wea. Rev*., **139**, 3694-3710.

Johnson, A. and X. Wang, 2012: Verification and calibration of neighborhodd and object-based probablistic precipitation forecasts from a multi-model convection-allowing ensemble. *Mon. Wea. Rev*., **140**, 3054-3077

Johnson, A. and X. Wang, 2013: Object-based evaluation of a storm scale ensemble during the 2009 NOAA Hazardous Weather Testbed Spring Experiment. *Mon. Wea. Rev.*, **141**, 1079-1098.

Jones, T. A., D. Stensrud, L. Wicker, P. Minnis, R. Palikonda, 2014: Simultaneous radar and satellite data storm-scale assimilation using an ensemble Kalman filter approach for 24 May 2011. *Mon. Wea. Rev*., In Press.

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

Kong, F., M. Xue, K. W. Thomas, Y. Wang, K. Brewster, X. Wang, J. Gao, S. J. Weiss, A. Clark, J. S. Kain, M. C. Coniglio, and J. Du, 2010: Evaluation of 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., P452.

Kühnlein, C, C. Keil, G. C. Craig, C. Gebhardt, 2014: The impact of downscaled initial condition perturbations on convective-scale ensemble forecasts of precipitation. *Q. J. R. Meteorol. Soc*., **140**, 1552-1562.

Lakshmanan, V., T. Smith, G. Stumpf, K. Hondl, 2007a: The Warning Decision Support System–Integrated Information. *Wea. Forecasting*, **22**, 596–612.

Lakshmanan, V., A. Fritz, T. Smith, K. Hondl, and G. J. Stumpf, 2007b: An automated technique to quality control radar reflectivity data. *J. Applied Meteorology*, **46**, 288-305.

Lakshmanan, V., J. Zhang, and K. Howard, 2010: A technique to censor biological echoes in radar reflectivity data. *J. Applied Meteorology*, **49**, 435-462.

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

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

Li, X., M. Charron, L. Spacek, G. Candille, 2008: A regional ensemble prediction system based on moist targeted singular vectors and stochastic parameter perturbations. *Mon. Wea. Rev*., **136**, 443–462.

Li, Y., X. Wang, M. Xue, 2012: Assimilation of radar radial velocity data with the WRF hybrid ensemble–3DVAR system for the prediction of Hurricane Ike (2008). *Mon. Wea. Rev*., **140**, 3507–3524.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci*., **20**, 130–141.

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

Lu, C., H. Yuan, B. E. Schwartz, S. G. Benjamin, 2007: Short-range numerical weather prediction using time-lagged ensembles. *Wea. Forecasting*, **22**, 580–595.

Luo, Y., L. Zhang, 2011: A case study of the error growth and predictability of a Meiyu frontal precipitation event. *Acta Meteorological Sinica*, **25:4**, 430-440.

Maddox, R. A., 1983: Large-scale meteorological conditions associated with midlatitude, mesoscale convective complexes. *Mon. Wea. Rev*., **111**, 1475–1493.

Marsigli, C., A. Montani, F. Nerozzi, T. Paccagnella, S. Tibaldi, F. Molteni, and R. Buizza, 2001: A strategy for high-resolution ensemble prediction. II: Limited area experiments in four Alpine flood events. *Quart. J. Roy. Meteor. Soc*., **127**, 2095–2115.

McAnelly, R. L., J. E. Nachamkin, W. R. Cotton, M. E. Nicholls, 1997: Upscale evolution of MCSs: Doppler radar analysis and analytical investigation. *Mon. Wea. Rev.,* **125**, 1083-1110.

Meng, Z., and F. Zhang, 2008: Tests of an ensemble Kalman filter for mesoscale and regional-scale data assimilation. Part IV: Comparison with 3DVAR in a month-long experiment. *Mon. Wea. Rev*., **136**, 3671–3682.

Michel, Y., T. A., T. Montmerle, 2011: Heterogeneous convective-scale background error covariances with the inclusion of hydrometeor variables. *Mon. Wea. Rev*., **139**, 2994–3015.

Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *J. Geophys. Res*., **102**, 16,663-16,682.

Morrison, H., and W. W. Grabowski, 2008: A novel approach for representing ice microphysics in models: Description and tests using a kinematic framework. *J. Atmos. Sci*., **65**, 1528–1548.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor*., **12**, 595–600.

NCAR, 2011: Gridpoint Statistical Interpolation (GSI) Version 3.0 User's Guide. Available at http://www.dtcenter.org/com-GSI/users/docs/.

Noh, Y., W. G. Cheon, S. Y. Hong, and S. Raasch, 2003: Improvement of the K-profile model for the planetary boundary layer based on large eddy simulation data. *Bound.-Layer Meteor*., **107**, 421–427.

Palmer, R. D., and Coauthors, 2011: Observations of the 10 May 2010 tornado outbreak using OU-PRIME: Potential for new science with high-resolution polarimetric radar. *Bull. Amer. Meteor. Soc*., **92**, 871–891.

Pan, Y., K. Zhu, M. Xue, X. Wang, M. Hu, S. G. Benjamin, S. S. Weygandt, J. S. Whitaker, 2014: A GSI-based coupled EnSRF-En3DVar hybrid data assimilation system for the operational Rapid Refresh model: Tests at a reduced resolution, *Mon. Wea. Rev*. In Press.

Peña, M. and E. Kalnay, 2004: Separating fast and slow modes in coupled chaotic systems. *Nonlin. Processes Geophys*., **11**, 319-327.

Peralta, C., Z. B. Bouallegue, S. E. Theis, C. Gebhardt, M, Buchhold, 2012: Accounting for initial condition uncertainties in COSMO-DE-EPS. *J. Geophys. Res*., **117**, 1-13.

Perkey, D. J., R. A. Maddox, 1985: A numerical investigation of a mesoscale convective system. *Mon. Wea. Rev*., **113**, 553–566.

Potvin, C. K., L. J. Wicker, M. I. Biggerstaff, D. Betten, A. Shapiro, 2013: Comparison between dual-doppler and EnKF storm-scale wind analyses: The 29–30 May 2004 Geary, Oklahoma, supercell thunderstorm. *Mon. Wea. Rev*., **141**, 1612–1628.

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

Rogers, E., and Coathors, 2009: NCEP North American mesoscale modeling system: Recent changes and future plans. Extended Abstracts, *19th Conf. on Numerical Weather Prediction and 23rd Conf. on Weather Analysis and Forecasting*, Omaha, NE, Amer. Meteor. Soc., 2A.4. [Available online at http://ams.confex.com/ams/23WAF19NWP/techprogram/paper_154114.htm].

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*., In Press.

Rotunno, R., 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.

Schenkman, A. D., M. Xue, A. Shapiro, K. Brewster, J. Gao, 2011: Impact of CASA radar and Oklahoma Mesonet data assimilation on the analysis and prediction of tornadic mesovortices in an MCS. *Mon. Wea. Rev*., **139**, 3422–3445.

Schwartz, Craig 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.

Schwartz, C. S., Z. Liu, 2014: Convection-permitting forecasts initialized with continuously cycling limited-area 3DVAR, ensemble Kalman filter, and "Hybrid" variational–ensemble data assimilation systems. *Mon. Wea. Rev*., **142**, 716–738.

Simonin, D., S. P. Ballard, Z. Li, 2014: Doppler radar radial wind assimilation using an hourly cycling 3D-Var with a 1.5km resolution version of the Met Office Unified Model for nowcasting. *Quart. J. Roy. Meteor. Soc*., In Press.

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

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, and J. G. Powers, 2005: A description of the advanced research WRF version 2. NCAR Tech Note NCAR/TN-468_STR, 88 pp. [Available from UCAR Communications, P.O. Box 3000, Boulder, CO 80307.].

Skamarock, W. C., and J. B. Klemp, 2007: A time-split nonhydrostatic atmospheric model for research and NWP applications. *J. Comput. Phys*., **135**, 3465–3485.

Snook, N., M. Xue, Y. Jung, 2011: Analysis of a tornadic mesoscale convective vortex based on ensemble Kalman filter assimilation of CASA X-Band and WSR-88D radar data. *Mon. Wea. Rev*., **139**, 3446–3468.

Snyder, C. and F. Zhang, 2003: Assimilation of simulated Doppler radar observations with an ensemble Kalman filter. *Mon. Wea. Rev*., **131**, 1663-1677.

Sobash, R. A., D. J. Stensrud, 2013: The impact of covariance localization for radar data on EnKF analyses of a developing MCS: Observing system simulation experiments. *Mon. Wea. Rev*., **141**, 3691–3709.

Stensrud, D. J., H. Brooks, J. Du, M. S. Tracton, E. Rogers, 1999: Using ensembles for short-range forecasting. *Mon. Wea. Rev*., **127**, 433-446.

Stensrud, D. J., J. M. Fritsch, 1993: Mesoscale convective systems in weakly forced large-scale environments. Part I: Observations. *Mon. Wea. Rev*., **121**, 3326–3344.

Stensrud, D. J. and coauthors, 2009: Convective-scale warn-on-forecast system. *Bull. Amer. Meteor. Soc*., **90**, 1487–1499.

Stephan, K., K., S. and Schraff, C., 2008: Assimilation of radar-derived rain rates into the convective-scale model COSMO-DE at DWD. *Q.J.R. Meteorol. Soc*., **134**: 1315–1326. doi: 10.1002/qj.269.

Stratman, D. R., M. C. Coniglio, S. E. Koch, and M. Xue, 2013: Use of multiple verification methods to evaluate forecasts of convection from hot- and cold-start convection-allowing models. *Wea. Forecasting*, **28**, 119–138.

Tao, W.-K., and Coauthors, 2003: Microphysics, radiation, and surface processes in the Goddard Cumulus Ensemble (GCE) model. *Meteor. Atmos. Phys*., **82**, 97–137.

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 in the initial state as a factor in the 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.

Thompson, T., 2014: Ensemble Kalman Filter Methods for Convective-Scale Radar Data Assimilation and Multi-Scale Data Assimilation of the 13 June 2010 Tornadic Supercell Environment. Ph.D. Dissertation. School of Meteorology, University of Oklahoma.

Tong, M., M. Xue, 2005: Ensemble Kalman filter assimilation of doppler radar data with a compressible nonhydrostatic model: OSS Experiments. *Mon. Wea. Rev*., **133**, 1789–1807.

Toth, Z., E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev*., **125**, 3297–3319.

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

Walser, A., D. Lüthi, C. Schär, 2004: Predictability of precipitation in a cloud-resolving model. Mon. Wea. Rev., 132, 560–577.

Wang, H., J. Sun, X. Zhang, X.-Y. Huang, and T. Auligné, 2013: Radar data assimilation with WRF 4D-Var. Part I: System development and preliminary testing. *Mon. Wea. Rev*., **141**, 2224–2244.

Wang, Y., M. Bellus, J.-F. Geleyn, X. Ma, W. Tian, and F. Weidle, 2014: A new method for generating initial condition perturbations in a regional ensemble prediction system: Blending. *Mon. Wea. Rev*., **142**, 2043–2059.

Wang, X., C. H. Bishop, 2003: A comparison of breeding and Ensemble Transform Kalman Filter ensemble forecast schemes. *J. Atmos. Sci*., **60**, 1140–1158.

Wang, X., C. H. Bishop, S. J. Julier, 2004: Which Is better, an ensemble of positive–negative pairs or a centered spherical simplex ensemble?. *Mon. Wea. Rev*., **132**, 1590–1605.

Wang, X., D. Barker, C. Snyder, T. M. Hamill, 2008a: A hybrid ETKF-3DVAR data assimilation scheme for the WRF model. Part I: observing system simulation experiment. *Mon. Wea. Rev*., **136**, 5116-5131.

Wang, X., D. M. Barker, C. Snyder, T. M. Hamill, 2008b: A hybrid ETKF–3DVAR data assimilation scheme for the WRF model. Part II: Real observation experiments. *Mon. Wea. Rev*., **136**, 5132–5147.

Wang, X., 2011: Application of the WRF hybrid ETKF–3DVAR data assimilation system for hurricane track forecasts. *Wea. Forecasting*, **26**, 868–884.

Wang, X., D. Parrish, D. Kleist, J. Whitaker, 2013: GSI 3DVar-based ensemble–variational hybrid data assimilation for NCEP Global Forecast System: Single-resolution experiments. *Mon. Wea. Rev*., **141**, 4098–4117.

Wang, X., T. Lei, 2014: GSI-based four dimensional ensemble-variational (4DEnsVar) data assimilation: formulation and single resolution experiments with real data for NCEP Global Forecast System. *Mon. Wea. Rev*., **142**, 3303-3325.

Wang, X. and X. Lu, 2014: Improving vortex scale hurricane prediction using the hybrid EnKF-Var data assimilation method. *World Weather Open Science Conference*, Montreal, QC.

Wee, T.-K., Y.-H. Kuo, D.-K. Lee, Z. Liu, W. Wang, S.-Y. Chen, 2012: Two overlooked biases of the Advanced Research WRF (ARW) model in geopotential height and temperature. *Mon. Wea. Rev*., **140**, 3907–3918.

Whitaker, J. S., T. M. Hamill, 2002: Ensemble data assimilation without perturbed observations. *Mon. Wea. Rev*., **130**, 1913–1924.

Whitaker, J. S., T. M. Hamill, X. Wei, Y. Song, Z. Toth, 2008: Ensemble data assimilation with the NCEP Global Forecast System. *Mon. Wea. Rev*., **136**, 463–482.

Whitaker, J. S., G. P. Compo, J.-N. Thépaut, 2009: A comparison of variational and ensemble-based data assimilation systems for reanalysis of sparse observations. *Mon. Wea. Rev*., **137**, 1991–1999.

Whitaker, J. S., T. M. Hamill, 2012: Evaluating methods to account for system errors in ensemble data assimilation. *Mon. Wea. Rev*., **140**, 3078–3089.

Wilks, D. S., 2006: Statistical Methods in the Atmospheric Sciences: An Introduction. 2nd ed. Academic Press, 467 pp.

Wu, W., R. J. Purser, and D. F. Parrish, 2002: Three-dimensional variational analysis with spatially inhomogeneous covariances. *Mon. Wea. Rev*., **130**, 2905–2916.

Xiao, Q. and J. Sun, 2007: Multiple-radar data assimilation and short-range quantitative precipitation forecasting of a squall line observed during IHOP_2002. *Mon. Wea. Rev*., **135**, 3381–3404.

Xie, Y., S. Koch, J. McGinley, S. Albers, P. E. Bieringer, M. Wolfson, M. Chan, 2011: A space–time multiscale analysis system: A sequential variational analysis approach. *Mon. Wea. Rev*., **139**, 1224–1240.

Xu, M., D. J. Stensrud, J.-W. Bao, T. T. Warner, 2001: Applications of the adjoint technique to short-range ensemble forecasting of Mesoscale Convective Systems. *Mon. Wea. Rev*., **129**, 1395–1418.

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. Physics*, **82**, 139-170.

Xue, M., F. Kong, K. W. Thomas, Y. Wang, K. Brewster, J. Gao, X. Wang, S. Weiss, A. Clark, J. Kain, M. Coniglio, J. Du, T. Jensen, and Y.-H. Kuo, 2010a: CAPS realtime storm scale ensemble and high resolution forecasts for the NOAA Hazardous Weather Testbed 2010 Spring Experiment. *25th Conf. Severe Local Storms*, Amer. Meteor. Soc., Paper 7B.3.

Xue, M., and Coauthors, 2010b: CAPS realtime storm scale ensemble and high resolution forecasts for the NOAA Hazardous Weather Testbed 2010 Spring Experiment. Preprints, 25th Conf. on Severe Local Storms, Denver, CO, Amer. Meteor. Soc., 7B.3. [Available online at https://ams.confex.com/ams/25SLS/techprogram/paper_176056.htm.]

Yang, S.-C., M. Corazza, A. Carrassi, E. Kalnay, T. Miyoshi, 2009: Comparison of local ensemble transform Kalman filter, 3DVAR, and 4DVAR in a quasigeostrophic model. *Mon. Wea. Rev*., **137**, 693–709.

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

Zhang, D.-L., J. M. Fritsch, 1988: Numerical sensitivity experiments of varying model physics on the structure, evolution and dynamics of two mesoscale convective systems. *J. Atmos, Sci*., **45**, 261-293.

Zhang, F., C, Snyder, R, Rotunno, 2003: Effects of moist convection on mesoscale predictability. *J. Atmos. Sci*., **60**, 1173–1185.

Zhang, F., A. M. Odins, J. W. Nielsen-Gammon, 2006: Mesoscale predictability of an extreme warm-season precipitation event. *Wea. Forecasting*, **21**, 149–166.

Zhang, F., N. Bei, R. Rotunno, C. Snyder, C. C. Epifanio, 2007: Mesoscale predictability of moist baroclinic waves: Convection-permitting experiments and multistage error growth dynamics. *J. Atmos. Sci*., **64**, 3579–3594.

Zhang, F., Y. Weng, J. A. Sippel, Z. Meng, and C. H. Bishop, 2009: Cloud-resolving hurricane initialization and prediction through assimilation of doppler radar observations with an ensemble Kalman filter. *Mon. Wea. Rev*., **137**, 2105–2125.

Zhang, F., Y. Weng, Y.-H. Kuo, J. S. Whitaker, and B. Xie, 2010: Predicting typhoon Morakot's catastrophic rainfall with a convection-permitting mesoscale ensemble system. *Wea. Forecasting*, **25**, 1816–1825.

Zhang, J., and Coauthors, 2011: National Mosaic and Multi-Sensor QPE (NMQ) system: Description, results and future plans. *Bull. Amer. Meteor. Soc*., **92**, 1321-1338.

Zhang, M., F. Zhang, X.-Y. Huang, X. Zhang, 2011: Intercomparison of an ensemble Kalman filter with three- and four-dimensional variational data assimilation methods in a limited-area model over the month of June 2003. *Mon. Wea. Rev*., **139**, 566–572.

Zhao, Q., J. Cook, Q. Xu, P. R. Harasti, 2008: Improving short-term storm predictions by assimilating both radar radial-wind and reflectivity observations. *Wea. Forecasting*, **23**, 373–391.

Zhu, K., Y. Pan, M. Xue, X. Wang, J. S. Whitaker, S. G. Benjamin, S. S. Weygandt, M. Hu, 2013: A regional GSI-based ensemble Kalman filter data assimilation system for the Rapid Refresh configuration: Testing at reduced resolution. *Mon. Wea. Rev*., **141**, 4118–4139.