DISSERTATION

FEATURES BASED ASSESSMENTS OF WARM SEASON CONVECTIVE PRECIPITATION FORECASTS FROM THE HIGH RESOLUTION RAPID REFRESH MODEL

Submitted by

Janice L. Bytheway

Department of Atmospheric Science

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2017

Doctoral Committee:

Advisor: Christian Kummerow

Russ Schumacher
David Randall
V. Chandrasekar
Curtis Alexander

ABSTRACT

FEATURES BASED ASSESSMENTS OF WARM SEASON CONVECTIVE PRECIPITATION FORECASTS FROM THE HIGH RESOLUTION RAPID REFRESH MODEL

Forecast models have seen vast improvements in recent years, via increased spatial and temporal resolution, rapid updating, assimilation of more observational data, and continued development and improvement of the representation of the atmosphere. One such model is the High Resolution Rapid Refresh (HRRR) model, a 3 km, hourly-updated, convection-allowing model that has been in development since 2010 and running operationally over the contiguous US since 2014. In 2013, the HRRR became the only US model to assimilate radar reflectivity via diabatic assimilation, a process in which the observed reflectivity is used to induce a latent heating perturbation in the model initial state in order to produce precipitation in those areas where it is indicated by the radar.

In order to support the continued development and improvement of the HRRR model with regard to forecasts of convective precipitation, the concept of an assessment is introduced. The assessment process aims to connect model output with observations by first validating model performance then attempting to connect that performance to model assumptions, parameterizations and processes to identify areas for improvement. Observations from remote sensing platforms such as radar and satellite can provide valuable information about three-dimensional storm structure and microphysical properties for use in the assessment, including estimates of surface rainfall, hydrometeor types and size distributions, and column moisture content.

A features-based methodology is used to identify warm season convective precipitating objects in the 2013, 2014, and 2015 versions of HRRR precipitation forecasts, Stage IV multisensor precipitation products, and Global Precipitation Measurement (GPM) core satellite observations. Quantitative precipitation forecasts (QPFs) are evaluated for biases in hourly rainfall intensity, total rainfall, and areal coverage in both the US Central Plains (29-49N, 85-105W) and US Mountain West (29-49N, 105-125W). Features identified in the model and Stage IV were tracked through time in order to evaluate forecasts through several hours of the forecast period. The 2013 version of the model was found to produce significantly stronger convective storms than observed, with a slight southerly displacement from the observed storms during the peak hours of convective activity (17-00 UTC). This version of the model also displayed a strong relationship between atmospheric water vapor content and cloud thickness over the central plains. In the 2014 and 2015 versions of the model, storms in the western US were found to be smaller and weaker than the observed, and satellite products (brightness temperatures and reflectivities) simulated using model output indicated that many of the forecast storms contained too much ice above the freezing level.

Model upgrades intended to decrease the biases seen in early versions include changes to the reflectivity assimilation, the addition of sub-grid scale cloud parameterizations, changes to the representation of surface processes and the addition of aerosol processes to the microphysics. The effects of these changes are evident in each successive version of the model, with reduced biases in intensity, elimination of the southerly bias, and improved representation of the onset of convection.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisor, Prof. Chris Kummerow, who saw something in me many years ago that I was unable to see myself, and who had the patience to guide me and help me to grow as a scientist. Curtis Alexander at NOAA not only provided all of the data I needed, but his guidance and discussions were instrumental to this research. I would like to thank my committee members, Curtis, Dave Randall, Russ Schumacher, and Chandra, for taking time out of their schedules to serve on my committee and for their valuable feedback during this process. I would also like to thank Chris O'Dell; though he served on my committee for only a brief time, his input early in this process was greatly appreciated. Having my results endorsed by the model developers themselves was not only helpful in advancing this research, but also served to build my confidence, and for that I again thank Curtis as well as Stan Benjamin at NOAA. Many thanks to Jamie Schmidt, Sarah Tisdale, and the entire department staff for their ever-helpfulness in navigating the University paper trail. To the Kummerow research group, thank you for your "Kummraderie" over the past 11 years. I can only hope to have such wonderful colleagues in my future endeavors. Lastly, I've been fortunate to have the love and support of my family, friends, and their families who have accepted me as one of their own over the years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Motivation and Previous Work

Quantitative precipitations forecasts (QPFs) from numerical weather prediction (NWP) models are employed by a variety of users with a wide range of meteorological understanding. They are created on time scales ranging from climatological outlooks to nowcasting on the order of tens of minutes. Applications of QPFs include water management planning for agricultural and urban applications, electricity generation, flood and drought forecasting, and planning of activities by the general public (Fritsch and Carbone, 2004). QPFs, particularly those with high spatial and temporal resolution, have the potential to be used in conjunction with hydrologic models to improve predictions of streamflow and potential warm-season flooding (Gilmore et al., 2004). Accurate forecasts of convective precipitation are also necessary to meet the National Weather Service's (NWS) goal of "warn-on-forecast" for severe weather events including severe thunderstorms, tornadoes, and flash flooding (Stensrud et al., 2009).

Ideally, precipitation forecasts would be both very high resolution and very accurate, however NWP model performance is dependent on many factors, including the season, grid resolution, availability of observations for assimilation, and the meteorological processes being modeled (Warner, 2011). Successful forecasts of any atmospheric variable require a good representation of that variable's initial state (Rogers et al., 2000; Sugimoto et al., 2009; Shrestha et al., 2013), since a poor initial state will likely propagate through the forecast period, degrading the forecast quality (i.e. "garbage in = garbage out"). Many models obtain this initial state via n-dimensional variational data assimilation (nDVAR), which calculates an atmospheric state that is optimally consistent with data from a variety of observational networks (Lopez, 2007). These observations are generally direct measurements of, or at least linearly related to, the model prognostic variables (e.g. pressure, temperature, wind, moisture) and can be assumed to have normally distributed errors. On the other hand, precipitation is a diagnostic variable; it is not explicitly forecast, but rather calculated as a function of hydrometeor concentrations and size distributions. The diagnostic nature of precipitation makes it notoriously difficult to accurately initialize in the model fields.

Precipitation is a highly heterogeneous variable in both time and space, and surface observations of precipitation from gauges are too widespread to accurately capture this variability and represent it in forecast initial states. Higher resolution observations of precipitation can be obtained from radar and satellite, however the observed variables (reflectivity and radiance) are both indirectly and non-linearly related to in-cloud hydrometeor distributions and precipitation rate, and have non-normally distributed uncertainties (Lopez, 2007). Furthermore, precipitation forecasts depend on more than just the environment characterized by the prognostic variables, because accurate modeling of precipitation requires the representation of condensation and evaporation, heat and moisture transport through the atmosphere, hydrometeor phase changes, and other phenomena not explicitly calculated by most NWP models (Shrestha et al., 2013). Thus, QPFs rely heavily on indirect relationships between prognostic variables and precipitation that are estimated using simplified assumptions and parameterizations (Lopez, 2007; Hou and Zhang, 2007). Additionally, if the model is strongly constrained by the initial state of the prognostic variables, surface precipitation will not be produced without a conducive atmospheric state, regardless of whether its presence is indicated by the assimilated observations (Hou and Zhang, 2007; Rogers et al., 2000).

The ability of NWP models to accurately predict precipitation also depends on the horizontal scale and longevity of the precipitation being forecast. Generally, models tend to be most skillful when forecasting large, synoptically forced precipitation, and least skillful at the convective scale (Ebert et al., 2007; Wernli et al., 2008: Shrestha et al., 2013). Mesoscale precipitation can be somewhat accurately forecast for relatively short timescales on the order of 1-2 days (Davis et al., 2006), but on shorter time scales, the spatially discontinuous and rapidly varying nature of convective precipitation make accurate prediction via NWP difficult. Fritsch and Carbone (2004) and Lin et al. (2005) suggest that, on time scales of a 1-3 hours, simple advection of radar echoes tends to produce more accurate precipitation forecasts than NWP output.

Cloud resolving models (CRMs) that produce frequently-updated forecasts at spatial resolutions on the order of a few kilometers are generally better at resolving the dynamics of mesoscale convective systems (MCSs) than their regional or global-scale counterparts. While the increased resolution and update frequency of such models is expected to mitigate some of the challenges to short-term convective precipitation forecasting, CRMs still have many sources of uncertainty, including radiative interaction, surface-atmosphere exchanges, and the microphysical, radiative, and dynamic processes that occur on sub-grid scales (Bryan and Morrison, 2012; Shrestha et al., 2013). The representation of ice in the model (i.e. number and type of species, particle size distributions (PSDs), particle fall speed) has been shown to play a large role in the success of a convective precipitation forecast (Gilmore et al., 2004, Bryan and Morrison, 2012). Given the continued challenges to convective-scale precipitation forecasting despite the improved capabilities offered by CRMs, Burghardt et al. (2014) suggest that further improvements to the representation of such systems by NWP models will require more than simply increasing the model resolution to the order of a few hundred meters or higher, but will require better understanding and parameterization of processes at those scales.

In addition to the challenges presented by the behavior of the convective system themselves and the difficulty obtaining an accurate initial state, NWP models need to “spin up” from their initial state (Sheng et al., 2006; Rogers et al., 2000; Errico et al., 2007; Sugimoto et al., 2009; Stensrud et al., 2009; Sun et al., 2014; Weygandt et al., 2007 and 2008). Including precipitation-effected observations in the initial model state with rapid updates can help to improve the representation of convective systems in the model analysis and reduce this spin-up time, however ground based radar is currently the only instrument capable of providing data at high enough temporal resolution for the frequent assimilation cycles used in modern storm-resolving CRMs (Errico et al., 2007; Sun et al., 2014).

Many studies over the past 10-15 years have now begun to evaluate the impact of assimilating ground-based radar reflectivity and/or radial velocities as part of the model initialization. For example, Rogers et al. (2000) used radar reflectivity data in a mesoscale model not to infer rain rates, but simply to turn the model's convective parameterization on or off, assuming that if a model is forced to have convection when and where it is observed, the initialization and therefore the forecast will be more accurate. Sheng et al. (2006) used 3-dimensional (3D) radar reflectivity and radial velocity data with infrared satellite radiances and background analyses to determine initial 3D cloud and hydrometeor fields as well as temperature and moisture perturbations due to clouds and rainfall. Their results indicate that a model that is correctly initialized with radar reflectivity can potentially predict short-term precipitation better than a standard reflectivity-rain rate (Z-R) relationship. Xaio and Sun (2007) and Sugimoto et al. (2009) assimilated radar reflectivity and radial velocity into the Weather Research and

Forecasting (WRF) 3DVAR assimilation system, while Stratman et al. (2013) compared WRF forecasts both with and without radar data assimilation. Those studies all found improvement in forecasts of convective precipitation as a result of including this data.

While assimilation of radar reflectivity has produced positive results, direct assimilation of reflectivity to calculate initial rain rates is not particularly desirable due to the non-linear relationship between the two discussed above. Instead, Sun et al. (2014) suggest diabatic assimilation, in which radar reflectivity is used to determine a perturbation in latent heating within the column. Diabatic assimilation has the benefits of imparting a change on one model variable (temperature) that translates to other, related model variables (e.g. vertical velocity, humidity) fairly quickly and avoiding the large uncertainty in the relationship between reflectivity and rain rate. The High Resolution Rapid Refresh (HRRR) model developed at the National Oceanic and Atmospheric Administration (NOAA) Assimilation and Modeling Branch (AMB) has employed diabatic assimilation of radar data in hourly forecasts since 2013 (Benjamin et al., 2016).

Increased resolution and the addition of radar data to the initial state have resulted in improved QPFs from NWP models. However, precipitation forecasts remain subject to errors. These errors are often described in terms of the timing, location, and intensity of precipitation (Marzban et al., 2009; Sugimoto et al., 2009; Shrestha et al., 2013). In order to evaluate model performance, some method to describe these errors is needed. This can be accomplished through validation studies, which evaluate the model performance with regard to these and other measures of accuracy and provide feedback to the model's developers and end users (Lack et al., 2010; Shrestha et al., 2013).

In order to perform verification of the model, one must first define what constitutes a ‘‘good forecast,’’ and then determine what metrics will be used to assess this quality. Verification studies focused on convection generally focus on how well the model represented the observed precipitation and are often aimed at informing users of the NWP output, for instance, public or private meteorologists who disseminate forecast information or hydrologists who use NWP output as inputs to hydrological models. While evaluations of model performance are beneficial to users and developers alike, they do not typically provide the developers with any information as to why the model did or did not produce an accurate forecast.

Future improvements to NWP models require information on what model processes are the underlying causes of a successful or failed QPF. Knievel et al. (2004) touched on this concept, showing that evaluating the modes of observed and forecast rainfall can point to regional, seasonal, and phenomenological dependencies in QPF quality. Bytheway and Kummerow (2015) suggest the idea of a model “assessment,” a process by which the model’s forecast of a particular atmospheric variable is not only validated, but an explanation for successful or poor forecast results is also sought. Both verification and assessment of NWP models follow the same initial steps: to define a successful forecast and the metrics that determine success or failure. Assessment goes a step further by attempting to relate the verification results to model assumptions (e.g., drop size distribution, Z-R relationships, etc.), parameterizations, or other forecast environment variables.

The assessment begins by selecting a verification technique, which can generally be categorized as either grid-point or spatial diagnostics. Grid-point methodologies compare the forecast and observed fields at each individual grid point, often employing a contingency table to categorize each pixel as either a successful forecast of an observed event, a missed forecast of an observed event, a “false alarm” forecast of an event that was not observed, or a successful “no event” forecast. Using the contingency table, many validation statistics can be calculated, including (but not limited to) Critical Success Index (CSI), False Alarm Ratio (FAR), Probability of Detection (POD), and Heidke Skill Score (HSS). Grid-point techniques, while informative, have a well-documented weakness in the “double penalty” problem, wherein an accurate but incorrectly located forecast is penalized twice: once as a missed forecast, and once as a false alarm (AghaKouchak et al., 2011; Ebert and McBride, 2000; Gilleland et al., 2009). Spatial diagnostic approaches, on the other hand, attempt to mimic the subjective ability of a human observer to identify similar “objects” in the observed and forecast fields and determine how well the forecast matches what was observed (Davis et al., 2006). These methods tend to be more intuitive to interpret (Ebert and Gallus, 2009), and allow significant flexibility in the measures of model performance that can be evaluated. This flexibility means that QPF quality can be evaluated with respect to other model variables, potentially providing information that can lead to model improvement. Therefore, spatial diagnostic techniques are determined to be a better fit for the goals of model assessments.

Gilleland et al. (2009) describe four general types of spatial verification techniques: neighborhood, scale separation, features based, and field deformation. The first two apply some type of spatial filtering to the observed or forecast field (or both) and calculate verification statistics on those filtered fields. One example of a neighborhood technique includes the Fractions Skill Score (FSS, Roberts and Lean, 2008), which can be used to determine the scales over which a model forecast can be considered “skillful.” The last two categories attempt to fit the forecast to the observations as well as possible, with the difference being that features-based methods identify features of interest and analyze each one separately, while field deformation techniques analyze the entire field or subsets therein. The studies described in this dissertation will follow a features or object-based verification method as a first step to model assessment.

Several features based methodologies for precipitation forecast verification have been developed in recent years. As Gilleland et al. (2009) point out, these methodologies usually differ in how a feature is defined, whether spatially discontinuous features in a field are considered together or separately, how features are matched from one field to the next, and which diagnostics are used.

Some examples of features based methodologies include the Contiguous Raining Area (CRA) described by Ebert and McBride (2000) and Ebert and Gallus (2009). This methodology isolates the rainfall of a particular system in both the forecast and model fields, and defines the CRA as the union of the forecast and observed precipitation areas. They then aim to identify the displacement error, correct for it, and validate the location-corrected forecast using a variety of statistics, including the FAR and POD, as well as errors in magnitude, raining area, and rain volume.

Wernli et al. (2008 and 2009) describe the Structure, Amplitude, Location (SAL) technique, which describes the model errors in each of these properties separately for each identified raining feature. The amplitude component measures the relative deviation of the domain-averaged QPF from the observed. The location and structure components require the identification of coherent precipitating objects to determine the displacement of their centers of mass, whether the object is an appropriate size, and whether the distribution of rainfall within it is too flat or peaked. Disadvantages to this method include the fact that since there are a small number of parameters being considered, the ability to uniquely describe model performance is diminished (i.e., different sorts of errors can result in identical values for the SAL parameters).

Although individual objects are identified, the statistics are calculated with respect to the entire domain of interest. Therefore, this methodology is difficult to apply in a large domain, particularly if there are several meteorologically distinct precipitation systems within it.

Lack et al. (2010) developed a verification scheme that used a weighted cost function to identify matches between observed and forecast precipitation features using a variety of criteria. For each forecast-observed feature pair, a Procrustes fit is calculated using a translation component, dilation component and rotation component. The fit is then used to calculate the forecast penalty. Disadvantages to this methodology include the inability to disallow matches with large separation distance when using the cost function to determine matches between the forecast and observations.

Davis et al. (2006 and 2009) developed the Method for Object-Based Diagnostic Evaluation (MODE) validation technique, wherein the precipitating fields are spatially smoothed to acquire a more contiguous raining area, then various statistics are calculated for each identified feature, including center of mass, axis angle, aspect ratio, and the cumulative distribution function (CDF) of rainfall. In Davis et al. (2006), they determine a match between an observed and forecast object using a distance threshold that takes into account the size of both the observed and forecast precipitation objects. Later, in Davis et al. (2009), the match criteria were strengthened into a parameter called Total Interest (TI), which describes the match between forecast and observed features in terms of distance separation of both the centers of mass and object boundaries, orientation angle difference, area ratio, and intersection area. In essence, the more alike two objects are, the more likely they are to be a match.

With recent advances in the assimilation of radar into NWP models and continuing increases in their spatial and temporal resolution, there is a clear need to evaluate model performance at high resolution and sub-daily time scales, and to understand how assumptions, parameterizations and physical representations of the atmosphere affect forecast quality. Specifically, since the HRRR is currently the only operational US model to assimilate radar data, there is a need to understand how its forecasts of precipitation perform under a variety of conditions, and how continuing updates to the model over time impact the quality of the QPFs. As such, this study will demonstrate the use of a features-based verification methodology to identify and assess QPFs from the HRRR. Given that the HRRR model was designed to provide rapid update model guidance on convective storms for use in severe weather forecasting, air traffic management and aviation hazards forecasting, and warning dissemination, as well as to eventually provide improved background fields for real-time mesoscale analysis (ESRL, 2013), this study will focus on the assessment of warm season convective precipitation features.

1.2 Outline of Dissertation

Features-based assessments of the HRRR were performed over the course of several years, during which time the model was undergoing continuous development. Thus, the ensuing chapters assess different versions of the model. Each chapter may be treated as an individual manuscript, either published or submitted for publication at the time of this writing.

Chapter 2 comprises the first study, “ Towards an Object-based Assessment of High Resolution Forecasts of Long-lived Convective Precipitation in the Central US,” which was published in the *Journal of Advances in Modeling Earth Systems* as Bytheway and Kummerow (2015). This study focuses on assessing forecasts of mature, long-lived (greater than 10 hours) warm season convection in the US Central Plains produced by the 2013 version of the HRRR, the first version to include the assimilation of radar data. Features were identified and tracked through time in order to examine model performance through the forecast period and to infer the longevity of influence of the assimilated radar. This chapter includes a comprehensive discussion of the features-based methodology used and referenced in ensuing chapters.

Chapter 3, “Consistency Between Cloud Resolving Model Output and Satellite Observations” was submitted to the *International Journal of Remote Sensing* in 2017. In reaction to reviewer comments regarding Bytheway and Kummerow (2015), the domain of interest was altered to include the complex terrain of the western US, where the terrain induces increased susceptibility to flash flooding events and can complicate the model’s ability to produce accurate forecasts. While ground based radar is available in this region, it is often less reliable due to beam blockage. The 2014 launch of the Global Precipitation Measurement (GPM) core satellite brought about a space-based radar source, not susceptible to the issues faced by ground-based radar in complex terrain, and the ability to infer additional information about the cloud systems using observations from the collocated GPM Microwave Imager (GMI). This study focuses on examining the microphysical properties of the clouds produced by the 2014 and 2015 versions of the HRRR by investigating the consistency between observed reflectivities and radiances and those simulated using model output.

Given the continuous development of the HRRR, and the implementation of several changes with the intent to improve precipitation forecasts, Chapter 4, “A Features-based Assessment of the Evolution of Warm Season Precipitation Forecasts from the HRRR Model over Three Years of Development,” (submitted to *Weather and Forecasting*, 2017) focuses on evaluating changes to the model performance through three years (2013, 2014, and 2015) of development. This study was motivated by the findings of previous validations and assessments, both those discussed herein and those performed by the model developers. This chapter seeks to correlate model forecast improvement to specific alterations that were made to the model in order to achieve them.

The results contained herein represent an in depth examination of HRRR forecasts of convective precipitation in the US Great Plains and mountain west. Key findings of the assessments presented in chapters 2-4 will be summarized in Chapter 5.

CHAPTER 2

TOWARDS AN OBJECT-BASED ASSESSMENT OF HIGH-RESOLUTION FORECASTS OF LONG-LIVED CONVECTIVE PRECIPITATION IN THE CENTRAL US

2.1 Introduction

Quantitative precipitation forecasts (QPFs) from numerical weather prediction (NWP) models are employed by a variety of users with diverse levels of meteorological understanding, from flood and drought forecasting by scientific experts, to daily activity planning by the general public. These forecasts provide estimates of precipitation expected to fall in the future, over time periods of a few hours to a few days, and those with high spatial and temporal resolution have the potential to be used in conjunction with hydrologic models to improve predictions of streamflow and potential warm-season flooding (Gilmore et al., 2004).

The accuracy of NWP models depends on the season, grid resolution, availability of observations, and the model's ability to represent the meteorological processes of interest (Warner, 2011). The final factor comes into play twice when forecasting precipitation: once when initializing the model with observations of precipitation (time=0), and again at each forecast time step (time>0). The successful forecast of any atmospheric variable requires a good representation of the initial state of the atmosphere (Rogers et al., 2000; Sugimoto et al., 2009; Shrestha et al., 2013), as a poor initial state will likely propagate through the forecast period, degrading the overall forecast quality. Many models obtain this initial state via n-dimensional variational data assimilation (nDVAR), which calculates an atmospheric state that is optimally consistent with data from a variety of observational networks (Lopez, 2007). The assimilated observations are often direct measurements of, or linearly related to, the model prognostic variables (e.g., pressure, temperature, wind), and can be assumed to have normally distributed errors. Precipitation fields are notoriously difficult to initialize because observations of precipitation are often obtained in the form of satellite radiances or radar reflectivities, which are both indirectly and nonlinearly related to hydrometeor distributions and precipitation rate, and usually have non-normally distributed uncertainties (Lopez, 2007). Additionally, a model that is strongly constrained by the initial state of the prognostic variables will fail to produce surface precipitation in the absence of a conducive atmospheric state, regardless of whether it is indicated by the assimilated observations (Hou and Zhang, 2007; Rogers et al., 2000). In the forecast itself, precipitation rate is generally treated as a diagnostic variable, calculated as a function of forecast hydrometeor concentration and size distribution using simplified assumptions and parameterizations that attempt to capture droplet growth processes and hydrometeor phase changes that are not explicitly calculated by NWP (Lopez, 2007; Hou and Zhang, 2007; Shrestha et al., 2013).

Thus, QPFs from NWP models rely first on indirect relationships between observed variables and precipitation at the initial time step, then on a different set of indirect relationships between the prognostic variables and precipitation rate. Generally, forecasts tend to be most skillful when forecasting large, synoptically forced precipitation fields, and least skillful at the convective scale (Ebert et al., 2007; Wernli et al., 2008; Shrestha et al., 2013). Mesoscale precipitation can be somewhat accurately forecast on relatively short time scales on the order of 1–2 days (Davis et al., 2006).

Cloud resolving models (CRMs) that produce forecasts at spatial resolutions on the order of a few kilometers are generally better at resolving the dynamics of mesoscale convective systems (MCSs) than their regional or global-scale counterparts, but are still subject to many sources of uncertainty, including subgrid-scale microphysical processes, radiative interaction, small-scale turbulence, and surface-atmosphere exchanges. In particular, the representation of ice in the model (number and type of species, particle size distributions, particle fall speed) plays a key role in the success of a convective precipitation forecast (Bryan and Morrison, 2012). In fact, Gilmore et al. (2004) found that changes in the parameterization of the hail/graupel category in a CRM simulation of midlatitude convection varied the amount of surface precipitation by a factor of 3 or more.

Although CRMs have improved mesoscale forecasting ability, Fritsch and Carbone (2004) and Lin et al. (2005) suggest that now-casting based on current radar observations still tends to be more accurate than the NWP forecast, at least on very short time scales of 1–3 h. These studies also suggest that the assimilation of observations from ground or space-based radars may reduce model spin-up time and improve short term forecasting. In fact, Fritsch and Carbone (2004) state that, in regards to convective precipitation forecasts, data assimilation ‘‘may be the most critical path through which the pace of forecast advances will be modulated.’’

The impact of assimilating radar reflectivity and/or radial velocities into NWP models has been an emerging topic of study over the last 10–15 years (Rogers et al., 2000; Sheng et al., 2006; Xiao and Sun, 2007; Sugimoto et al., 2009). More recently, the High-Resolution Rapid Refresh (HRRR) model developed at the National Oceanic and Atmospheric Administration (NOAA) Assimilation and Modeling Branch (AMB) has been assimilating 3 km radar data for use in 1 h forecasts since April 2013, and has been in operational use since September 2014 (ESRL, 2015). Despite the potential improvements in QPF resulting from increased resolution, the use of CRMs and the assimilation of radar data, model forecasts are still subject to errors in the timing, location, and intensity of precipitation (Shrestha et al., 2013; Sugimoto et al., 2009; Marzban et al., 2009).

The goal of NWP model verification is to evaluate model performance with regard to these and other measures of accuracy and provide feedback to the model's developers and end users (Lack et al., 2010; Shrestha et al., 2013). In order to perform verification of the model, one must first define what constitutes a ‘‘good forecast,’’ and then determine what metrics will be used to assess this quality. Verification studies generally focus on how well the model forecast the observed precipitation and are often aimed at informing users of the NWP output, for instance, public or private meteorologists who disseminate forecast information or hydrologists who use NWP output as inputs to hydrological models. While evaluations of model performance are beneficial to users and developers alike, they do not typically provide the developers with any information as to why the model did or did not produce an accurate forecast.

Future improvements to NWP models will rely on determining what model processes are the underlying cause of a successful or failed QPF. Therefore, in this study, we will differentiate between ‘‘verification’’—validation of the model with the intent to describe its performance capabilities, and ‘‘assessment’’—validation of the model with the intent to determine which variables or model processes are likely related to the model's performance. Both verification and assessment of NWP models follow the same initial steps: to define a successful forecast and the metrics that determine success or failure. Assessment goes a step further by attempting to relate the verification results to model assumptions (e.g., drop size distribution, Z-R relationships, etc.) or other forecast environment variables. The current study presents an assessment of the HRRR model.

The assessment begins by selecting a verification technique. Here we select a spatial diagnostic approach rather than a grid point comparison. Spatial diagnostic methods demonstrate several advantages over validation methodologies that compare individual grid points in both the observed and forecast fields. Grid point comparisons have a well-documented ‘‘double penalty’’ problem, wherein grid point verification schemes over-penalize an accurate but incorrectly located forecast twice: once as a missed forecast, and once as a false alarm (AghaKouchak et al., 2011; Ebert and McBride, 2000; Gilleland et al., 2009). Additionally, spatial diagnostic methods tend to be more intuitive to interpret (Ebert and Gallus, 2009) as they mimic the subjective ability of a human observer to identify similar ‘‘objects’’ in the observed and forecast fields and determine how well the forecast matches what was observed (Davis et al., 2006). Spatial diagnostic methods also allow significant flexibility in the measures of model performance that can be evaluated. This flexibility means that QPF quality can be evaluated with respect to other model variables, potentially providing information that can lead to model improvement.

Gilleland et al. (2009) describe four general types of spatial verification techniques: neighborhood, scale separation, features based, and field deformation. The first two apply some type of spatial filtering to the observed or forecast field (or both) and calculate verification statistics on those filtered fields. The last two categories attempt to fit the forecast to the observations as well as possible, with the difference being that features-based methods identify features of interest and analyze each one separately, while field deformation techniques analyze the entire field or subsets therein. The current study will follow a features or object-based verification method as a first step to the model assessment.

Several features-based methodologies for precipitation forecast verification have been developed, including the Contiguous Raining Area (CRA; Ebert and McBride, 2000; Ebert and

Gallus, 2009), Structure Amplitude Location (SAL; Wernli et al., 2008 and 2009), and Method for Object-based Diagnostic Evaluation (MODE; Davis et al., 2006 and 2009) techniques. These methodologies generally differ in how a feature is defined, whether spatially discontinuous features in a field are considered together or separately, how features are matched between forecast and observed fields, and which diagnostics are used (Gilleland et al., 2009). The work described herein partially follows the MODE validation technique (Davis et al., 2006 and 2009) in which the precipitating fields are spatially smoothed to acquire a more contiguous raining area and then various descriptive statistics are calculated for the identified feature.

Since HRRR is intended to be used as an operational forecast model and is likely to undergo many upgrades, there is a need to understand how its forecasts of precipitation perform under a variety of conditions and how they may be improved. This study aims to build on previous work to develop and demonstrate a features-based verification method for high-resolution NWP precipitation forecasts and relate the verification results to model variables and processes to assess potential reasons for the model's success or failure. The assessment is performed on convective precipitating systems in the central United States (U.S.) during the 2013 warm season (May–August).

Section 2.2 describes the HRRR model and the output that was used for this study, as well as the National Centers for Environmental Prediction (NCEP) Stage IV multisensor precipitation product that was used for validation. Section 2.3 outlines the features-based identification scheme, detailing how a feature is defined, how a match between forecast and observed features is determined, and the database of forecast and observed feature matches that were created to store the various statistics used to assess model performance. Results will be presented in section 2.4, followed by concluding remarks in section 2.5.

2.2 Data Sets

2.2.1. Stage IV Radar

The Stage IV multisensor precipitation analysis is available hourly over the contiguous U.S. (CONUS) at 4 km resolution (Lin and Mitchell, 2005), and serves as the reference precipitation data set in this study. This product is a mosaic of regional radar analyses that is adjusted using gauge information. Each National Weather Service (NWS) River Forecast Center (RFC) produces an automated version for their region shortly after the end of the accumulation hour. Then, several hours later, manual quality control is performed on the initial analysis to remove errors that may have made it through the automated processing. These analyses are then combined onto a national grid at NCEP, averaging points where data are provided by multiple RFCs.

This process generally begins 35 min past the end of the hour of accumulation, and the data may be available shortly thereafter. This near real-time analysis may not have contributions from all of the RFCs, which are filled in as they become available. The final mosaics containing data from all RFCs with manual quality control are generally available 12–18 h after accumulation time on a Hydrologic Rainfall Analysis Project (HRAP) grid and are archived at the National Center for Atmospheric Research (NCAR). The final quality controlled mosaics are used as the reference data set in this study.

Though the Stage IV product has a history of use as the reference product for validation of both models and other observational data sets, it is not without its own uncertainties, some discussion of which can be found in Smalley et al. (2014) and Prat and Nelson (2015).

2.2.2. High-Resolution Rapid Refresh

The HRRR NWP model is an hourly updated storm-resolving model running at 3 km horizontal resolution with 50 vertical levels over the CONUS. Its domain is nested within the 13 km Rapid Refresh mesoscale model, which also provides boundary conditions (Benjamin et al., 2013). The HRRR was designed to provide rapid update model guidance on convective storms for use in severe weather forecasting, air traffic management, aviation hazard forecasting, and warning dissemination, as well as to eventually provide improved background fields for real-time mesoscale analysis (ESRL, 2015).

The 2013 version of the HRRR used for this study uses version 3.4.1 of the Advanced Research version of the Weather Research and Forecasting (WRF-ARW) model with Thompson microphysics. Initial fields are created using 3D-VAR data assimilation. As of April 2013, latent heating profiles are calculated as a function of radar reflectivity, which is assimilated at 3 km resolution every 15 min (Benjamin et al., 2013). The assimilated reflectivity data come from the same NWS Weather Surveillance Radar 1988 Doppler (WSR-88D) network that is used to create the Stage IV reference data set.

The HRRR model has been in development and running experimentally for the last several years and transitioned to operational status in September 2014. This study uses the 3 km two-dimensional hourly data set in Grib2 format. The data fields being used are the geolocation information (grid latitude and longitude) and the precipitation accumulated over the forecast hour (FH). Additionally, forecast environmental variables related to precipitation such as total precipitable water (TPW), cloud base and top heights, 18 dBZ echo top heights, CAPE, and CIN were obtained to investigate model performance as a function of the near-storm environment. Model output for the 2013 warm season (May–August) was obtained for forecasts initialized every 6 h (00, 06, 12, and 18 UTC). The data were obtained after the inclusion of radar reflectivity began in April 2013, but prior to a major upgrade that was completed in May 2014.

An expectation of new or upgraded NWP models is that they outperform their predecessors, so HRRR QPF performance was compared to that from the Global Forecast System (GFS) model. GFS forecasts of accumulated rainfall at the first three hourly time step (i.e., 3 h after initialization) at half-degree grid resolution were obtained from the NOAA National Operational Model Archive and Distribution System (NOMADS) for the same temporal range and initialization times as the HRRR. Both the HRRR and Stage IV hourly output were accumulated over 3 h and averaged from their native resolution to match the GFS resolution. Grid point comparisons between the two models and the observations were made over the central U.S. (29.0^{o}N– 49.0^{o}N, 84.0^{o}W–105.0^{o}W) for each forecast. We opted to make grid point comparisons and apply a threshold of 1 mm/3 h for a grid box to be considered to have nonzero rainfall in consideration of the fact that averaging the high-resolution HRRR and Stage IV data to 0.5^{o} resolution could cause the well-defined edges of precipitation features in the high-resolution products to be smoothed to a feature with a significantly different shape than that predicted by the GFS (e.g., a single raining high-resolution pixel in a 0.5^{o} grid box results in very small nonzero mean precipitation for that box). The use of this grid point comparison and the 1 mm accumulation threshold also avoids over-penalizing the GFS for observed precipitation features on the order of a few kilometers that would be too small for the lower-resolution model to produce. A 1 mm threshold still represents a relatively small amount of precipitation when considered over a 3 h accumulation period, and therefore likely captures the majority of measurable precipitation.

Frequency distributions of correlation coefficient, bias-corrected root-mean-square error (RMSE), probability of detection (POD), and false alarm ratio (FAR) for grid boxes meeting the three hourly accumulation threshold of 1 mm are shown in Figure 2.1, which shows the significant improvement over the lower resolution model that the HRRR provides. The HRRR tends to be far more correlated to the observations than does the GFS, and is rarely, if ever, negatively correlated to the observed rainfall. While the bias corrected RMSE values are similar between the two models, vast improvement can be seen in the POD and FAR for grid boxes exceeding the 1 mm/3 h rainfall threshold. While the GFS forecasts exhibit PODs ranging from 0

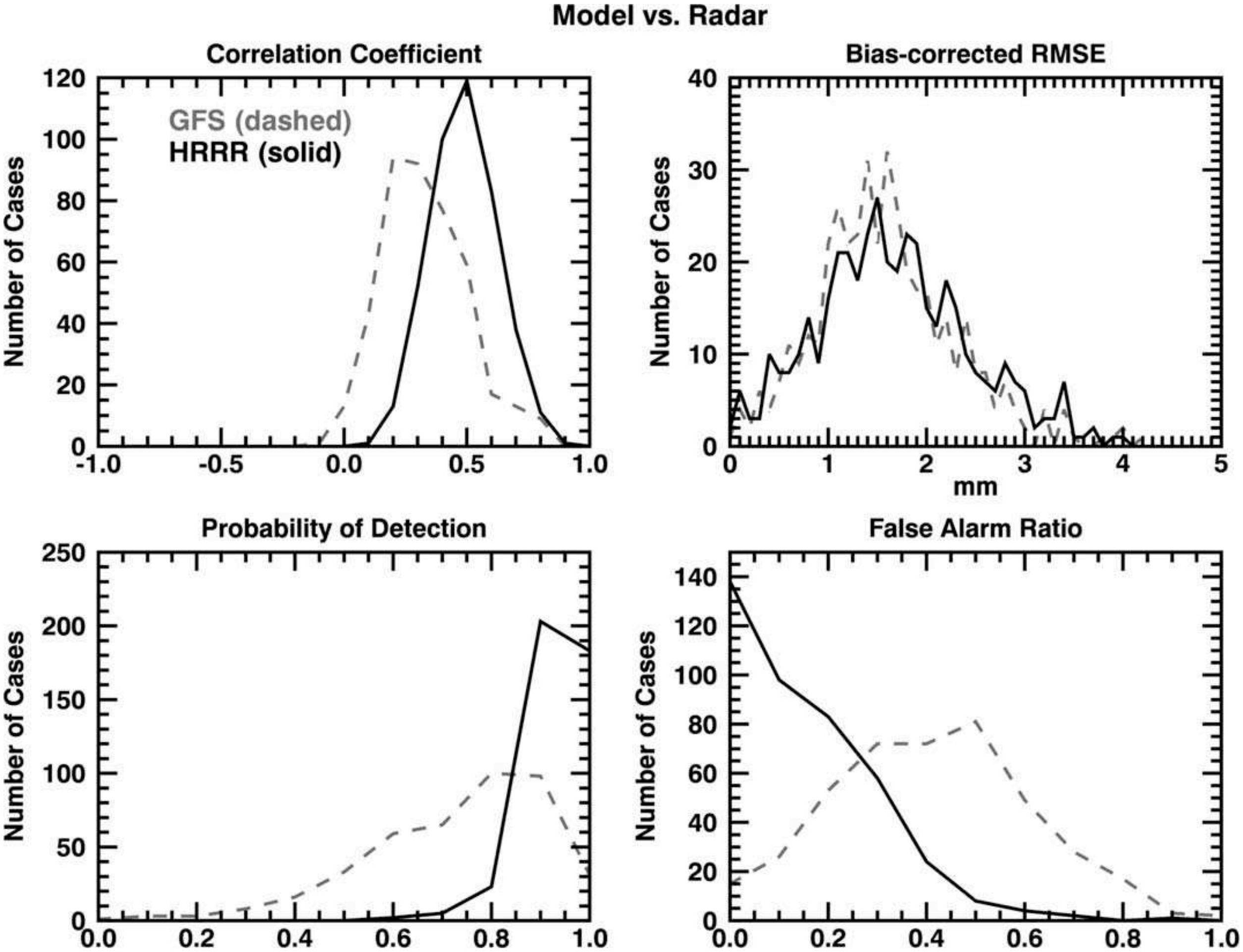

Figure 2.1. Comparison of validation statistics from the GFS and HRRR forecast models over the central US during the 2013 warm season.

to 1, HRRR forecasts rarely exhibit PODs less than 0.8 at this resolution. Similarly for FAR, the GFS forecasts often have FARs between 0.3 and 0.6, whereas the HRRR rarely has a FAR greater than 0.4. This seems to indicate that the averaged HRRR is creating precipitation in the correct locations, but does not provide any information as to why or under what circumstances the model performs well. Such questions can be better answered via object-oriented verification.

2.3 Methodology

To evaluate the performance of the HRRR model over its 15 h forecast period, model output from the 2013 warm season (May–August) at 6 h intervals (00, 06, 12, and 18 UTC initialization) was selected. The warm season was chosen to evaluate HRRR performance in long-lived convective situations, particularly MCSs, which are responsible for a significant portion of warm-season rainfall in the central United States (Fritsch et al., 1986). To focus on assessment of the model's performance with assimilated radar reflectivities, the evaluation is performed only when the Stage IV product indicated the presence of convective precipitation (i.e., conditional on observed convection).

Direct comparisons between the HRRR and the Stage IV require that they be available at the same resolution. Temporally this is not a problem, since both products are available hourly. In order to make comparisons at the same spatial resolution, both data sets were linearly averaged to a 0.05^{o} (~5 km) grid. This places both data sets on the same frame of reference with very little degradation of the spatial resolution and thus should maintain most of the variability within the observed and modeled precipitating features. Additionally, this slight degradation in resolution is not expected to have a large effect on the results for the mesoscale features of interest to this study.

The first step to the object-oriented validation scheme is to identify individual raining objects. As discussed, there are many different ways to accomplish this task, and the current paper follows the MODE methodology described by Davis et al. (2006 and 2009). There are two main steps to identifying raining objects using this method: (1) apply a spatial smoothing to the raining field and (2) create a binary mask of grid boxes where the smoothed rain field exceeds a given threshold to ''mask out'' precipitating objects. This methodology was selected for several reasons. First, the spatial smoothing of the rain field results in more contiguous raining features, and small areas of rainfall that may be detached from but are associated with a larger raining field would likely be incorporated with the parent feature as a result of the smoothing. Second, the binary mask is designed to allow for consideration of different rainfall thresholds. Therefore, the model can be examined from a variety of perspectives (e.g., performance with respect to the entire feature versus performance with respect to just strong convective cores). An example of smoothed rainfall and binary masks for a variety of masking thresholds is shown in Figure 2.2, with smoothing performed over three grid boxes in each direction (approximately 15 km). As expected, the lower thresholds result in larger raining areas and more contiguous features, whereas higher thresholds focus attention on convective centers.

The binary masks are used to identify individual raining features in both the radar and model data, and the smoothed rainfall maps are used to calculate properties of each identified feature within the selected domain (the central U.S., between 29.0^{o}N and 49.0^{o}N and 84.0^{o}W and 105^{o}W). This region was selected to avoid complications caused by coastlines and mountainous areas, and to ensure the best radar coverage for validation (Maddox et al., 2002). Because we are interested in evaluating model performance given the assimilation of precipitation-affected observations, we focus on preexisting, possibly mature, long-lived convection.

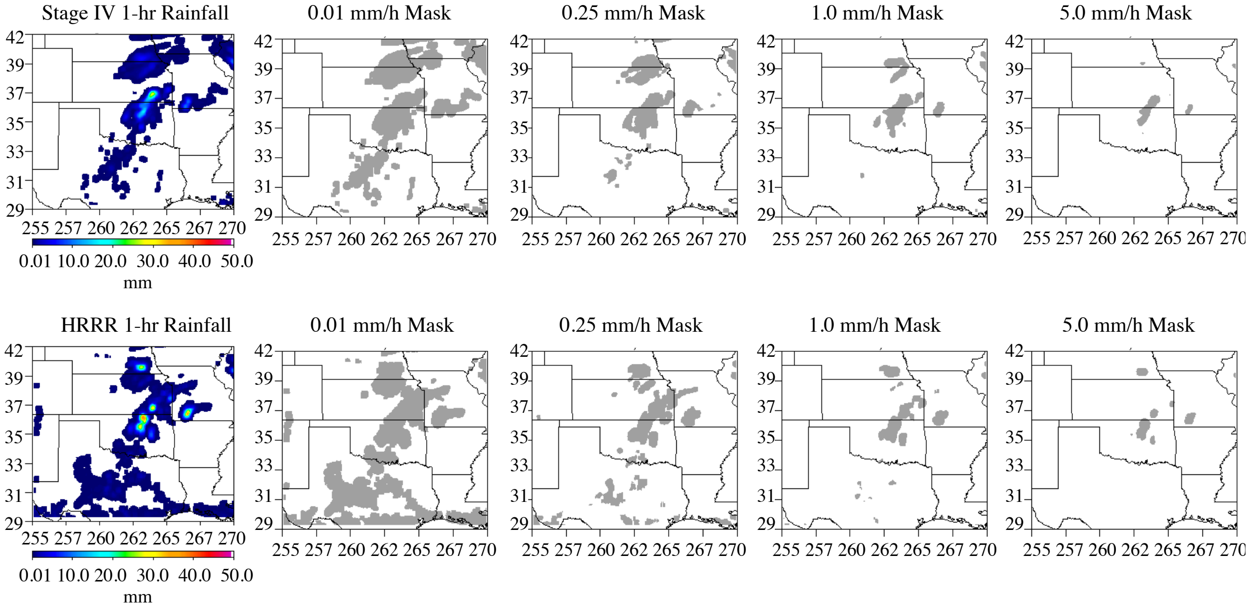

Figure 2.2. Example of (top) Stage IV and (bottom) HRRR rainfall fields with the ~15km smoothing applied, and the raining areas identified in each scene when 0.01, 0.25, 1.0, and 5.0 mm/h thresholds are applied.

Features to be considered for the assessment were selected from those identified in the radar fields using the following criteria:

1. Feature is present on the radar 1 h prior to forecast initialization (i.e., FH-1), ensuring that the feature was assimilated into the model initial state.
2. The maximum observed rain rate 1 h prior to forecast initialization is ≥10 mm/h, ensuring likely convection.
3. The feature identified in the binary mask has an areal extent ≥5000 km2. This should capture all but the smallest convective cells.
4. Feature is trackable for at least 70% of the forecast run (10 of the 15 h), allowing the evaluation of model performance over time.

The first two criteria are checked 1 h before forecast initialization, while the third criterion is checked at every forecast time step. Fulfillment of the final criteria is determined after all of the radar features have been identified at each time step and several attributes of each feature have been calculated and used to track the feature through time. The only requirement for the identification of a model forecast feature is that the smoothed raining field exceeds the selected masking threshold (that is, it is present on the binary mask). The binary masks, along with the unsmoothed rainfall maps are used to calculate properties of each identified feature within the study domain at each forecast hour. Properties calculated for each identified feature in the observations and forecast include

1. Bounding latitudes and longitudes of the feature.
2. Latitude and longitude of the feature center of mass.
3. Area of the feature exceeding the selected rainfall threshold.
4. Total, mean, and maximum rainfall.
5. Variance of the distribution of rainfall within the feature.

Using these attributes, the identified features were then tracked through the duration of the forecast. This was done in a manner similar to the Thunderstorm Identification, Tracking, Analysis and Nowcasting (TITAN) algorithm (Dixon and Wiener, 1993), using storm centroid and other properties to search within an expected ‘‘travel distance’’ for similar storms. It differs from TITAN in that it uses 2-D data rather than 3-D, and a simple decision tree rather than optimization. Starting at FH1 (FH-1 for radar observations), the algorithm searches for a feature in the next forecast hour with a center of mass within the effective radius of the current hour’s feature multiplied by 1.25. The effective radius of a feature is defined as the radius of a circle having the same surface area as that feature (Ebert and McBride, 2000). The multiplier of 1.25 was selected because the features of interest in this study generally have effective radii on the order of 100 km or more and propagate at speeds of less than 120 km/h (approximately 75 mph).

This also eliminates the chance of penalizing the model for an extremely fast-moving event that it may not have the physics to forecast correctly. This multiplier worked quite well for tracking observed features, but tended to result in model features that were very short lived. This was found to be a result of the low bias in the forecast feature size at FH1. Therefore, when tracking forecast features, the search radius was either 1.25 times the feature effective radius, or 105 km, whichever was greater. If more than one feature was identified within the search radius, the one with the most similar total rainfall to the current hour's feature was selected. This process is continued until no feature is found within the search radius, resulting in a time series of each observed and forecast feature.

If a radar feature was found to exist for 12 or more hours (1 h prior to initialization through 10 h of the forecast), then the search for a matching forecast feature is performed. This process attempts to replicate what a human observer might do if tasked with comparing two rainfall maps and is illustrated in Figure 2.3. First, using the binary masks for each rainfall feature, the algorithm determines if any of the forecast features overlap the observed feature (i.e., location of forecast rainfall is at least partially correct). If more than one forecast precipitation object overlaps the observed feature, the forecast feature overlapping the highest percentage of the observed precipitation is selected as a match.

If precipitation is observed but not forecast in a given location, the algorithm begins searching for a forecast precipitation object in the vicinity of the observed rainfall using a search radius of 2.0 times the observed feature effective radius. Much like the threshold used to create the binary mask, this multiplier can be increased or decreased to assess forecast location of convective precipitation. For example, a smaller search radius would be useful for assessing only those forecasts with very accurate placement of convective cores, whereas a larger search radius

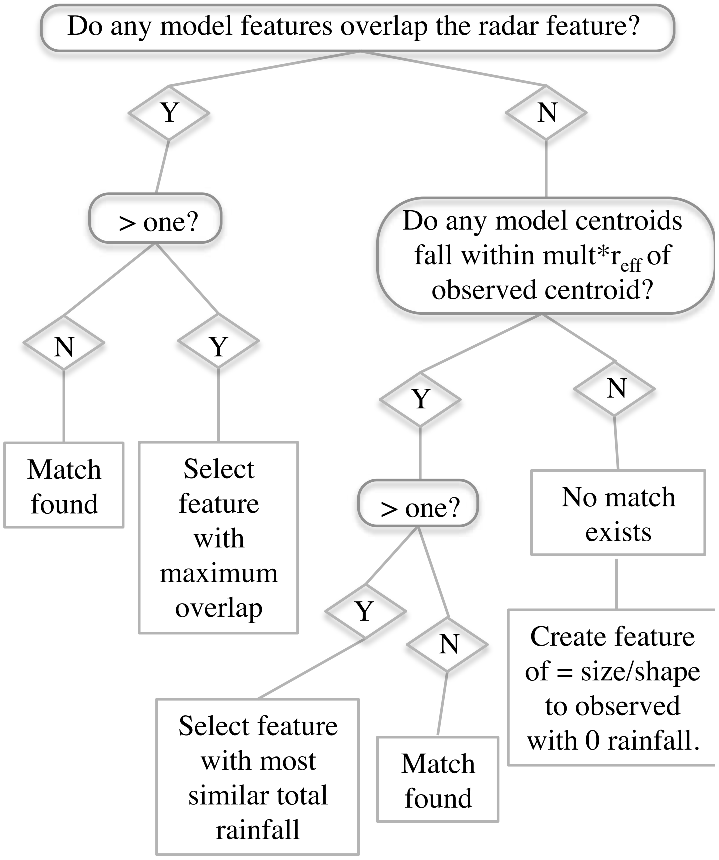
Do any model features overlap the radar feature?
Y
N
> one?
Do any model centroids fall within mult*r_{eff} of observed centroid?
N
Y
Y
N
Match found
Select feature with maximum overlap
> one?
No match exists
Y
N
Create feature of = size/shape to observed with 0 rainfall.
Select feature with most similar total rainfall
Match found

Figure 2.3. Flowchart illustrating the process by which a forecast precipitation feature is matched to an observed precipitation feature.

allows for evaluation of the model when it produces precipitation, but may not accurately predict the location.

If multiple forecast objects are present within the search radius, the feature with the most similar total rainfall to the observed is selected as a match. If there are no forecast objects within the search radius, the model is considered to have missed this rainfall event. Because we have tracked the observed and forecast features through the entire forecast period, we need only to identify a match at FH1 in order to evaluate how well the model's tracked feature represents the observed storm behavior through time.

The result of this methodology is a database of observed precipitation features and the model forecast for that feature, along with descriptive properties of each, through 15 h of the forecast. This database can be used to perform forecast verification either by comparing these properties or by calculating additional verification statistics. To avoid skewed results, if two or more features merged within the first 5 h of the forecast period, only the feature with the highest quality forecast match at FH 1 was considered for verification (i.e., after the hour of the merge, validation statistics would be identical for what were multiple features at FH 1). Most often, merging occurred within the first 2–3 h. The ability to consider multiple forecast features as potential matches to a single observed feature was not considered here.

2.4 Results

As shown in Figure 2.2, smoothing and masking the data following MODE (Davis et al., 2006 and 2009) successfully identifies precipitating areas at a variety of thresholds. While lower thresholds result in the identification of a large number of features and capture the vast majority of the precipitating area in both the observations and the model, the higher thresholds result in fewer identified areas of precipitation of smaller size, capturing only the heaviest precipitation. While the overlap and effective radius criteria for determining a match between the forecast and the model is relatively straightforward at lower masking thresholds, at higher thresholds significant feature overlap becomes less likely, and the effective radius of the identified feature decreases such that a higher multiplier would be needed to find a match. This complication has been noted by Davis et al. (2009), Ebert and Gallus (2009), and Wernli et al. (2008). All results presented in this paper were obtained using a 1 mm/h threshold.

Over the 2013 warm season, 467 observed precipitating features that met the defined criteria were identified in the region of interest. Comparisons between individual radar and model precipitation features were made and stratified by month, time of day, and the size and intensity of the observed system; however, the most robust indicators of overall model behavior were obtained when considering all of the identified features collectively. A variety of validation statistics were considered, including direct comparisons between forecast and observed feature mean rain rate, maximum intensity, total rainfall, and areal extent, as well as the biases in each of these feature properties. Location error, correlation between the forecast and observed features, POD, FAR, and RMSE were also calculated.

Here we will discuss the model performance over the duration of the 15 h forecast and attempt to relate these validation results to processes within the model itself. In the interest of space, statistical behavior will only be shown for every-other time step, and only through FH 11. While statistics were calculated for all 15 forecast time steps, the requirement that observed features last for at least 10 h means that later forecast time steps often lacked a feature to validate against, and therefore such statistics are less robust than those obtained earlier in the forecast. Many of the following discussions are based on results seen when examining the PDF of a given statistic over the forecast period. Emphasis will be given to results from early in the forecast when the influence of the assimilated radar data is the strongest.

2.4.1 Placement

As one would expect, the HRRR forecast places the precipitation features in very similar locations to where precipitation was actually observed, particularly at FH1. Figure 2.4a (Figure 2.4b) shows PDFs of the displacement of the HRRR forecast feature centroid from the observed in the east-west (north-south) direction, with negative values indicating that the forecast feature centroid is too far west (north).

In Figure 2.4a, at FH1, forecast feature centroids are generally located within 65 km from the observed, with 50% of feature centroids falling within 30 km (assuming 1^{o} longitude is approximately 85 km at 40^{o}N). At FH3, the majority of features still have centers of mass within

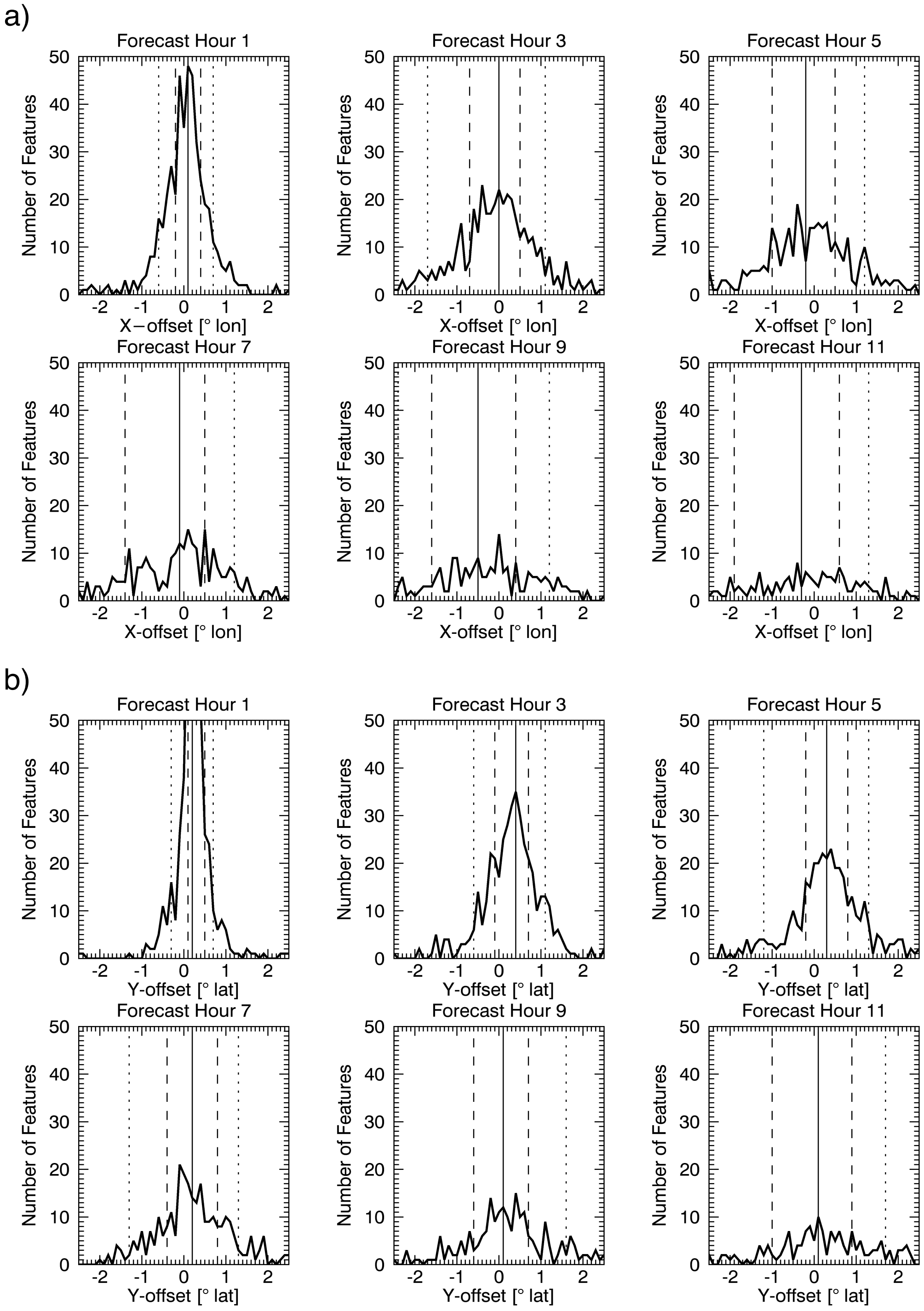

Figure 2.4. Distribution of the displacement of the forecast precipitation centroid from the observed precipitation centroid for every other hour over the first 11 h of the forecast in the (a) east-west direction and (b) north-south direction, where a negative offset indicates the model precipitation is offset toward the west and north, respectively. Dotted vertical lines represent the 10th and 90th percentiles, dashed vertical lines the 25th and 75th percentiles, and the solid vertical lines the 50th percentile.

80 km of the observed. However, as the forecast progresses, the median of the distribution shifts toward a more westward offset. In the north-south direction (Figure 2.4b), the offset distribution is narrower than in the east-west direction, with 80% of features falling within 65 km of the observed. There is a tendency for forecast feature centroids to be placed slightly too far to the north in early forecast hours, with placement improving somewhat through the duration of the forecast.

These results suggest that the model has some tendency to propagate precipitation too slowly in the east-west direction (assuming eastward propagation, which is typical in this region). This is possibly related to the model's forecast easterly wind component being somewhat too weak. The reason for the model's tendency to place feature centroids somewhat south of their observed position is more difficult to postulate, as eastward propagating systems in the central U.S. can take on both northerly and southerly motion components. Discussion with the model developers (C. Alexander et al., personal communication, 2015) indicated that the shift from a northerly offset to near 0^{o} or slightly southerly was not surprising, as during the Northern Hemisphere summer, more conducive air masses with more available latent heat to sustain or support further convective development are generally located to the south.

It should be noted that due to the design of the feature-matching algorithm, smaller systems require a closer centroid match, so a very large centroid offset implies a relatively large feature. This offset value only considers the difference between the centers of mass of the forecast and observed precipitating objects, and so for larger features a large offset could indicate misplacement of convective cores within an overall larger system, and not generally bad placement of the feature itself.

2.4.2 Spin-up and biases

Generally, forecasts tend to be of highest quality around FH3, with consistently smaller absolute biases in feature mean and maximum rainfall and areal extent, as shown by the 0-lag bias distributions in Figure 2.5 (black solid curves). While the bias in feature raining area shows a more broad distribution at FH3 than at FH1 (Figure 2.5a), the distribution shifts from strongly negative biases at FH1 to a distribution more evenly centered around 0% at FH3. With regard to intensity, in Figures 2.5b and 2.5c, the median bias in feature mean hourly rainfall and maximum intensity is high at FH1, with relatively wide distributions. By FH3, both distributions have narrowed and exhibit median values near 0%. This indicates that even with the assimilation of radar reflectivity, the model still requires 1–2 h of spin-up before it most accurately produces rainfall that is statistically similar to the observed.

Discussion with the model developers (C. Alexander et al., personal communication, 2015) suggested that biases early in the forecast may also be related to the model's reliance on reflectivity information acquired up to 1 h prior to model initialization. Therefore, evaluation of biases in forecast feature mean, maximum, and total rainfall as well as feature areal extent was performed at temporal lags of 0, 1, and 2 h (that is, comparing the forecast feature to the assimilated features). Figure 2.5 shows PDFs of these biases at FH1 and FH3 with the associated results at lags of 1 (pink dashed) and 2 (blue dash-dotted) h with median values of the distributions shown by vertical lines. As discussed, the model exhibits overall low biases in the feature areal extent at FH1, with significant improvement by FH3. When comparing to the observed feature raining area from 2 h prior, we see that the median bias is nearly 0% (Figure 2.5a), indicating that the model is not significantly changing the size of the precipitating feature

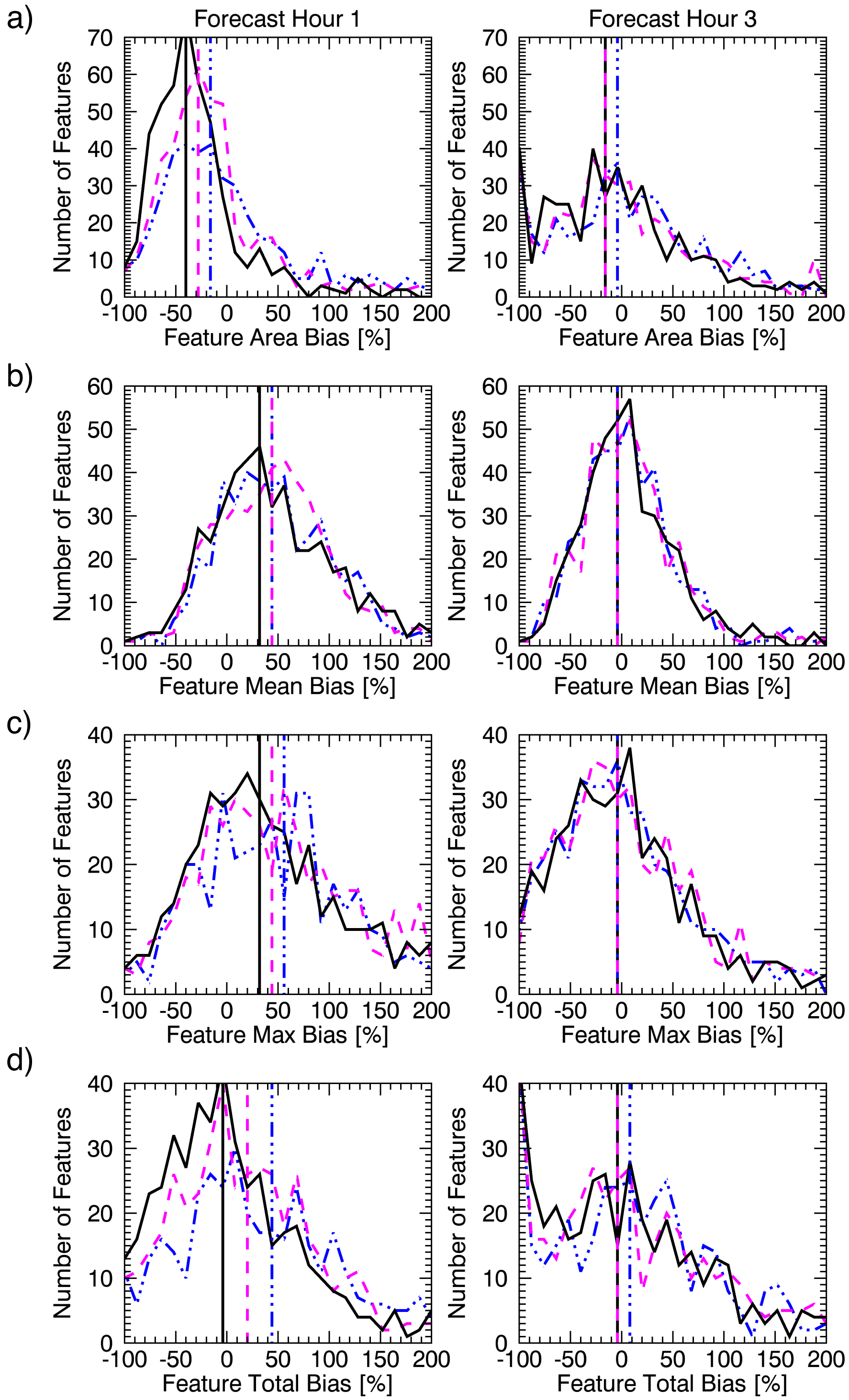

Figure 2.5. Distribution of the biases in (a) feature areal extent, (b) mean hourly rainfall, (c) maximum rainfall intensity, and (d) feature total rainfall at forecast hours (left) 1 and (right) 3 for temporal lags of 0 (black solid), 1 (pink dashed), and 2 (blue dash-dot) h. Vertical lines indicate median values of each distribution.

between assimilation time and FH1. By FH3, the median feature developing in the model is much closer in size to what was observed, with a median bias similar to that observed 1 h prior.

With respect to the statistics indicating rainfall intensity, the best results are seen at a lag of 0h. Forecast feature mean and maximum rainfall are 30-40% higher than observed at FH1, with positive biases of 40–50% with respect to the 1 and 2 h lags (Figures 2.5b and 2.5c). Meanwhile, the bias in forecast feature total rainfall at FH1 is nearly 0% with no temporal lag, with high biases compared to the 1–2 h time lags. These results indicate that although the model does not appear to be altering the size of the assimilated precipitation feature much, there is some change in the strength of the latent heating profile being made between initialization and FH1.

These results suggest that the model has a general tendency to create rainfall that is more intense than observed over a smaller area than observed, creating a situation where the total accumulated rainfall is relatively accurate, but incorrectly distributed (similar results have been seen by Lean et al. (2008)). This concept is illustrated in Figure 2.6, which shows a composite of all 467 observed/forecast feature pairs at FH1 and FH3. Figure 2.6 illustrates not only the tendency for the model to concentrate very heavy rainfall in the convective cores, but also the tendency of improvement over the first few forecast hours, as indicated by the bias comparison given in Table 2.1. Figure 2.7 displays an example of this behavior during the 2013 warm season.

The increased similarity in the biases at all three lag time steps between FH1 and FH3 is not surprising considering the types of precipitating systems considered in this study. Such mesoscale systems with required lifetimes of 12+ h are likely in the developing stages when observed by the radar up to 2 h prior to model initialization, and are likely undergoing many

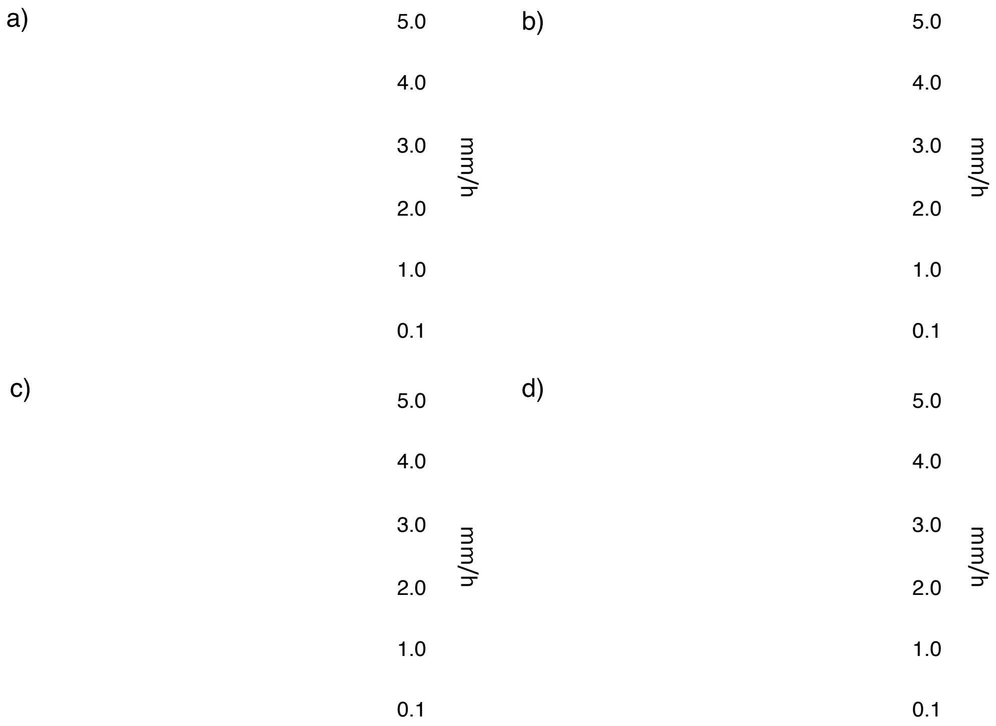

Figure 2.6. Composite of (a and c) observed and (b and d) forecast precipitation features at forecast hours (top) 1 and (bottom) 3.

Table 2.1. Bias statistics of composite observed and forecast rainfall features shown in figure 2.6 at forecast hours 1 and 3.

	Forecast Hour 1	Forecast Hour 3
Area Bias	-67%	-22%
Mean Bias	+61%	+25%
Max Bias	+303%	+125%
Total Bias	+8%	+2%

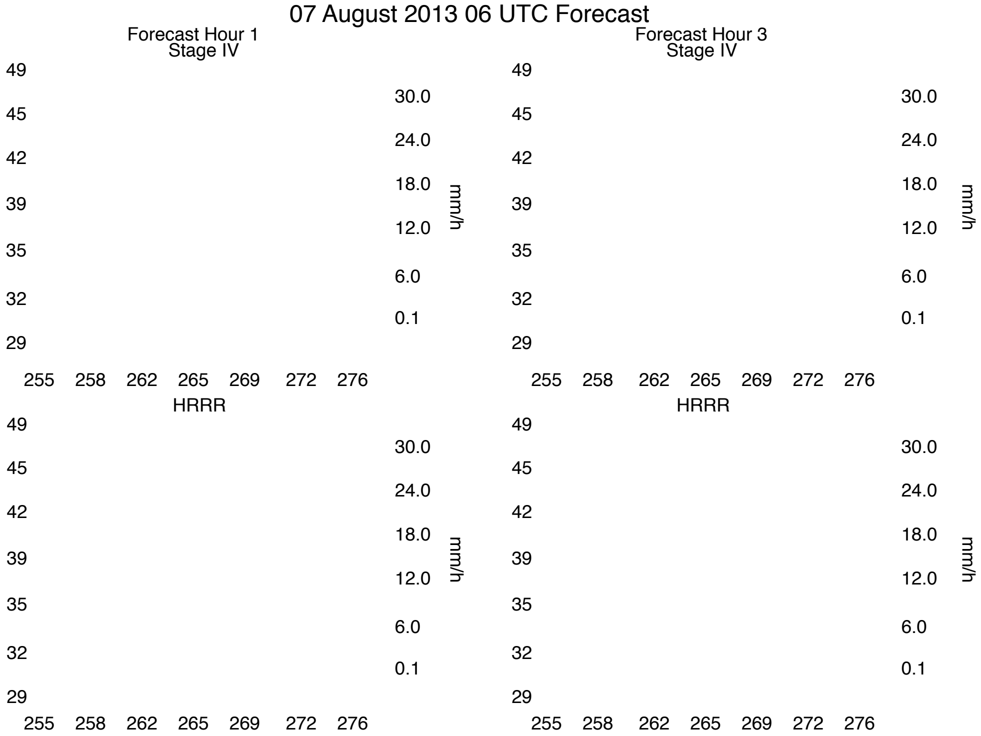

Figure 2.7. Observed and forecast accumulated rainfall for 7 August 2013 for hours 1 and 3 of the 06 UTC HRRR forecast run with validated precipitating feature over Lake Michigan indicated by grey outline. This case represents a typical occurrence in the model, with high maximum and mean rainfall biases and low area bias in hour 1 and improved representation of the observed feature by FH3.

changes in size and intensity during the assimilation period. A few hours into the forecast, these systems have likely reached their mature stage and changes in size and intensity are smaller, less frequent, and less rapid. Therefore, comparisons between size and intensity a few hours apart might not show much difference. This similarity in lag results as the forecast progresses is not seen in the location comparisons (not shown), in which the lag 0 feature is always to the east of the features from 1 to 2 h prior, indicating appropriate propagation of the assimilated reflectivities.

2.4.3 Probability of Detection

Thus, far we have shown that the model generally predicts accurately located systems with low biases in areal coverage and a tendency to concentrate moderate to heavy rainfall in convective cores, whereas the observed features generally have more moderate rain rates over a larger area. This is further supported when examining the POD (Figure 2.8). With the assimilation of radar reflectivities and the good location statistics, one would expect high POD values, particularly early in the forecast. However, we see that at FH 1, the HRRR has a POD of 0.65 or less 75% of the time, with a similar distribution at FH3, and decreasing median values through the remainder of the forecast period. This seems to suggest that one or a combination of factors in the model will not allow for the production of light to moderate precipitation, even when reflectivities indicating its presence have been assimilated.

To test this theory, several model variables related to precipitation were chosen, and the probability of precipitation (POP) as a function each variable was calculated for a variety of precipitation thresholds, given that radar indicated rainfall of 1 mm/h or more was present. The

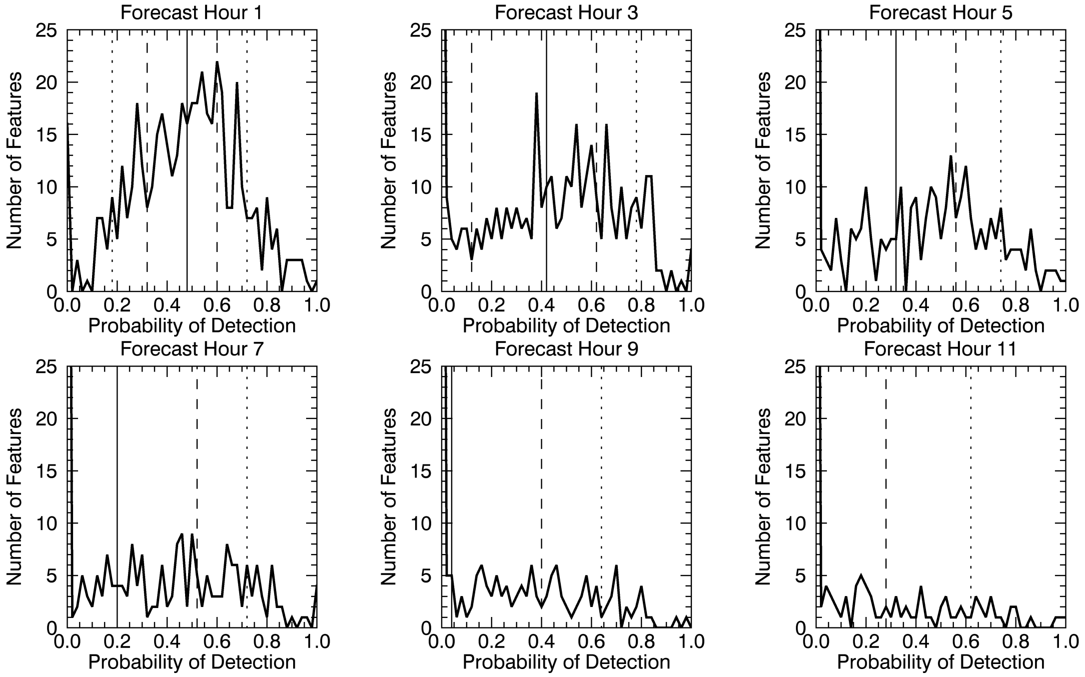

Figure 2.8. Same as figure 2.1, but for Probability of Detection.

environmental variables considered included TPW, cloud top and base heights, forecast 18 dBZ echo top height, CAPE, CIN, and the observed rain rate, with results from the variables that showed a strong relationship to precipitation production shown in Figure 2.9 for all forecast hours. Figure 2.9a shows that, given an observed rain rate of 1 mm/h or greater, there is an 80% or higher probability that the model will produce at least some rainfall. As the thresholds for forecast precipitation increase, however, there is a significant decrease in POP. In fact, regardless of observed rain rate, there is less than a 50% chance that the model will forecast hourly rainfall of 1 mm or more. This could partially explain the relatively low area bias we see using a 1 mm/h threshold. Figures 2.9b–2.9d provide some insight as to why this might occur, indicating that while the HRRR can produce very light precipitation in environments forecast to have TPW

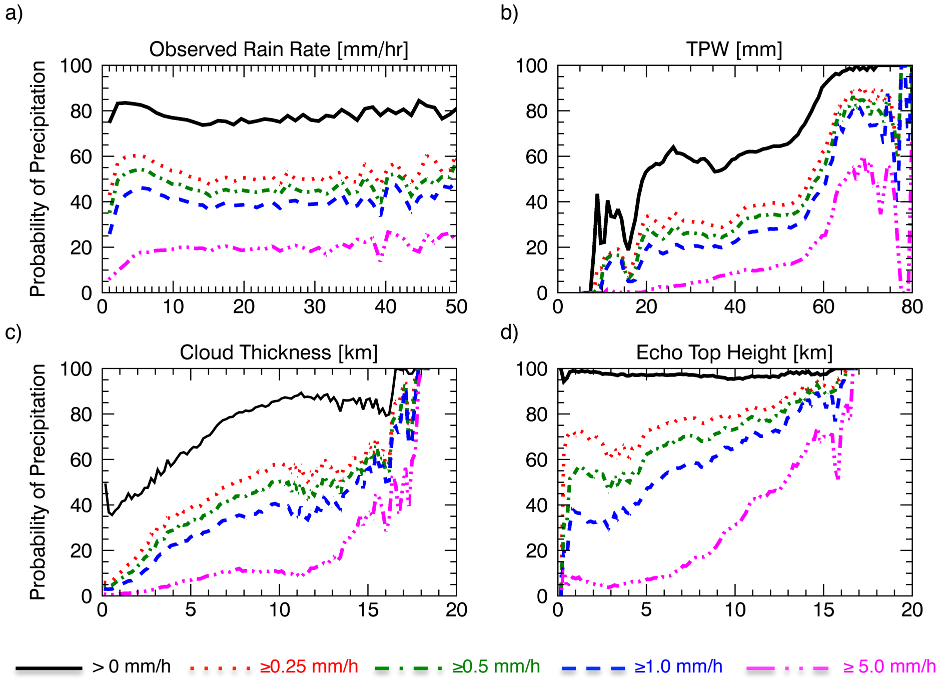

Figure 2.9. Probability of forecast precipitation rate exceeding a variety of thresholds given (a) observed rain rate and (b) the forecast precipitable water, (c) cloud thickness, and (d) 18 dBZ echo top height for all forecast hours.

between 25 and 45 mm, it does not reliably produce moderate to heavy precipitation (≥5 mm/h) without TPW values exceeding 50–60 mm. Such values, while possible over the study domain, especially the southeastern portion, are more typical of tropical environments and are not often seen in the northern high plains. Very high moisture contents in turn appear to be related to the depth of the storms being observed, as significant rainfall is most likely in the deepest forecast clouds with the highest forecast 18 dBZ echo top heights (Figures 2.9c and 2.9d).

Figure 2.10 shows the bias in feature mean forecast rainfall as a function of the mean TPW forecast for that feature considering only rainfall exceeding 0, 5, and 10 mm/h thresholds, with 0% bias and 40 mm TPW (the mean TPW forecast by HRRR in the identified precipitating features) indicated by red dashed lines for reference. While there is some degree of scatter, there is a clear tendency for the HRRR forecast to exhibit high biases in mean hourly rainfall at high values of feature-mean TPW. This trend appears even stronger when considering only the area within the 10 mm/h isohyet (Figure 2.10c), further supporting the connection between the production of moderate to heavy precipitation and the amount of column moisture the model produces.

2.5 Conclusions

This study builds on the features-based precipitation verification method described by Davis et al. (2006 and 2009), using a technique similar to MODE to identify precipitating features in both forecast and observations. Features are identified and tracked over the duration of the forecast period, enabling the evaluation of forecast model performance through time. Information about each identified feature in the domain from both the forecast and observations is stored in a database that can be used to perform direct comparisons between each matched feature pair, or to calculate additional statistics in order to assess the model’s performance.

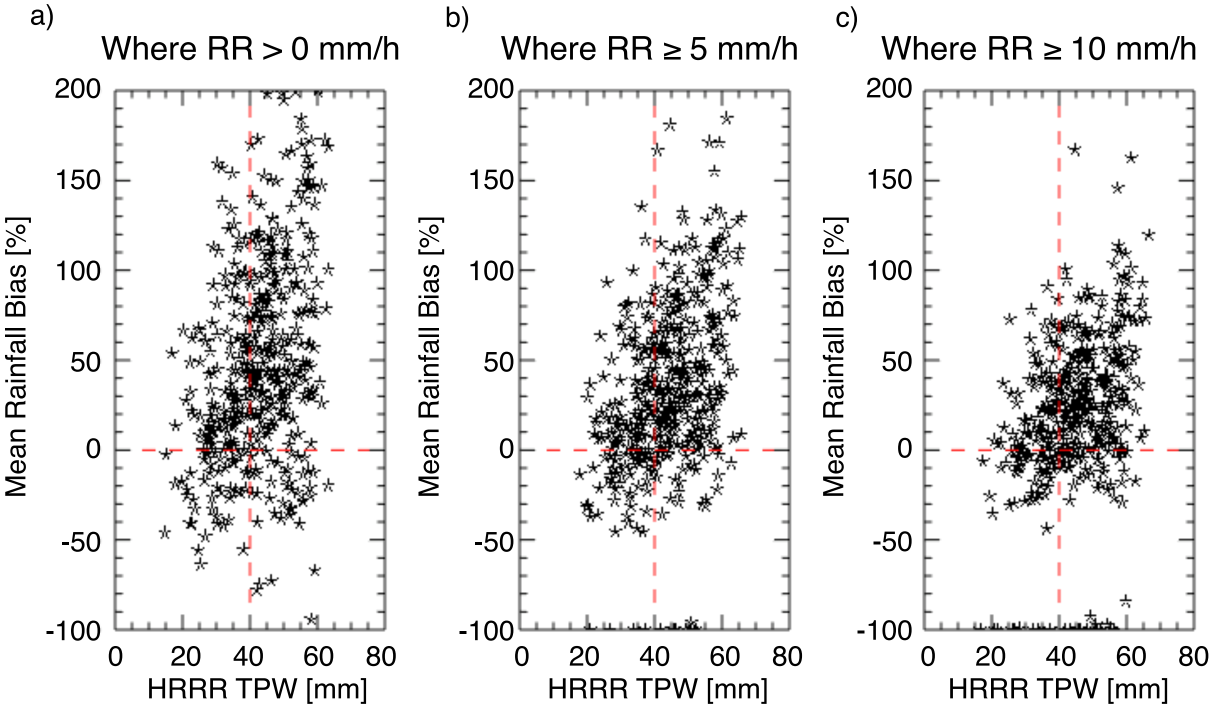

Figure 2.10. Mean rainfall biases as a function of feature average TPW for raining areas exceeding 0, 5, and 10 mm/h thresholds. Red dashed lines are overlaid at 0% bias and 40 mm TPW for reference.

While object-oriented verification methods such as this present many advantages over grid point validation methods, they are not guaranteed to work as desired every time. For example, Figure 2.11 shows a case where the model forecast captures the precipitation system extending from Wisconsin southwest through Oklahoma relatively accurately. The observations indicate a large swath of precipitation with many embedded areas of stronger precipitation, and the model indicates a similar feature that is somewhat wider in east-west extent and with more widespread heavy precipitation. As a result of the distribution of heavy rainfall within the forecast feature, the 1 mm threshold used in this study has masked out nearly the entire system, whereas the algorithm considers only the northernmost area of observed heavy precipitation as one of

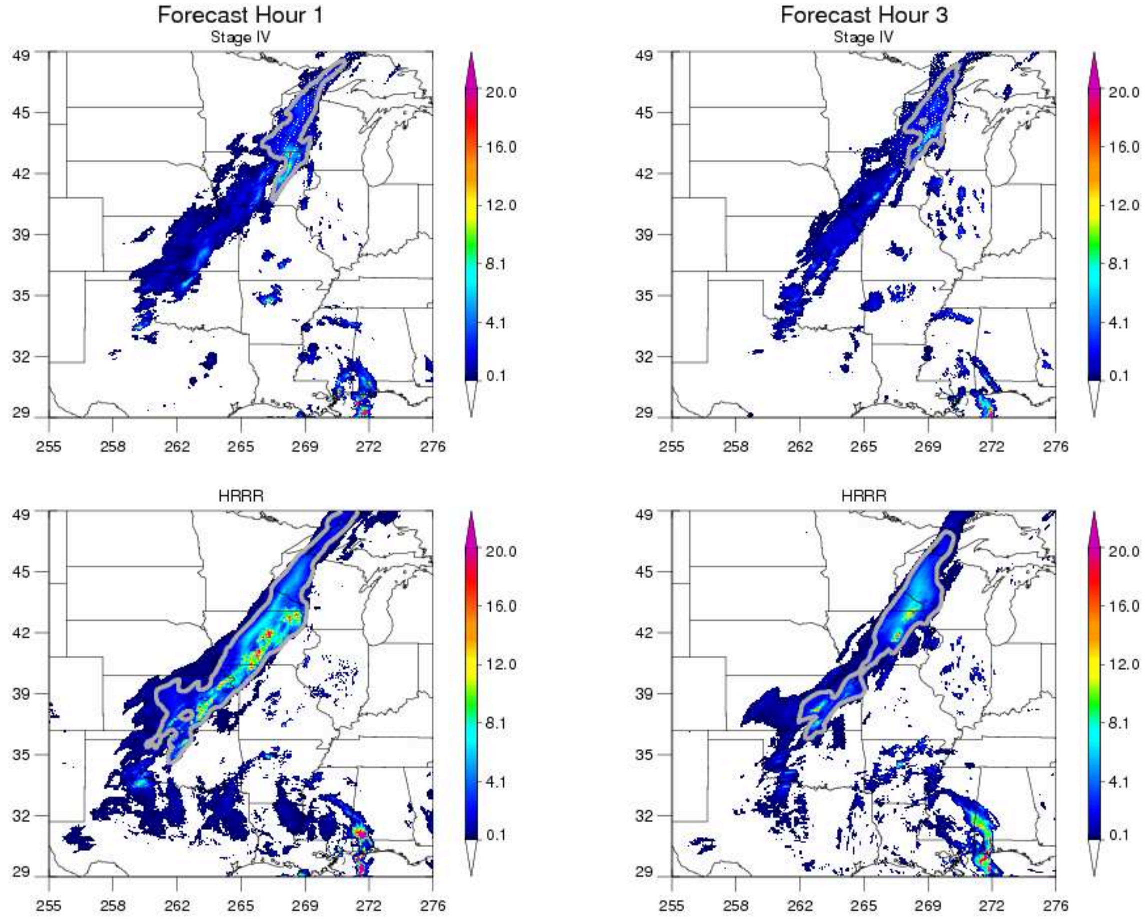

Figure 2.11. Same as Figure 2.7 but for 2 May 2013 12 UTC forecast.

multiple individual features within the larger line. This is a case wherein the forecast is not necessarily bad, but with extremely large biases in feature total rainfall and areal extent as well as very large differences between the observed and forecast feature centroids. Such occurrences would be more or less common depending on the chosen masking threshold and are one reason why we chose to highlight overall model behavior in this study.

The technique developed in this study was applied to evaluate the performance of HRRR forecasts of precipitation from large, long-lived MCS-like systems over the 2013 warm season and relate verification results to relevant model variables and processes. The results of this work indicate that the assimilation of radar does not eliminate the need for the model to spin-up before achieving its highest forecast quality. The HRRR tends to place the centroid of precipitating systems relatively well, usually within approximately 50 km of the observed precipitation, although there is some tendency for the HRRR to place features slightly west of their observed positions, particularly later in the forecast, suggesting winds that are not quite strong enough to accurately propagate precipitation features in the eastward direction. North-south placement is generally skewed toward feature centers of mass that are farther north than observed.

While the location of the precipitating systems' centers of mass is relatively well forecast by HRRR, the areal extent of region within the 1 mm/h isohyet is frequently too small, especially early in the forecast. Feature mean and maximum intensity is often biased high. The distribution of rainfall within the forecast precipitation features appears to be due to the lack of growth of the precipitation feature between assimilation and FH1 and the model's tendency to concentrate heavy precipitation in convective cores. While very light precipitation is frequently forecast by the model, moderate to heavy precipitation is generally not produced in environments with precipitable water below roughly 50 mm. Very high TPW values were seen to be necessary in order for the model to reliably produce rain rates greater than 1 mm/h, suggesting model processes that do not efficiently produce clouds and precipitation in the absence of significant moisture.

As the goal of this model assessment is to inform both the users and the developers of NWP models, the results of this study were shared with several members of the HRRR model development team. They indicated that our results were consistent with their expectation that the model would rely more heavily on the assimilated reflectivity to tell it where the rain is located, and thus the better statistics when comparing the areal extent to lagged observations. They were also aware of the model's tendency to produce small areas of very intense rainfall as well as precipitation features with lower cloud top heights than observed/expected. The 2014 version of the model employs a different reflectivity-latent heat relationship to attempt to mitigate these issues.

The results of this paper focused on validating the forecast model using ground-based radar as a reference data set. The methodology discussed herein focused on relating the performance of HRRR hourly rainfall forecasts to two-dimensional model variables. This process could easily be adapted to three-dimensional quantities, including assessments of forecast and observed three-dimensional reflectivity profiles that can be related to additional microphysical processes and assumptions within the model.

CHAPTER 3:

CONSISTENCY BETWEEN CONVECTION-ALLOWING MODEL OUTPUT AND SATELLITE OBSERVATIONS

3.1 Introduction

Recent advances in numerical weather prediction (NWP) models have resulted in increased model resolution, both spatially and temporally. In the United States, the National Centers for Environmental Prediction (NCEP) High Resolution Rapid Refresh (HRRR) model is run operationally every hour at 3km horizontal resolution, providing 18-hour forecasts over the continental United States (CONUS; Benjamin et al., 2016). The HRRR model aims to provide accurate forecasts of convective precipitation in the short term (on the order of a few hours). These forecasts are needed for a variety of uses with varying degrees of desired accuracy, including urban and rural water management, flood forecasting, and initialization of stream flow forecast models (Shrestha et al., 2013; Fritsch and Carbone 2004; Ebert et al., 2007). Accurate forecasts of convective precipitation are also necessary to meet the National Weather Service's (NWS) "warn-on-forecast" goals for such severe weather events as severe thunderstorms, tornadoes, and flash flooding (Stensrud et al., 2009).

The spatially discontinuous and rapidly varying nature of convective precipitation makes such short-term forecasts difficult (Shrestha et al., 2013; Sun et al., 2014). In fact, Lin et al. (2005) point out that on scales of a few hours, simple advection of radar echoes tends to provide a better forecast than NWP output. While increased model resolution and update frequency is expected to mitigate some of these challenges, many microphysical, radiative, and dynamic processes within convective systems occur on sub-grid scales (Shrestha et al., 2013; Bryan and

Morrison, 2012), and Burghardt et al. (2014) point out that further improvement of the representation of convective precipitation systems by NWP models will require more than simply increasing the model resolution to the order of a few 100 meters.

In addition to the difficulties caused by the behavior of convective precipitation systems themselves, such systems are also difficult to forecast due to the challenges of accurately representing precipitation in the initial forecast state and the need for NWP models to “spin-up” from their initial state (Sheng et al., 2006; Rogers et al., 2000; Errico et al., 2007; Sugimoto et al., 2009; Stensrud et al., 2009; Sun et al., 2014; Weygandt et al., 2007 and 2008). Including precipitation-effected observations in the initial model state with rapid updates can help to improve the representation of convective systems in the model analysis and reduce spin-up time, however ground-based radar is currently the only instrument available for frequent assimilation cycles of precipitation used in the HRRR model (Sun et al., 2014; Errico et al., 2007).

Studies by Sheng et al. (2006); Xaio and Sun (2007); Sugimoto et al. (2009); Sun et al. (2014); and Craig et al. (2012), among others, have shown improved short-range forecasts as a result of the assimilation of radar data. Direct assimilation of reflectivity to calculate rain rates is not particularly desirable, however, due to the non-linear relationship between the two. Instead, Sun et al. (2014) suggest diabatic assimilation, in which radar reflectivity is used to determine a perturbation in latent heating within the column. This methodology is preferred because the change in latent heating translates to other model variables, and introduction of the large uncertainty in the relationship between reflectivity and rain rate can be avoided. Diabatic assimilation of radar data has been employed in the HRRR model since 2013 (Benjamin et al., 2016).

With the increased resolution and assimilation of radar data into the HRRR, improved forecasts of precipitation amount can be expected. Such improvements can be documented through validation studies, yet the key to continued model advancement lies in understanding why the model does or does not perform well. As such, Bytheway and Kummerow (2015, hereafter BK15) suggest the idea of a model "assessment," a process by which the model forecast of a particular atmospheric variable is not only validated, but an explanation for successful or poor forecast results is also sought.

The key to improving forecasts of precipitation lies in understanding the physical and microphysical processes occurring within the precipitating cloud, and accurately representing those processes within the cloud resolving model (CRM). It is difficult to evaluate the accuracy of the CRM in-cloud processes, however, since direct observations of cloud processes are not common, typically being limited to aircraft measurements taken during field campaigns. Remote sensing observations are thus the next-best option for assessing CRM in-cloud behavior.

The launch of the GPM core satellite in early 2014 added a valuable source of observations of the 3-dimensional structure of storms. While ground-based radars are generally useful for identifying larger hydrometeors (such as rain drops, large snow-flakes, and hail), their data are not as suitable for discriminating between smaller particles (such as cloud drops vs. cloud ice). Additionally, ground-based radars tend to be less reliable in complex terrain, where beam blockage by orographic features can become an issue. The GPM core satellite carries the first Dual Frequency Precipitation Radar (DPR), which provides 3D hydrometeor storm characteristics and rainfall information, without being susceptible to interference from terrain and ground clutter. Collocated radiances measured by the GPM Microwave Imager (GMI) provide even more information about the storm structure, particularly with respect to ice content.

Used individually and in combination, observations from the instruments on the GPM core satellite can be used to infer both profiles and integrated quantities of liquid and ice as well as hydrometeor size distribution parameters. These properties can then be compared to those produced by the model, in an attempt to understand instances when the model produces storms that are inconsistent with the observed.

This study will aim to continue the work of BK15, evaluating model performance in the early hours of the forecast, and seeking to understand differences between forecast and observed precipitating objects. Differences between forecast hydrometeor profiles and those retrieved using GPM observations, and between GMI observed radiances and those simulated using HRRR hydrometeor and atmospheric output will be examined. As in BK15, we continue to focus on the warm season (June, July, August) convective precipitation. While reference data for model validation is widely available in the US Great Plains, the reliability of the GPM instruments in complex terrain provides an opportunity to assess HRRR QPFs in the mountainous western US (west of 105°W). This area is particularly susceptible to terrain-induced precipitation and flash flooding, and as such understanding model behavior with respect to QPFs in this area is critical to meeting the NWS's warn-on-forecast goals. Because GPM overpasses are generally available only twice daily at a given location, this study will focus on evaluating QPFs and CRM output in the first hour of the forecast only.

The current study will continue with the features-based methodology described by BK15, in which precipitating objects are identified in both the observations and the forecast, and then matched to one another. Biases with respect to feature mean rain rate, maximum rainfall intensity, raining area, and storm total rainfall are calculated, as are location offsets of feature centers of mass. Then, using profiles of temperature, pressure, and hydrometeor mixing ratios from the model, brightness temperatures (Tbs) and 3D reflectivity profiles are calculated and compared to the observed. Where large inconsistencies are found, they are examined further, and possible explanations for the differences are explored.

3.2 Data

3.2.1 High Resolution Rapid Refresh Model

The HRRR model has served as the NWS's operational rapid update forecast model since September 2014. It is a storm-resolving model running at 3km grid spacing with 50 vertical levels at native resolution, or 40 vertical pressure levels, output from the latter being used for this study. Its domain covers the CONUS and is nested within the 13km Rapid Refresh mesoscale model, from which the HRRR receives its boundary conditions (Benjamin et al., 2013). The HRRR model was designed to provide rapid update model guidance on convective storms in order to improve severe weather forecasting, air traffic management and aviation hazards forecasting, and warning dissemination (ESRL, 2015). Full details of the HRRR model are given in Benjamin et al. (2016).

This study makes use of both 2 and 3-dimensional output from the 2014 and 2015 experimental versions of the model. Because the model did not become operational until late 2014, only the experimental version was available for use in both years. Output variables of interest include profiles of temperature, dewpoint, winds, and mixing ratios for the five hydrometeor classes (cloud and rain water, cloud ice, snow, and graupel), as well as surface rain rate. While there were several upgrades made to the model in April 2015 (most notably upgrading from WRF-ARW v3.5.1 to v3.6 and upgrading to aerosol-aware microphysics (Thompson and Eidhammer, 2014)), the behaviors observed in the forecast precipitation discussed herein were evident in both years studied.

3.2.2 GPM Core Observatory

The GPM Core Observatory was launched on February 27, 2014 from Tanagashima Island, Japan, and represents a joint effort between the US National Aeronautics and Space Agency (NASA) and the Japan Aerospace Exploration Agency (JAXA). It flies in a non-sun-synchronous orbit with a 65-degree inclination and altitude of 407 km. The satellite carries a passive microwave imager and dual-frequency precipitation radar (DPR) operating at both Ka (35 GHz) and Ku (13.6 GHz) bands, the data from which can be used to retrieve atmospheric parameters of interest. These retrieved parameters, as well as observed reflectivity and Tbs are available as a variety of products from the NASA Precipitation Measurements Mission (PMM) website (https://pmm.nasa.gov/data-access). Those products relevant to the current research are described below.

3.2.2.1 GPROF Precipitation Estimates

The Goddard Profiling Algorithm (GPROF, (Kummerow et al., 1996; 2001; and 2015)) provides estimates of instantaneous precipitation rates using passive microwave radiances at a nominal resolution of approximately 5km. GPROF is a Bayesian-type algorithm that searches an a-priori database of precipitating profiles that result in simulated Tbs that are similar to the observed, and then calculates a precipitation rate using a weighted average of all similar profiles.

GPROF precipitation rates were used to identify reference precipitating features (i.e., observed features) and estimate the observed hourly precipitation (assuming consistent rain rates over the hour of satellite overpass). Only overpasses occurring within 15 minutes from the top of the forecast hour were used in this study in an attempt to ensure the observed/retrieved cloud properties would be most representative of those output by the model.

3.2.2.2 GMI Brightness Temperatures

The GPM Microwave Imager (GMI) is a conical scanning microwave radiometer featuring channels ranging from 10 to 183 GHz and earth incidence angle of 52.8 degrees. The GMI swath is 904 km (562 miles) wide, and is overlapped in the middle by both the Ka and Ku band radars. The 1CGMI product contains calibrated brightness temperatures from the GMI instrument.

This study focuses on comparing observed and simulated brightness temperatures at 89 GHz. This channel was selected for its relative insensitivity to surface parameters, such that assumptions about surface emissivity in the Tb simulations would not be as important as they might be at lower frequency channels, as well as for its sensitivity to the ice hydrometeors that are typically abundant in convective storms. GMI footprints at 89 GHz are 4.4km in the along-scan direction and 7.2km in the cross track direction.

3.2.2.3 Ku band Reflectivity

The Dual-frequency Precipitation Radar (DPR) carried by the GPM core observatory operates at both Ku and Ka band and provides profiles of reflectivity through the atmosphere with 5km horizontal resolution and 250m vertical resolution.

For this study we use only the KuPR, which is more sensitive to precipitation than the Ka band radar, and has a wider swath (245km compared to only 120km for KaPR). The wider swath makes it more likely that an observed precipitation feature will have been measured by both the GMI and the Ku radar, as well as more likely to have sampling over the majority of the feature. Reflectivities from the 2A_Ku product were used in this study for comparison of observed reflectivities with those simulated using the HRRR output hydrometeor profiles.

3.2.2.4 2B Combined Product

The 2B Combined product (2BCMB) combines observations from GMI and DPR to retrieve information about surface rainfall, profiles of cloud and rain water content, surface elevation, storm top height, precipitation type, and hydrometeor phase, as well as providing Tbs simulated using this information (Olson and Masunaga, 2016). The product is released at the same horizontal and vertical resolutions as the 2AKu product, with an additional data field including those footprints where the Ka-band radar was available. Products produced using only KuPR are used in this study. Retrieved profiles of cloud and rain water content were used for comparison with the forecast model output, and retrieved surface rain rate and simulated brightness temperatures were used as an additional reference estimate of those parameters, which were generally similar to those obtained using GPROF and GMI.

3.3 Methodology

In order to compare the CRM output to the observed and retrieved products of the GPM satellite in raining areas, precipitating features must be identified and matched between the two datasets. Once identified, the cloud and atmospheric information from the output is used to simulate 89GHz Tbs and Ku-band reflectivity profiles.

3.3.1 Feature Matching and Evaluation

As a first step to feature matching and evaluation, instances of the GPM core satellite passing over the study domain were identified. Only those cases where the overpass occurred within 15 minutes of the top of the hour were selected for study. By enforcing this criterion, it is likely that there would be minimal change in the feature structure between the satellite overpass and the valid forecast hour. Next, both the HRRR precipitation field and GPROF retrieved precipitation were linearly averaged onto a 0.05 degree (~5km) grid in order to place both datasets on the same frame of reference, while maintaining the majority of heterogeneity within the observed and forecast features. Since the HRRR is produced at 3km resolution, and GPROF is produced at approximately 5km resolution, this re-gridding serves as only a minor degradation in resolution.

Observed and forecast features were identified and matched in a manner similar to the Method of Object-based Diagnostic Evaluation (MODE, Davis et al., 2006 and 2009) as detailed in BK15. A 15km smoothing is applied to both the observed and forecast fields, with the forecast field being “masked” to include only those regions within the GPM Microwave Imager (GMI) swath. Precipitating objects are then defined as contiguous regions within the smoothed field having rain rates exceeding 0.5 mm/h and an aerial extent greater than 250 km^2. These are lower values than those selected by BK15 (1mm/h and 5000km^2, respectively), and were chosen because the convective precipitation features over the US mountain west are generally not as large or intense as those convective precipitation systems common over the US Great Plains, however these thresholds should omit the smallest, least mature features that the model would be unable to represent.

After precipitating objects are identified, descriptive statistics about each feature are stored in separate databases for the observed and forecast features, including bounding latitude and longitude, latitude and longitude of the feature center of mass, aerial extent, and feature total, mean, and maximum rainfall. These statistics were then used to search for a forecast precipitation feature to correspond to each observed feature. The matching algorithm first searches for forecast features that overlap observed features, selecting the forecast feature that overlaps the highest percentage of the observed feature as a match. If there are no forecast features collocated with an observed feature, the algorithm then searches for forecast precipitation in the vicinity, seeking features with centers of mass within a search radius of 2.0 times the effective radius of the observed feature (effective radius defined as the radius of a circle having the same area as the precipitating feature (Ebert and McBride 2000)). If any forecast precipitation features are identified within the search radius, the feature with the most similar total rainfall to the observed is selected as a match. If no forecast features are found that match an observed feature, the model is considered to have missed that rainfall event.

3.3.2 Simulated Reflectivity and Radiances

The HRRR model output contains three-dimensional fields of cloud water, rain water, cloud ice, graupel and snow mixing ratios which can be used in forward radiative transfer models to simulate reflectivity profiles and Tbs at frequencies matching those used on the GPM core satellite instruments. Reflectivity simulations were performed using the QuickBeam software (Haynes et al., 2007), while Tb simulations were performed using the slant-path plane-parallel Eddington approximation (Kummerow, 1993).

Care was taken in the simulations to match the microphysical properties as closely as possible to those used in the HRRR. This included using the mixing ratios of graupel and rain to calculate the intercepts (N_o) of those particle size distributions (PSDs), and the layer temperature and snow mixing ratio to determine the number concentration of snow particles, as described in Thompson et al. (2008). Additional properties, such the slope and shape parameters of each species' PSDs and particle densities were also given in Thompson et al. (2008), or culled directly from the WRF v3.6.1 code. Surface emissivity for the Tb calculations was assumed to be 0.95, a reasonable assumption over land surfaces that will generally affect the simulated Tbs at lower frequency radiometer channels (e.g. 10 and 19 GHz) much more strongly than those at higher frequencies.

While the HRRR model operates with double moment microphysics for rain hydrometeors (Benjamin et al., 2016), the hydrometeor number concentrations were not included in the model output at the time of this writing. Therefore, with the exception of snow number concentrations, which are calculated based on mixing ratio and temperature, (Thompson et al., 2008) the Tb and reflectivity simulations were performed using single moment microphysics. Since we are focusing on convective storms that are expected to have at least some ice in the column, it is probable that the ice scattering signal will dominate any uncertainties in the simulated radiances caused by the lack of explicit rain drop size distributions. The simulated reflectivities will be more sensitive to the assumed PSD and therefore only general rather than direct comparisons between simulated and observed reflectivity profiles will be used to infer inconsistencies between the modeled and observed storms.

3.4. Results and Discussion

Over the course of the 2014 and 2015 warm seasons in the western US, 585 observed and forecast feature pairs were identified in the first forecast hour. Comparisons were made between the storm center of mass, mean and maximum rainfall intensity, total hourly rainfall, and areal extent, with results shown in figure 3.1. Overall, when compared to retrieved rainfall from GPROF, the HRRR tends to place the center of mass of raining features within 20-30 km of the observed. Larger offsets, while possible, generally indicate larger features with the forecast and observed features having a different distribution of rainfall within the feature. With respect to intensity and areal extent of the forecast features, the HRRR model tends to create smaller features that are less intense than those retrieved by the GPROF algorithm. These results were consistent with those obtained when comparing the HRRR features to those identified in the NCEP Stage IV radar product (Lin and Mitchell, 2005) in the same region (not shown). Within

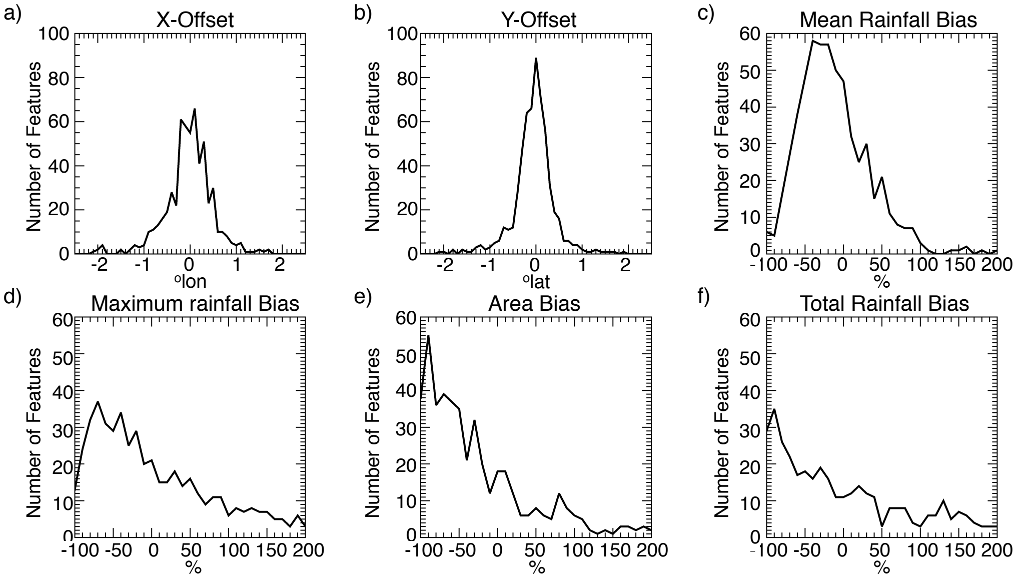

Figure 3.1. Displacement of forecast storm center of mass from the observations in the a) east-west and b) north-south direction and forecast biases in c) mean and d) maximum rainfall intensity, e) areal extent, and f) storm total rainfall.

the scope of this study, it is possible that the inconsistencies between the observed and forecast cloud structure could at least partially explain some of these biases, which will be examined further in the next sub-sections.

3.4.1 Observed and Simulated Tbs

Brightness temperatures in the microwave frequencies can be used to infer properties of the observed storms, for example, Tbs at 18 and 23 GHz can be used to infer integrated water vapor in the atmosphere, while those at higher frequencies (37 and 89 GHz) can be used to infer the presence of scattering by ice particles. In this study we will focus on the Tbs at 89GHz, a channel that is both sensitive to the large concentrations of column ice typical of strong convective storms, and insensitive to the assumptions made about the surface emissivity. Unlike IR satellite observations that measure only the temperature at the top of the clouds, the 89GHz microwave imager channel accounts for integrated hydrometeor quantities.

Comparisons were made between the observed and simulated 89GHz brightness temperatures as a function of observed and forecast rain rate. Figure 3.2 shows the mean rain rate forecast by the HRRR and retrieved by the Ku radar algorithm for a range of simulated and observed 89 GHz brightness temperatures. The sensitivity of the 89 GHz channel to frozen hydrometeors can be seen, as the strong relationship between observed rain rate and Tbs becomes apparent for Tbs less than 270K. In contrast, there is no apparent relationship between HRRR forecast rain rate and the Tbs simulated using the HRRR model output. Instead, simulated Tbs are slightly warmer than expected for light rain on the order of 1 mm/h. For light to moderate rainfall, however, simulated Tbs are significantly colder than observed. For example, rain rates of 5 mm/h are typically associated with observed Tbs in the 220-230K range, but simulated Tbs are approximately 60K colder.

In an to attempt to determine possible reasons for the very low Tbs at relatively low rain rates, seven observed/forecast feature pairs (four from 2014, three from 2015) were selected for

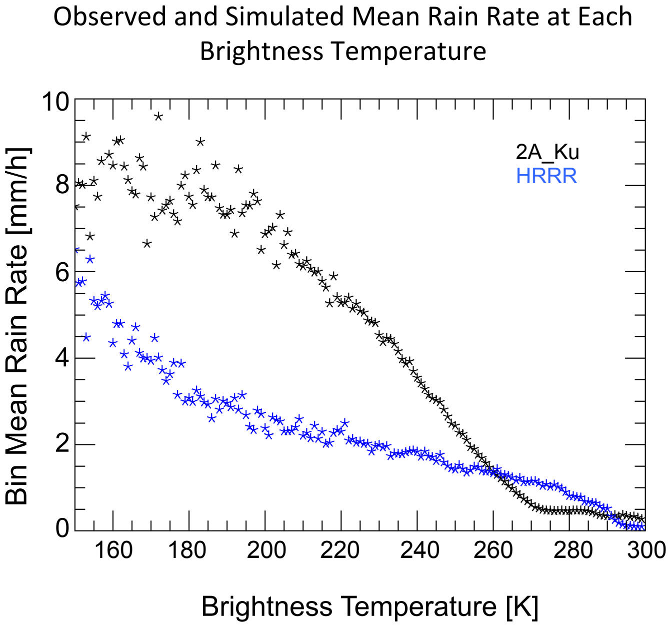

Figure 3.2. 89 GHz brightness temperatures and the corresponding mean rain rate from the observations (black) and forecast (blue).

further study. The selected features were chosen based on two criteria: A probability density function (PDF) of simulated Tbs indicating a significant number of grid boxes with Tbs substantially colder than observed, and an observed feature located entirely or at a large percentage within the swath of the Ku band radar, to enable the use of the 2BCMB product and Ku observed reflectivities

Figures 3.3 and 3.4 display a typical example of a selected case for a storm occurring along the Utah/Nevada border on July 26, 2014 at 06 UTC. In figure 3.3, only the features being considered are shown, with any surrounding precipitation masked out. This figure highlights the low bias in areal coverage by the HRRR when compared to GPROF, as the GPROF rainfall product indicates a contiguous feature appearing to consist of two convective lines, while the HRRR forecast results in two separate features, with only one of them allowed to be considered a match (At this time the matching algorithm can only select one forecast feature as a match for

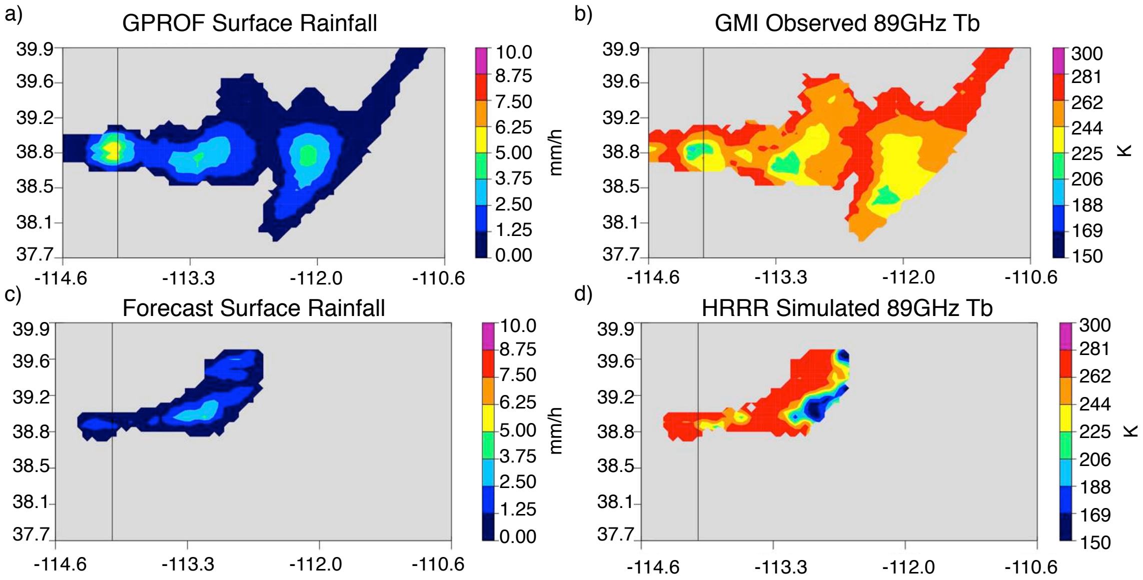

Figure 3.3. Maps of observed (top) and forecast/simulated (bottom) surface rainfall (left) and 89 GHz brightness temperatures (right) for a case occurring along the Nevada-Utah border on July 26, 2014 at 06UTC.

each observed feature, a trait common to features-based validation algorithms and a potential weakness of this methodology). The Ku radar covers all but approximately the western third of this storm, and agrees well with GPROF, with Ku indicating some areas of higher intensity and slightly different structure, likely due to the higher spatial resolution of the radar product (not shown).

In the region where both the HRRR and GPROF products indicate rainfall, we see that the HRRR feature is well placed and has a similar distribution of rainfall, but is smaller and less intense that the observed feature (Fig 3.3 a and c), which is unsurprising given the results shown in figure 3.1.

When attention is turned to the observed and simulated brightness temperatures (Fig 3.3 b and d), the GMI typically displays cooler Tbs in regions of heavier rainfall, as expected. However, the Tbs simulated using the HRRR atmospheric and hydrometeor profiles are significantly colder than observed, with minimum temperatures more than 50K less than the minimum Tbs from the observations. They are also offset towards the leading edge of the storm, which results in the colder Tbs being located in regions of light rain, as was seen in Figure 3.2. Additionally, outside of the cold cores, Tbs tend to be somewhat warmer than observed, also consistent with fig. 3.2.

Figure 3.4 displays PDFs of simulated Tbs and forecast surface rainfall, column rain water, and column cloud water from the HRRR, observed Tbs from GMI, retrieved rainfall from GPROF, and simulated Tbs and retrieved surface rainfall and column rain and cloud water from the GPM 2BCMB product for the case shown in figure 3.3. The tendency for the simulated HRRR Tbs to be somewhat warmer in warm regions and significantly colder in cold regions is evident (fig 3.4a), despite having a fairly similar distribution of surface rainfall to the retrieved

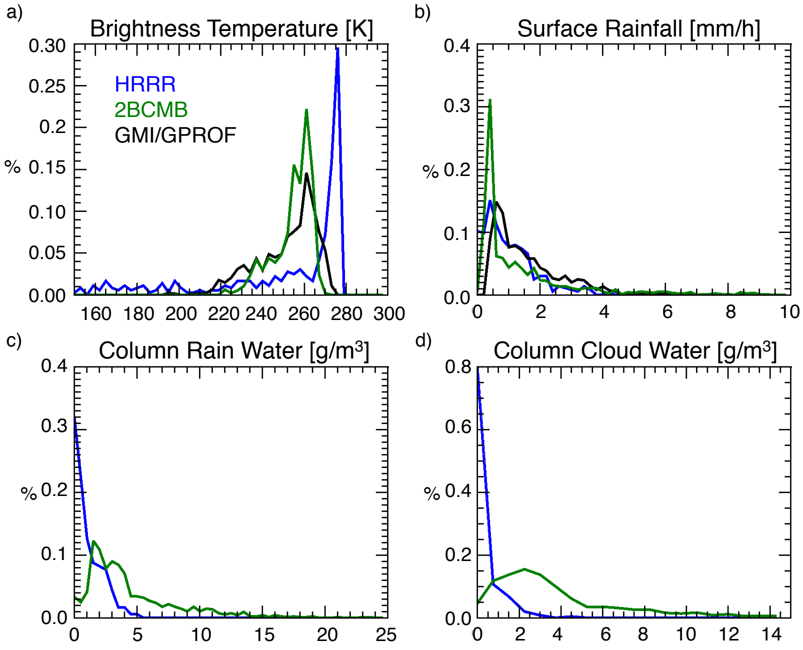

Figure 3.4. Probability Distribution Functions of a) Brightness Temperature, b) Surface Rainfall, c) Column Rain Water and d) Column Cloud Water from the HRRR forecast output (blue), 2BCMB product (green) and GMI/GPROF observations (black) for the case shown in Figure 3.3.

products (fig 3.4b). It is also evident that the model creates a large portion of the storm with lower integrated rain and cloud water than is retrieved by the 2BCMB product. It is possible that such low integrated water values are contributing to the slight warm bias at the higher end of the Tb scale by allowing more influence from the warm surface to be seen by the radiometer.

As mentioned previously, cold Tbs at 89GHz are typically associated with large amounts of ice. While the GPM products did not provide any retrieved ice information at the time of this writing, figure 3.5 shows that the HRRR forecast does indeed predict large amounts of snow, and to a lesser extent graupel, in the regions with the lowest simulated Tbs.

The results seen in figures 3.3-3.5 were typical of the seven selected cases, with the HRRR features showing regions with high concentrations of integrated snow associated with simulated Tbs significantly lower than observed values and slightly offset from maximum rainfall, as well as column rain and cloud water distributions with a high percentage of low

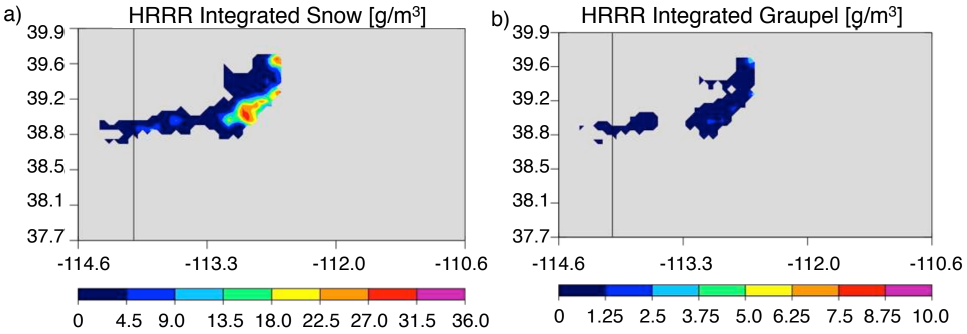

Figure 3.5. Forecast a) integrated snow water and b) integrated graupel water for the case shown in Figure 3.3.

values compared to the 2BCMB product. There are several possible explanations for why this is happening, including the microphysical scheme simply creating too much ice (a documented problem with previous versions of the Thompson microphysics scheme (Aligo, 2011; Lin et al., 2005), improper distribution of water content among liquid and frozen hydrometeors, or improper parameterization of drop size distribution parameters.

In order to try to gain a better understanding of the inconsistencies between observed and simulated 89 GHz radiances, a series of experiments were performed, altering one microphysical parameter at a time in the Tb simulations in order to try to obtain simulated Tbs and that more closely match the observed. These experiments are listed in Table 3.1, with those experiments resulting in improved representation of the Tbs indicated in bold.

Those experiments that were found to be successful were repeated, both by increasing the magnitude of the individual change that caused the improvement and by combining them with other successful experiments to see if further improvement could be achieved. The greatest improvements in the representation of simulated brightness temperatures were found when decreasing the density of snow hydrometeors by 75%, thereby creating unphysically “fluffy”

Table 3.1. List of changes made in the brightness temperature simulations to achieve simulated Tbs more similar to the observed. Experiments that were successful are highlighted in italics, and were repeated at higher orders of magnitude as well as combined with the other successful experiments.

Experiment Type	Microphysical Property Change	Re-distribution of Water
	Increase/*decrease* snow density	Transfer all snow to rain at T > 270K
	Increase/decrease ice density	Transfer all snow to graupel
	Increase/*decrease* graupel density	*Transfer of snow to cloud:* *100% at T>270K* *50% 260K < T < 270K*
	Increase/decrease rain intercept parameter (No)	At T>270K, evenly distribute snow water among snow, rain, and cloud. At 260K < T < 270K transfer 10% snow water to cloud and rain.
	Increase/decrease snow intercept parameter (No)	
	Increase/decrease graupel intercept parameter (No)	

snow flakes with density 0.025g/cm^3, and when transferring snow to cloud at varying proportions depending on temperature as shown in table 3.2, essentially increasing the concentration of supercooled liquid in the cloud. Figure 3.6 shows the improved PDFs of Tbs and column cloud water achieved when combining these two effects, i.e., transferring some or all of the snow to cloud water depending on temperature, and decreasing the density of the remaining snow particles by a more reasonable 40% (0.060 g/cm^3). While the slight bias in Tbs still exists at warmer temperatures, the skewness of the distribution of simulated Tbs into very cold temperatures has been eliminated. Meanwhile, the distribution of integrated cloud water is also much improved, with the distribution following that of the 2BCMB product very closely at values greater than $6g/m^3$, and a 30% decrease in pixels with very low integrated cloud water amounts.

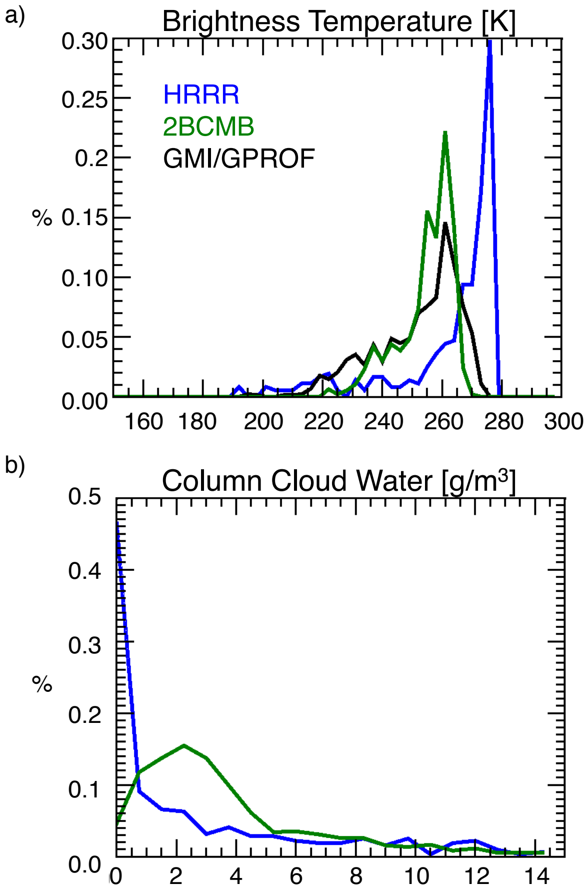

Figure 3.6: Same as Figure 3.4 b) and d), but for simulations in which snow water is transferred to cloud water in proportions given in table 3.2 and the density of the remaining snow is decreased by 40%.

Table 3.2. Amount of snow water transferred to cloud water for as a function of temperature for the results shown in figure 3.6.

Temperature Range	Amount of Snow to Cloud
T > 265K	100%
260K < T < 265K	75%
255K < T < 260K	40%
T < 255K	0%

3.4.2 Supercooled water

The results discussed above, as well as cross sections of the HRRR forecast hydrometeors through the centroids of storms (not shown) tend to indicate an absence of liquid hydrometeors above the freezing level. There are several possible explanations for the apparent improper partitioning of water content between the liquid and frozen hydrometeor species. At the root of some possible explanations lies the fact that the Thompson microphysics scheme "cuts off" the cloud ice category at particles reaching 200 μm. This serves to eliminate the need to consider growth of cloud ice by riming (thus considering only growth by deposition), but also creates many extremely small snow particles. The apparent excess snow and resultant cold Tb biases found outside the precipitation cores could be a result of these tiny particles being lofted out of the main cores in the direction of storm propagation. Another possibility is described in Thompson et al. (2008), wherein snow forms with the given particle size distribution in the upper levels of the cloud, growing by vapor deposition and falling through the cloud as they grow larger. When they descend to the levels where supercooled liquid water is typically present, they are large enough to grow efficiently by riming, thus depleting much of the supercooled liquid. It may also be posited that, if these were actively developing storms, the snow would likely be more widely distributed through the storm in later hours of the forecast, and, since snow reaches the surface more slowly than liquid particles, the cold Tbs accompanying low rain rates would make sense in this scenario (Greg Thompson, personal communication, 2016).

An additional possible explanation for the apparent lack of supercooled water above the freezing level takes a more ground-up approach. Luo et al. (2014) showed that storms with stronger vertical velocities lifted liquid hydrometeors and precipitation-sized particles to higher altitudes, resulting in increased growth of the droplets by collision processes and thus heavier rainfall. If forecast updrafts were weak, hydrometeors would freeze within the freezing level, reducing the amount of liquid water in the layers directly above it.

While updraft speeds are difficult to observe and therefore not typically measured, Szoke et al. (1986) suggest two ways in which profiles of radar reflectivity can be used to infer updraft strength. First, cells with stronger updrafts would be expected to reach their maximum reflectivity at higher altitudes due to the lofting of raindrops. Second, since cloud ice and snow flakes generally result in lower reflectivity than liquid hydrometeors, storms with a large quantity of lofted liquid drops would be expected to have a slower decrease in reflectivity with height above the freezing level than storms whose hydrometeors freeze at or just above the freezing level. This was demonstrated by Zipser and Lutz (1994), who emphasized the vertical gradient of reflectivity in the layer from 0^o to -20^oC (a layer approximately 3km thick). Their Figure 5 showed higher reflectivity at the freezing level and a much smaller decrease in reflectivity with height over the 3km above the freezing level for mid-latitude convective storms having stronger updrafts than their tropical oceanic counterparts. In general, they showed mid-latitude convective storms to have a change in reflectivity between -5 and +10 dBZ in the 3km above the freezing level, while tropical oceanic convection had reflectivity decreases of 10-25 dBZ over the same layer.

Figure 3.7 shows the PDF of the height of maximum reflectivity for the QuickBeam simulations using the HRRR hydrometeor output and the observations from the Ku radar for the seven selected cases. While the majority of observed storms have maximum reflectivity heights in the 3-4 km range, most of the simulated storms have maximum reflectivity values much closer to the surface. Figure 3.8 replicates figure 5 of Zipser and Lutz (1994) with the current data, showing the reflectivity at the freezing level in convective cores (defined as those grid boxes

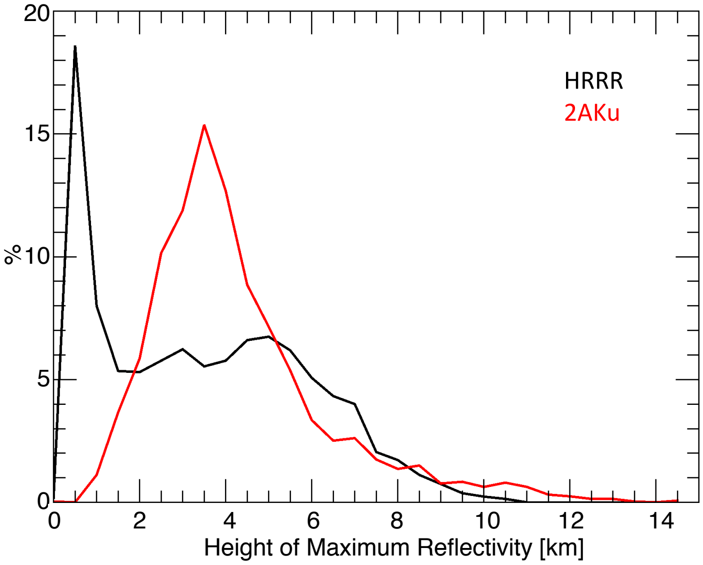

Figure 3.7. PDF of the height of maximum reflectivity for the Ku observations (red) and the HRRR simulations (black).

having freezing level reflectivity greater than 35 dBZ), and the change in reflectivity in the 3km above. While there is significant noise when freezing level reflectivity is between 35-40 dBZ, as the reflectivity at the freezing level increases, there is a distinct tendency for the vertical gradient in the observed features to change very little, remaining within the -5-+10 dBZ values reported by Zipser and Lutz (1994). The simulated reflectivities from the forecast features do not follow this trend, however. From these experiments, it can be inferred that there is some merit in the updraft hypothesis: that the HRRR model updrafts are not strong enough to loft liquid hydrometeors above the freezing level, resulting in excess ice hydrometeors in the column, colder-than-observed Tbs, and under-prediction of rainfall. Updraft strength may be a difficult problem to solve, given that it is highly dependent on the resolution of the model, as shown by Weisman et al. (1997) and Bryan et al. (2003).

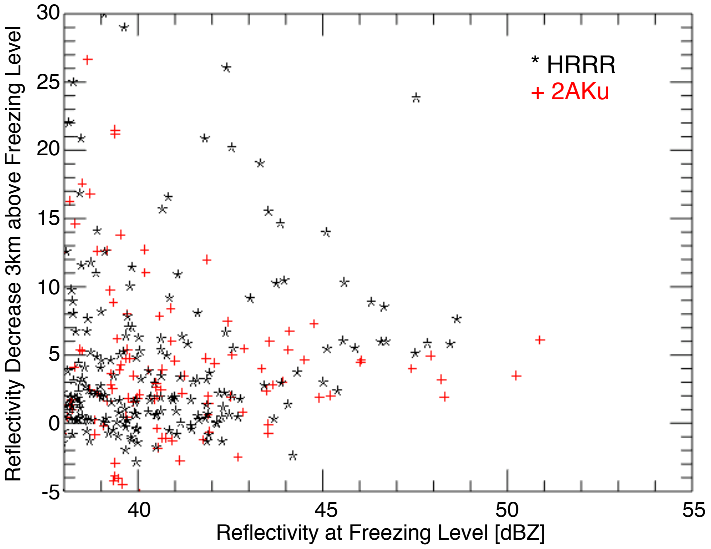

Figure 3.8. Change in reflectivity in the 3km above the freezing level as a function of reflectivity at the freezing level for the 2AKu observations (red plusses) and HRRR simulations (Black asterisk).

3.5 Conclusions

A features-based approach was used to evaluate HRRR 1 hour forecasts of precipitation against observations and products from the GPM core satellite over the western US. When compared to GPROF precipitation, the HRRR was found to under-predict precipitation amount, intensity, and areal extent. Radiances and reflectivities at frequencies matching those of the GPM core satellite instruments were simulated while maintaining microphysical properties (e.g. DSD parameters) as close to those used in the model microphysics package as possible. These simulated quantities were then used to evaluate the consistency between forecast and observed/retrieved cloud properties.

It was shown that the Tbs simulated using the forecast hydrometeor profiles had a significant low bias, often in regions with low hourly rainfall, whereas very cold Tbs in the observations were indicative of heavy convective rainfall. Seven cases with very cold simulated

Tbs in regions with light rainfall and corresponding observed features within the Ku radar swath were selected for further scrutiny, and the HRRR Tbs and reflectivities for these cases were re-simulated under varying conditions in an attempt to gain representation more consistent with the observations. Over a course of 20+ experiments, it was found that the closest match to observed Tb distributions was achieved when reducing the density of particles in the snow category while simultaneously transferring a fraction of the snow to cloud water in temperature-dependent proportions. This redistribution of water between hydrometeor classes also had the benefit of bringing the integrated cloud water content of the features more in line with what was retrieved in the 2BCMB product.

The results from the simulation experiments imply a lack of supercooled liquid hydrometeors above the freezing level. Several possible causes for this phenomenon were discussed. Conflicting behavior between observed and simulated reflectivity profiles provides some support for the updraft hypothesis, wherein updrafts in the HRRR model were theorized to be too weak, such that lofted hydrometeors traversed the freezing level slowly and were therefore completely frozen shortly after crossing it. A stronger updraft would result in the lofting of liquid hydrometeors above the freezing level, allowing for increased growth by collection and heavier rainfall. Since updrafts are difficult to measure, the height of maximum reflectivity and the vertical gradient of reflectivity in the 3km above the freezing level were used as proxies to infer updraft strength.

The results discussed herein seem to paint a more dismal picture of HRRR QPF quality over the western US than may be true. The results of figures 3.1 and 3.3 show that the model is producing precipitation in the correct places and often with similar shape and structure to observed storms, with some underestimation of the intensity and areal extent of precipitation.

Ultimately, the inconsistencies between the observed storms and those produced by the model appear to be related to an underestimation in the amount of supercooled liquid water near the freezing level. The results of the experiments outlined in Table 3.1 indicate that simply transferring some of the column water mass from frozen hydrometeors to liquid can bring the characteristics of forecast storms and their simulated radiative properties more in line with those observed by the GPM satellite. A systematic examination of the circumstances that produce the apparent underestimate of supercooled cloud water, and whether forcing the model to either create more supercooled liquid or transfer water content from other hydrometeor species results in an improved forecast of warm season convective precipitation will be left to future study.

It should be noted that the over-production of snow and ice hydrometeors is not a problem unique to the Thompson microphysics scheme. In WRF simulations run by Han et al. (2013), both the Morrison et al. (2005 and 2009) and Goddard (Tao and Simpson, 1993; Tao et al., 2003) microphysical schemes produced as much and more snow than the Thompson scheme. Similar behavior was also seen by Gallus and Pfeifer (2008), Wu et al. (2013), and Morrison and Milbrandt (2011). These results suggest that the ice production issues and inconsistencies between observed and simulated satellite products discussed herein are not limited to the HRRR model, and would likely be found in cloud resolving models using other microphysics schemes as well.

The goal of a model assessment, as described in BK15 is to take validation a step further, providing potential explanations as to why a forecast was or was not successful. The assessment process can also be used to explore inconsistencies between observations and model forecasts to provide insight into model behavior. The results discussed herein demonstrate the ability to use observations and retrieved parameters from space-borne instruments to infer potential weaknesses in model processes, which could be applied to other regional or global models to assess model forecasts in regions with sparse surface measurements.

CHAPTER 4:

A FEATURES-BASED ASSESSMENT OF THE EVOLUTION OF WARM SEASON PRECIPITATION FORECASTS FROM THE HRRR MODEL OVER THREE YEARS OF DEVELOPMENT

4.1 Introduction

The High Resolution Rapid Refresh (HRRR) model is an hourly-updated storm resolving model that was designed to provide rapid update model guidance on convective storms in order to improve severe weather forecasting, air traffic management, aviation hazards forecasting, and dissemination of severe weather warnings (ESRL, 2015). The HRRR became the National Weather Service's (NWS) operational rapid update forecast model in September 2014, and has undergone continuous development since before its transition to operational status. Upgraded, experimental versions of the model are run concurrently with the operational version in order to test updated model performance and measure improvements resulting from updates.

Bytheway and Kummerow (2015) (hereafter BK15) performed a detailed features-based assessment of quantitative precipitation forecasts (QPFs) of long-lived warm season convective storms over the central US using the 2013 experimental version of the HRRR. Bias statistics and composite features indicated that the model produced smaller, more intense storms than observed, particularly early in the forecast. Additionally, although the center of mass of precipitating features was well-placed with respect to the observed precipitation, model features were often produced somewhat farther south.

Since BK15, there have been number of specific areas where the HRRR has been updated, including changes to the assimilation and microphysical schemes, to potentially reduce spin-up and better represent initial precipitation fields early in the forecast. Changes were also instituted to reduce premature erosion of the cap in southern portions of the domain, with the intent of reducing the southward bias in precipitation production and improving the representation of the onset of convection.

With several years of experimental HRRR output now available, this study replicates the methodology of BK15 over the same domain (85-105 W, 29-49N, shown in Figure 4.1) in order to determine whether changes to the model had the expected outcomes, looking specifically for improvements to the problems mentioned above. In particular, answers to the following questions are sought:

- Have biases in intensity and areal extent of precipitation been reduced?
- Has spin-up time been reduced?
- Has the slight southward bias been reduced or eliminated?
- Has the tendency for over-prediction of heavy rain been reduced?
- How well does the HRRR capture the onset and development of convection?

The remainder of this manuscript will be structured as follows: Descriptions of the HRRR model and the National Centers for Environmental Prediction (NCEP) Stage IV multisensor precipitation product used as a reference are given in section 2. A brief review of the features-based methodology will be given in section 3. Section 4 will provide assessment results that indicate changes in model performance through multiple years of development, and will attempt to relate these results to specific changes that were made in the model. Concluding remarks will be presented in section 5.

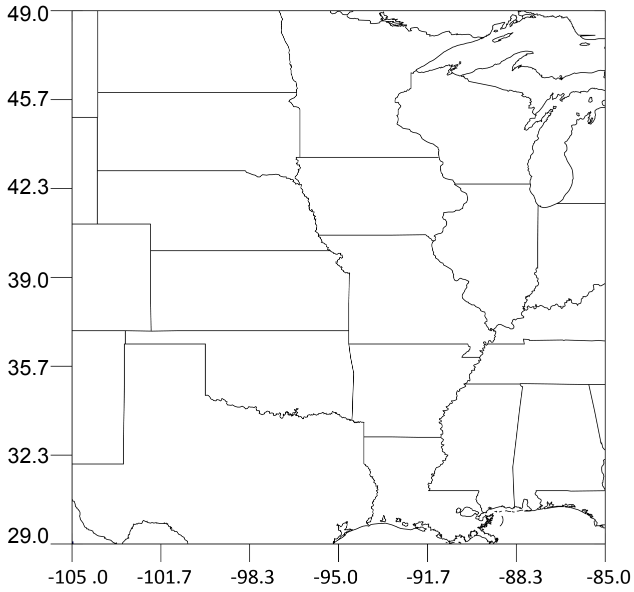

Figure 4.1. Map of the study domain.

4.2 Data

4.2.1 High Resolution Rapid Refresh Model

The HRRR model is an hourly-updated convection-allowing model with hourly forecasts produced over the contiguous US (CONUS) at 3km horizontal resolution with 50 vertical levels at native resolution, or 40 vertical pressure levels. The HRRR domain is nested within the 13km Rapid Refresh mesoscale model, which also provides boundary conditions (Benjamin et al., 2013). Full details of the HRRR model can be found in Benjamin et al. (2016). Forecasts from the 2013, 2014, and 2015 experimental versions of the model were evaluated, with the model output files obtained from National Oceanic and Atmospheric Administration Earth System Research Lab (NOAA ESRL) servers. While the 2016 experimental version of the model became available during the course of this research, the updates made were not expected to have a significant effect on precipitation forecasts (Stan Benjamin, personal communication).

While a maximum number of cases for study is desirable, local data storage for the large model output files was limited, and therefore only a portion of the available forecast runs were used in this study. Specifically, in 2013, forecasts initialized every other hour were used, with odd (even) numbered hours obtained on odd (even) numbered Julian days. For 2014 and 2015, data was collected at Global Precipitation Measurement (GPM) core satellite overpass times as well as the four hours leading up to the overpass. The GPM data is being used to evaluate the HRRR in a separate research study (Bytheway and Kummerow, 2017, submitted). This provided a large number of cases for study in each year (1108, 1207, and 744 matched feature pairs in 2013, 2014, and 2015, respectively. Lower numbers in 2015 are a result of several missing days of model output) and an adequate sampling of the diurnal cycle.

The HRRR has undergone continuous development and improvement, with new experimental versions of the model released yearly. Some of the changes expected to have an impact on QPFs are listed in table 4.1 and described below. Early versions of the model were found to have a warm, dry bias during the day in the warm season, which was traced to a positive bias in incoming solar radiation due to the inability of the model to represent sub-grid scale clouds (e.g. shallow, fair-weather cumulus). As a result, the boundary layer scheme was altered to include a sub-grid scale cumulus parameterization that had a significant impact on incoming sensible heat flux. Additionally, the Thompson microphysics scheme (Thompson et al., 2008) was updated to account for aerosols (Thompson and Eidhammer 2014). The addition of aerosol “awareness” resulted in both increased reflection of incoming shortwave radiation and increased

Table 4.1. HRRR upgrades expected to effect QPF quality.

HRRR Version	Assimilation	Microphysics	WRF Version	PBL
2013	GSI 3DVAR with Radiances	Thompson et al. (2008) with enhancements	3.2.1+	MYJ
2014	GSI with hybrid ensemble-variational (0.5/0.5)	Same as 2013, with minor adjustments	3.4.1+	MYNN
2015	GSI with hybrid ensemble-variational (0.75/0.25)	Thompson and Eidhammer (2014)	3.6.1+	MYNN 2015

cloudiness. These changes served to increase stability at the top of the mixed layer and slow boundary layer growth in the 2015 version of the model.

Additional changes to account for the warm dry bias in the warm season included changes to the land surface model both to account for irrigated crop land and reduce the wilting point of transpiration, essentially making it more difficult for the model to allow agricultural land to go dormant and thus increasing surface latent heat flux. These changes to the model physics in the boundary layer and at the surface resulted in a cooler, moister mixed layer, which in turn reduced the daytime warm bias by 2-3C and reduced high biases in convective initiation and production.

In addition to physics changes, many changes to the model data assimilation (DA) system were made from 2013-2015. The treatment of mesonet and METAR surface data was altered in the assimilation process to produce pseudo-observations: vertical profiles of temperature and dewpoint calculated based on the surface observations. These pseudo-observations were given increased weight in the assimilation optimization, essentially bringing the profiles at grid points with automated surface stations closer to the observed state.

The 2013 HRRR was the first version of the model to assimilate radar reflectivity via diabatic assimilation, a process by which the reflectivity measurement is used to calculate a latent heating perturbation to induce precipitation in the model initial state. The 2013 version of the model was fairly aggressive in adding latent heating in regions with observed reflectivity greater than 35 dBZ, and, as BK15 showed, this resulted in very high biases in extreme rainfall early in the forecast. In the 2014 version of the model, the threshold for perturbing latent heating was reduced to 28 dBZ, but the strength of the forcing was decreased by a factor of four, thus increasing the area of influence of the radar observations, but reducing the tendency to induce explosive convection.

Also in 2014, the DA system was upgraded from a 3D-Variational (3D-VAR) system to a hybrid system that included both 3D-VAR and GFS Ensemble Kalman Filter (EnKF) systems, each having equal influence on the covariance matrices. This resulted in increased assimilation skill with respect to standard atmospheric observations. The EnKF system weight was increased to 75% in the 2015 HRRR.

4.2.2 NCEP Stage IV Multisensor Rainfall Product

The NCEP Stage IV multi-sensor precipitation analysis (Lin and Mitchell, 2005, Nelson et al., 2016) is a 4km hourly precipitation product available over the CONUS. This product serves as the reference precipitation data set for this study, and consists of a mosaic of regional radar analyses produced by individual NWS River Forecast Centers (RFCs) and adjusted to gauge measurements. Each RFC produces an automated version shortly after the end of each hour of accumulation, which is followed several hours later by manual quality control and placement on a national grid at NCEP. Data is available in near-real time, however it may not contain information from all RFCs. Final mosaics become available 12-18 hours after the accumulation period and can be obtained from the National Center for Atmospheric Research (NCAR). These final quality controlled full mosaics are used in the current study. Smalley et al. (2014), Prat and Nelson (2015), and Nelson et al. (2016) provide some discussion of uncertainties in the Stage IV product.

In order to evaluate whether changes in the HRRR performance over three years are actually due to model changes or due to differences in the type of storms being forecast, the observed warm season accumulated precipitation for 2013, 2014, and 2015 is shown in figure 4.2. The maps in figure 4.2 indicate that the 2013 warm season precipitation was concentrated more over the southern half of the domain, with large amounts of rainfall along the Gulf Coast and a secondary maximum over the Southern Great Plains states of Oklahoma, Kansas, and Missouri. The accumulated rainfall distributions in 2014 and 2015 are more similar to each other, with rainfall maxima extending farther north into Iowa and Nebraska, with the Gulf Coast maximum still present, but in smaller magnitude than in 2013.

For a more statistical view of rainfall over the three-year period of study, figure 4.3 displays PDFs of the storm area, maximum rain rate, mean rain rate, and total rainfall for identified observed features over the three-year period, for both the northern and southern halves of the domain. Separating the domain into northern and southern sub-domains allows for an examination of whether the different pattern of accumulated rainfall in 2013 is a result of different types of precipitating features. The distributions in figure 4.3 generally have the same shape, and the year-to-year difference in the distributions is a result of the difference in the number of identified radar features year to year (14708, 16244, and 10659 in 2013, 2014, and 2015, respectively). The means of the distributions of the mean and maximum intensity are similar in both sub-domains in all three years, (mean maximum rain rate of 12.48 ± 0.47 mm/h

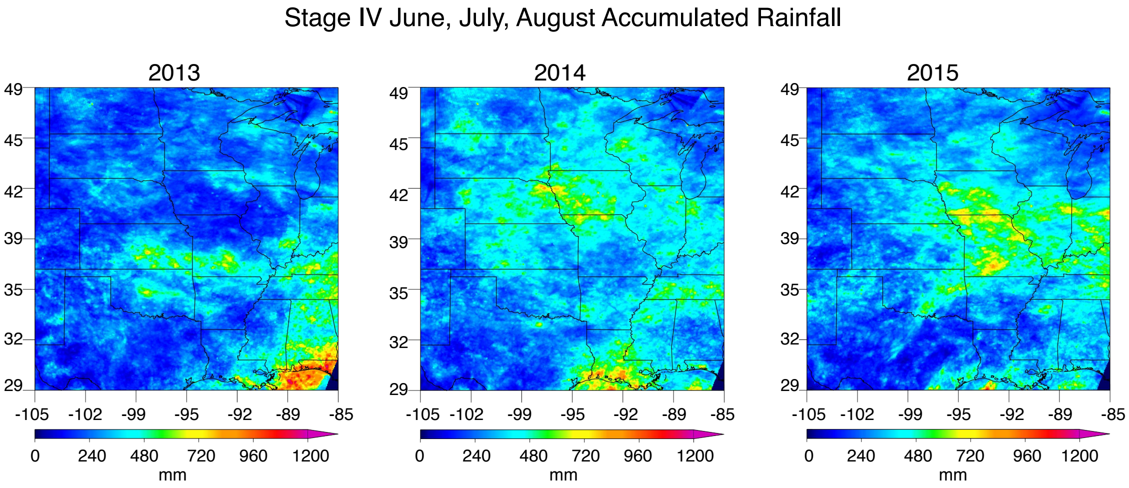

Figure 4.2. Seasonal accumulated rainfall from the Stage IV product for June, July, and August of 2013, 2014 and 2015.

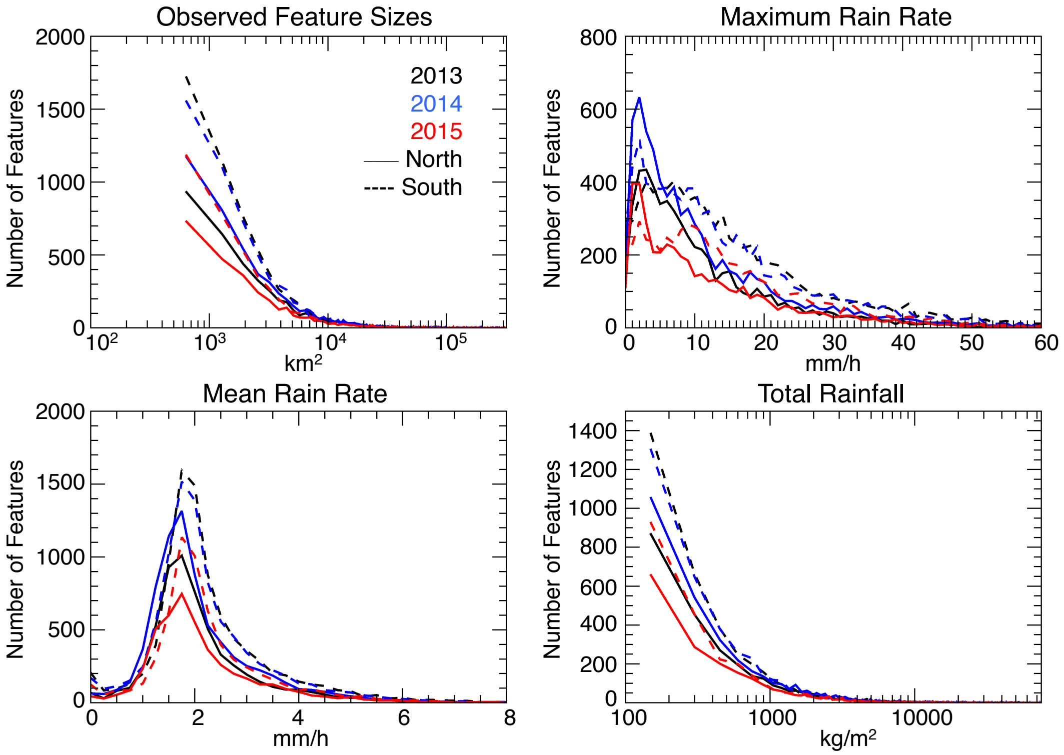

Figure 4.3. Distribution of observed feature sizes, maximum rain rate, mean rain rate, and total rainfall from the Stage IV product for the 2013 (black), 2014 (blue), and 2015 (red) warm seasons (JJA) in the northern (solid) and southern (dashed) halves of the study domain.

(15.88 ± 0.49 mm/h) in the northern (southern) domains, and average mean rain rate of 2.35 ± 0.1 mm/h in both domains). Therefore, any significant changes in forecast intensity biases can be assumed to be due to model changes and not to differences in the types of convective storms occurring. With respect to precipitating area, storms in the northern part of the domain were on average somewhat larger in 2014 and 2015 (5807km^2, 7354km^2, and 8184 km^2 in 2013, 2014, and 2015 respectively), but similar in the southern domain over the 3 years examined (~5800 km^2). Given that the precipitating area of the examined storms spans several orders of magnitude, the differences in the northern sub-domain can also be considered relatively small, and, at least in 2015, may be affected by the periods of missing model output. Thus, bias changes with respect to precipitating area are likely due to model changes, though impacts from natural variability cannot be completely ruled out.

4.3 Methodology

The assessment in this study follows the methodology of BK15, wherein features are identified in both the Stage IV and HRRR precipitating fields in a manner similar to the Method for Object-Based Diagnostic Evaluation (MODE; Davis et al., 2006 and 2009). Full details of the feature identification, tracking, and matching algorithms can be found in BK15.

The Stage IV and HRRR hourly precipitation fields are linearly averaged onto a 0.05-degree (~5km) grid, which places the fields on the same frame of reference while maintaining the majority of heterogeneity within the fields. A 15-km smoothing is then applied to the regridded fields, and a precipitating feature is defined as any contiguous area within the smoothed field exceeding a given rain rate (1mm/h) with a maximum rain rate exceeding 10 mm/h (to ensure likely convection). Features are identified through each of the first six hours of the forecast, and a database of statistical properties for each feature is created.

Using the database of statistical properties, both observed and forecast precipitating features are tracked through consecutive forecast hours. If an observed feature lasts for 6 hours or more, a matching forecast feature is sought at the beginning of the forecast. A match is defined as those forecast features that are either collocated with the observed feature, or nearby and statistically similar. Matching is done only at the first forecast hour, and calculated biases track how well the forecast features' behavior follows that of the observed features. If multiple forecast features match a given observed feature at the first forecast hour, only the most statistically similar forecast feature at that hour is used. If no matches are found, it is considered a missed forecast.

4.4 Results

4.4.1 Significant changes in model biases in 6-hour forecasts

As in BK15, distributions of the bias in the feature mean, maximum, and total rainfall, and areal extent were created for each forecast hour for all three years. Because BK15 indicated that these distributions were most useful early in the forecast, only the first 6 hours will be examined here. Significant changes between the three years, as determined by the Student's T-test, were mostly found to affect the feature mean and maximum rainfall. Significant improvements between the 2013 and 2014 versions of the HRRR were also found with respect to total rainfall during the first forecast hour (FH).

Figure 4.4 displays the bias distribution for maximum rainfall intensity at forecast hours 1 (top) and 6 (bottom) for the three years of interest, with solid vertical lines indicating the distribution median, dashed vertical lines indicating the interquartile range (IQR), and dotted vertical lines indicating the 10th and 90th percentiles. While the median bias in maximum rainfall intensity remains approximately 20% at FH1 in both 2013 and 2014, both the 75th and 90th percentiles decrease 20% in the 2014 HRRR, with the 75th (90th) percentile decreasing from 90% (160%) in 2013 to 70% (140%) in 2014. The 75th and 90th percentiles at FH1 do not change from 2014-2015, though the median bias in maximum intensity at FH1 decreases to just under 10% in 2015, a significant change over the 2013 version of the model.

At FH6, median biases in maximum rainfall intensity are again similar in 2013 and 2014, with a 20% decrease in the 75th percentile between the 2 years, and an overall wider distribution than that at FH1, consistent with BK15 results. From 2014-2015, the distribution continues shifting to the left, with 2015 median bias decreasing from ~+10% to ~-18%. The most obvious change at FH6 is the decrease in IQR, with the 75th percentile decreasing from +105% in 2014 to +57% in 2015.

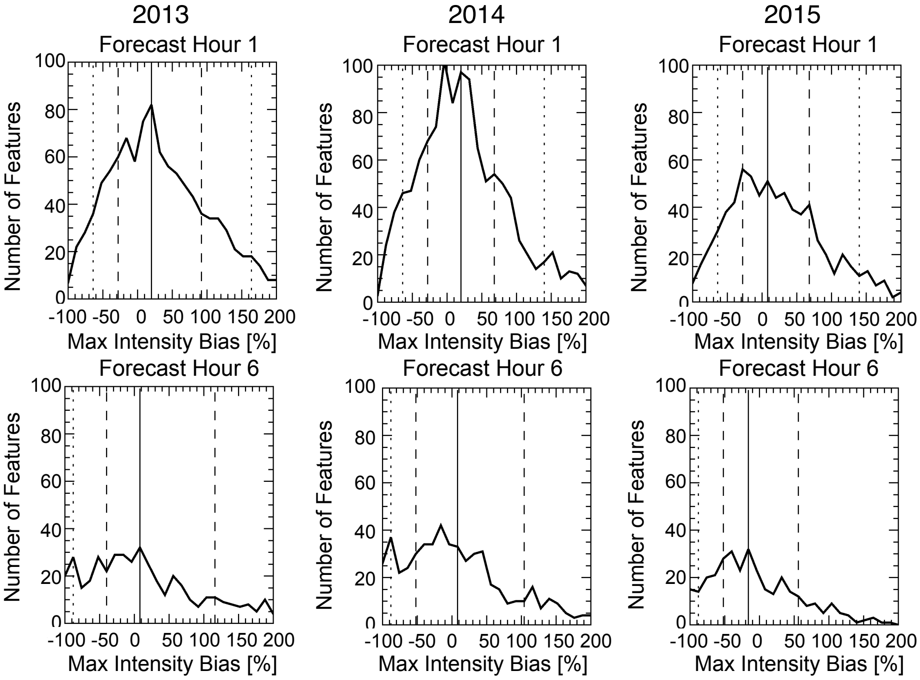

Figure 4.4. Distribution of maximum intensity bias in the central US domain for (left) 2013, (center) 2014, and (right) 2015 at (top) forecast hour 1 and (bottom) forecast hour 6. Solid vertical lines represent the median of the distribution, while dashed vertical lines represent the interquartile range and dotted vertical lines indicate the 10th and 90th percentiles.

Both DA and physics changes likely contribute to the significant improvement in the representation of precipitation intensity. The 2014 DA changes that lessened the imposed latent heat perturbation are expected to have the largest impact early in the forecast period, initially producing less intense storms than previous versions of the model. The improvement at FH6 is likely related more to the modifications made to decrease incoming shortwave and stabilize the boundary layer, since physics changes are expected to have impacts lasting longer into the forecast period than changes to DA techniques.

Mean bias distributions differ significantly at all forecast hours between the 2013 HRRR and the next two versions (excepting FH6 in 2015), while changes to the mean bias distribution between 2014 and 2015 are only significant at FH1, shown in figure 4.5. At FH1, the median bias in mean rainfall decreases from +20% in 2013 to +9% in 2014 and 2015. As in figure 4.4, the mean bias distributions show reduced skewness towards positive biases year to year. The 75th percentile decreases from +69% in 2013 to +42% in 2014 and 2015. While the 10th percentile remains stationary in all three years, the 90th percentile decreases from +115% in 2013, to +90% in 2014 and +80% in 2015.

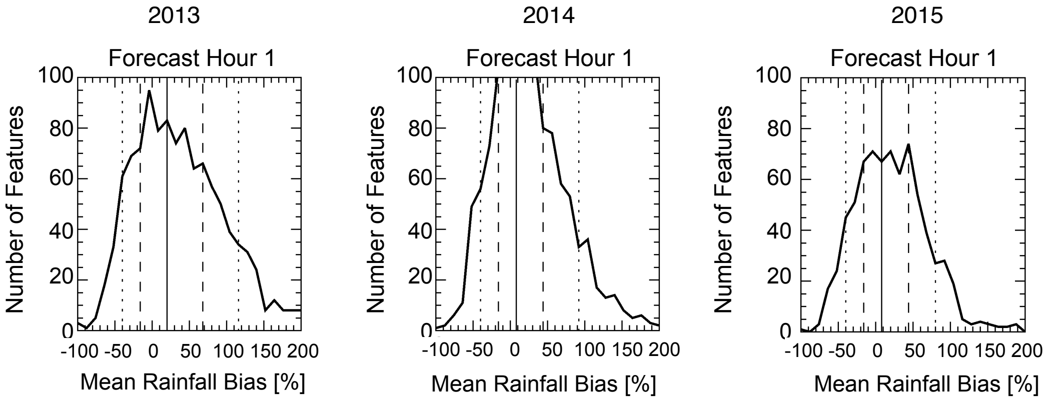

Figure 4.5. Same as figure 4.2 for mean rainfall bias in the central US domain at forecast hour 1 only.

Overall, changes in both the median intensity biases and the reduction in the skewness of the bias distributions towards large positive biases can likely be attributed to both the changes to the model physics that increased boundary layer stability as well as the alterations to the radar DA, with the DA upgrades having more impact early in the forecast, and the effects of physics changes carrying forward longer through the forecast period.

4.4.2 Model Spin-up time

BK15 showed that in the early hours of the forecast, the best bias statistics were obtained at FH3. Additionally, when comparing FH1 with the observations from either the hour of initialization or the hour when assimilation began (i.e. a 1-2 hour lag in observations to the forecast valid time), the lagged biases were smaller than those for the valid hour, particularly with respect to areal extent. By forecast hour 3, there was no longer a need to compare the forecast with lagged observations in order to obtain relatively low (generally within ±10%) median biases in feature mean, maximum and total rainfall and areal extent. These results indicated the need for several hours of spin-up prior to the HRRR producing the most accurate forecasts, and that in some cases, the assimilated observations were still heavily influencing the precipitation forecast in the first few hours of the model run. In particular, it seemed that the model produced storms with similar size to the assimilated storms with considerably more intense rainfall. Given that the HRRR is designed for short term forecasting of convection, it is desirable to have as short a spin-up time as possible.

Upgrades to the HRRR DA system between 2013 and 2015, particularly those reducing the strength of the latent heating perturbation induced in regions of observed reflectivity in the 2014 version, were inferred in the previous section to improve intensity biases in the early hours of the forecast. Figure 4.6 shows the median bias values for feature areal extent, mean rainfall, maximum intensity, and total rainfall for the first 6 hours of the forecast in 2013 (black x's), 2014 (blue triangles), and 2015 (red stars). Given that the BK15 results generally found median bias values within ±10% by forecast hour 3, here we will define the model as having spun-up at the hour when the median bias of a given property reaches ±10% (shown by vertical gray lines in figure 4.6), or when the absolute value of the median bias reaches its minimum during the 6 hours examined. The total spin-up time is then defined as the average number of hours needed for each of the four bias values to reach the ±10% (or absolute minimum) threshold, given in table 4.2. Using this definition of spin-up, the model on average spins up 45 minutes faster in 2015 than in 2013.

From the point of view of the model developers, the model is considered spun up when the size distribution of the model features matches that of the observed features, that is, an appropriate representation of the scales that the model is able to resolve. In order to examine model performance from this perspective, the distribution of feature sizes identified in both the model and observations is shown in figures 4.7, 4.8, and 4.9 for 2013, 2014, and 2015, respectively. In 2013 (fig 4.7), the model distribution indicates a nearly +10% bias in the

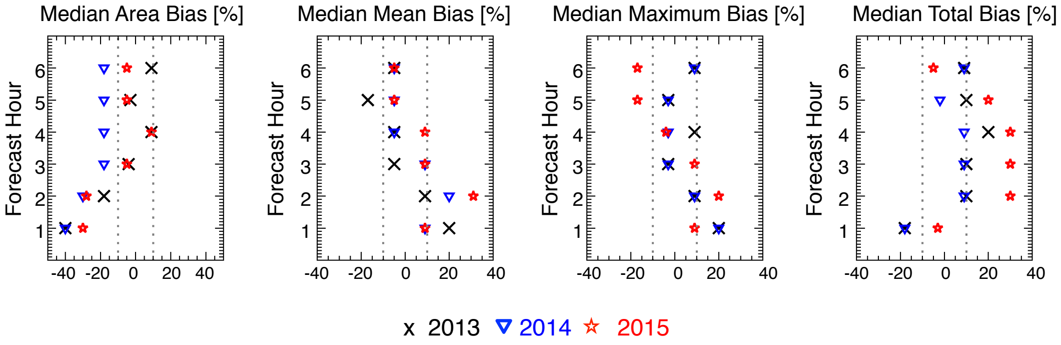

Figure 4.6. For each of the first six hours of forecast, median values of the distribution of area, mean, maximum, and total rainfall biases in (black x's) 2013, (blue triangles) 2014, and (red stars) 2015. Dotted grey lines outline the ±10% threshold.

Table 4.2. Number of hours needed for the model to reach ±10% median bias in mean, maximum, and total rainfall and raining area in each year, as well as the average total spin-up time.

	Area Bias	Mean Bias	Max Bias	Total Bias	Total Spin-up Time
2013	3	2	2	2	2.25
2014	3	1	2	2	2.0
2015	3	1	1	1	1.5

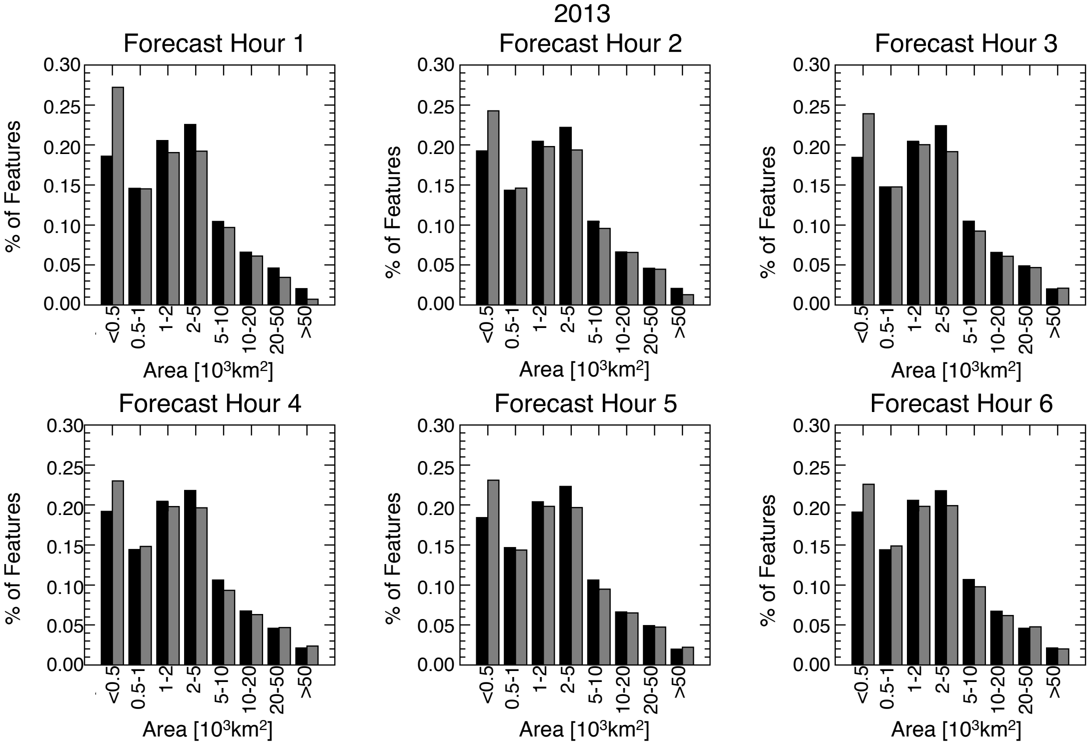

Figure 4.7. Distribution of feature sizes from the (black) observations and (gray) model at the first 6 forecast hours for the 2013 warm season.

smallest size category, with under-prediction of larger features, in particular those between 2000 and 5000 km^2. The large over-prediction of the smallest category remains over 5% until FH 4, when the under-prediction in the 2000-5000km^2 category is also smallest. By 2014 (fig 4.8) and 2015 (fig 4.9), the difference between the fraction of observed and forecast features in each size category is less than 2% in almost all forecast hours and size categories. Based on this criterion, the model does a better job representing the observed size distribution at early forecast hours in 2014 and 2015 than it does in 2013, indicating a spin-up time reduction from 4 hours to just 1. This definition of spin-up doesn't take into account the model's ability to represent the precipitation processes taking place within the raining area, however (i.e. just because the size is accurately represented does not mean the intensity and distribution of precipitation within the feature is accurately represented as well). As such, we can say that it appears that model spin-up time is decreasing with continued development, but by how much depends on how one defines spin-up.

The relatively small changes in overall model spin-up given by the median bias definition and the over-prediction (under-prediction) of features in the smallest (moderate) size category are not surprising given the larger, presumably more mature storms considered in this study.

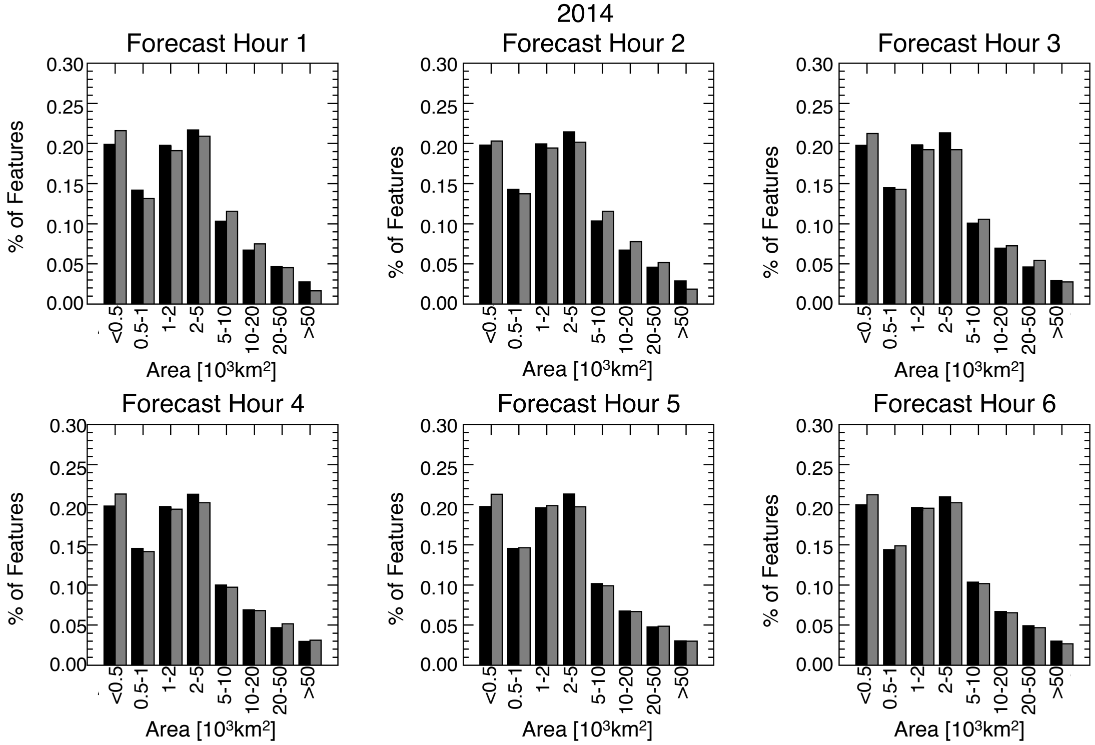

Figure 4.8. Same as figure 4.7 for the 2014 warm season.

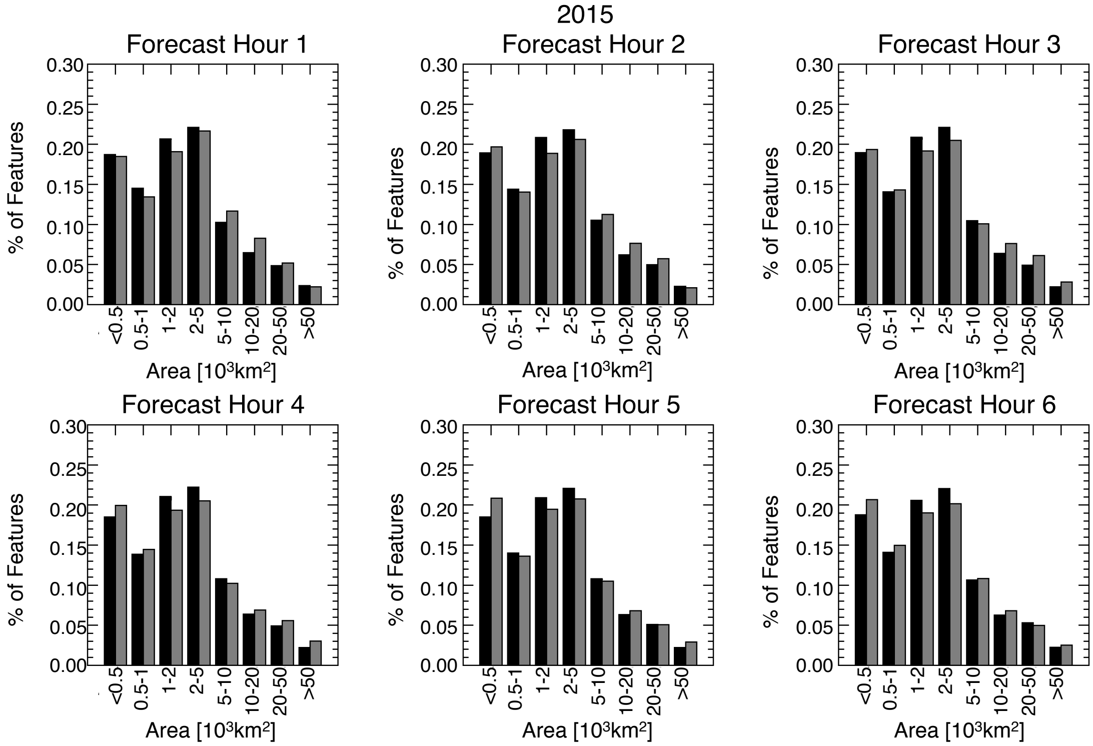

Figure 4.9. Same as figure 4.7 for the 2015 warm season.

Because the DA period is fairly short (1 hour) and the model output is produced at hourly time steps, the model is given just one hour to produce mature, possibly intense storms, which is unlikely given the model's current capabilities.

4.4.3 Southward biases

Figure 4 of BK15 showed the HRRR producing storms with centers of mass within about 25km of the centers of mass of observed storms, but with a slight southward bias that increased through the first few hours of the forecast (i.e. storms formed too far to the south, then either didn't correct position in ensuing forecast hours or propagated northward too slowly). This bias was attributed to both the generally larger instability of air masses closer to the warm, moist Gulf of Mexico and rapid erosion of the cap by the model. Such diurnally driven boundary layer processes are generally more prevalent to the south where synoptic forcing is not as strong as it

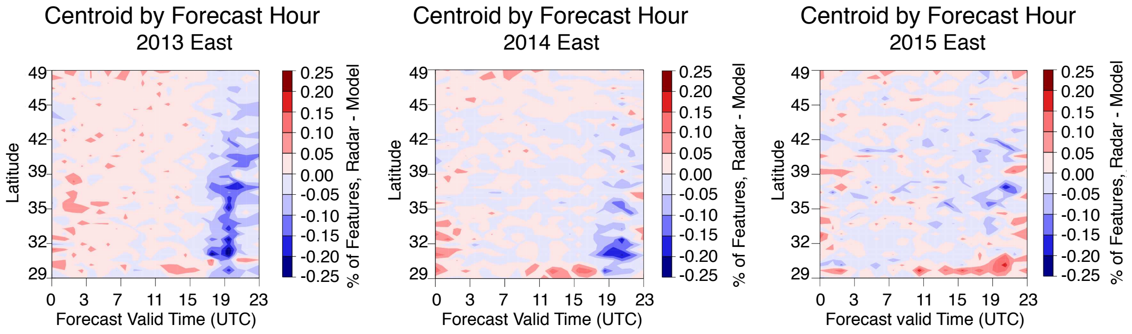

Figure 4.10. Difference in the fraction of observed and identified features in 0.5° latitude bands by time of day for the first hour of the forecast in the 2013, 2014 and 2015 warm seasons.

tends to be farther to the north, closer to the warm season storm tracks. Figure 4.10 shows the difference in the fraction of identified features between the radar and the model in 0.5° latitude bands by time of day for the first hour of the forecast (i.e., forecast valid time of 18 UTC corresponds to 17 UTC model run) for the eastern domain, calculated as

$$\left(\frac{\sum Radar\ features\ in\ 0.5^{o}\ band}{\sum All\ Radar\ features} - \frac{\sum Model\ features\ in\ 0.5^{o}\ band}{\sum All\ Model\ features}\right) \times 100.$$

Blue shading indicates a larger fraction of identified forecast features in a latitude band than observed, and appears starting around 17 UTC in the southernmost part of the domain in 2013. The over-prediction by the model propagates northward over the next several hours, however the strongest biases are found south of 39°N. This figure also indicates when the model begins to strongly initiate warm season convection, which will be discussed further in section 4.4.5.

In 2014, the over-prediction of convection in the southern part of the domain appears to begin about an hour later (~18 UTC), suggesting that updates to the 2014 version of the model delayed convective initiation slightly. The biases are limited mostly to areas south of 35°N, with strongest biases between 29°N and 32°N. By 2015, only a small region of over-prediction remains between 35°N and 39°N, and the tendency for over-prediction has been replaced by slight under-prediction in the southernmost part of the domain, indicating that the boundary layer physics changes and addition of the sub-grid scale cumulus parameterization made in the 2015 HRRR have in fact significantly reduced the tendency for the model to initiate strong convection in the south.

The tendency for the model to over-forecast convection in the southern US was not only dependent on time of day, but also on forecast hour. Figure 4.11 shows the same data as figure 4.10, but for the third hour of the forecast run. Overall, the biases are smaller in magnitude, and the excess initiation of convection was limited to regions south of 39°N in 2013, with some small biases being introduced in the overnight hours (00-05 UTC). The slight under-prediction seen in FH1 of the 2015 HRRR becomes evident in FH3 in the 2014 version. Combined with the FH1 results shown in figure 4.10, this suggests that the 2014 version of the model creates a large number of convective features early in the forecast that either a) die off rapidly or b) coalesce into a smaller number of larger features. By 2015, the under-prediction in the southern half of the domain seen in FH1 has started to increase towards the north with time, as the storms not produced at FH1either fail to materialize later in the forecast or those that do produce either do not propagate northward or dissipate.

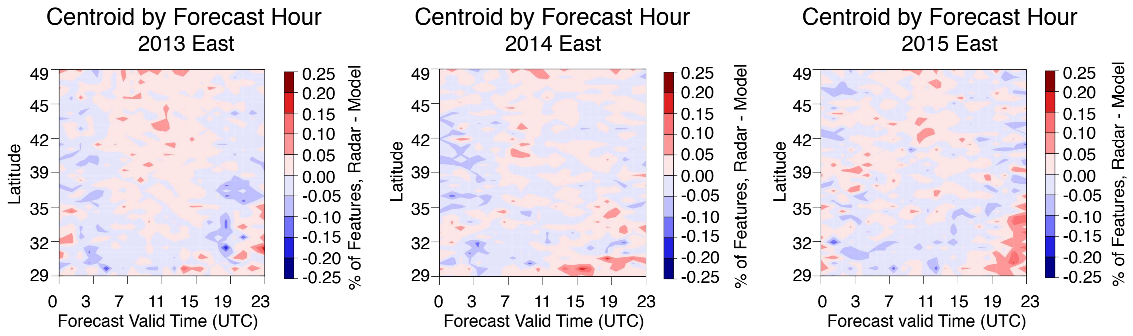

Figure 4.11. Same as figure 4.10 for forecast hour 3.

4.4.4 Overproduction of extreme rainfall

While not a direct result of the BK15 paper, high biases in maximum rainfall, excessively strong convective cores shown in composite features, and results from the developers' in-house validation work suggested that the HRRR was significantly over-predicting very heavy rainfall. The combination of reduced latent heat forcing in the DA along with changes to the model physics are one possible way to reduce this over-prediction. Figure 4.12 shows the difference in the fraction of identified features with maximum rain rates exceeding 5, 10, 25, 50, 75, and 100 mm/h (calculated in the same manner as equation 1) in the first 6 hours of the forecast for the three years of study.

The top panel of figure 4.12 shows that the 2013 HRRR usually under-forecast the fraction of features with rain rates exceeding 5 and 10 mm/h. This was also indicated in BK15 by a low probability of precipitation for features exceeding 5 mm/h in the absence of very large amounts of atmospheric moisture. For features with rain rates exceeding 25-75 mm/h, however, the 2013 HRRR had a tendency to over-predict at all of the first 6 hours of the forecast. This suggests that the 2013 HRRR needed some catalyst to produce moderate to heavy rainfall, but once that requirement was met, it was able to produce and sustain more intense rainfall than observed.

By 2014, the pattern at 5-10 mm/h reversed, with the HRRR over-predicting the fraction of features exceeding these thresholds by a significant margin, particularly at forecast hour 3. The tendency to over-predict the fraction of features with maximum rain rate exceeding 5 mm/h increases over the first 3 hours of the forecast, then decreases through hours 4-6. This is possibly related to assimilation of lighter rainfall induced when the reflectivity threshold for assimilation was reduced from 35 to 28 dBZ. In 2014, FH6 has only very small biases at all thresholds, a likely result of physics changes, as DA impacts are typically expected to be strongest early in the

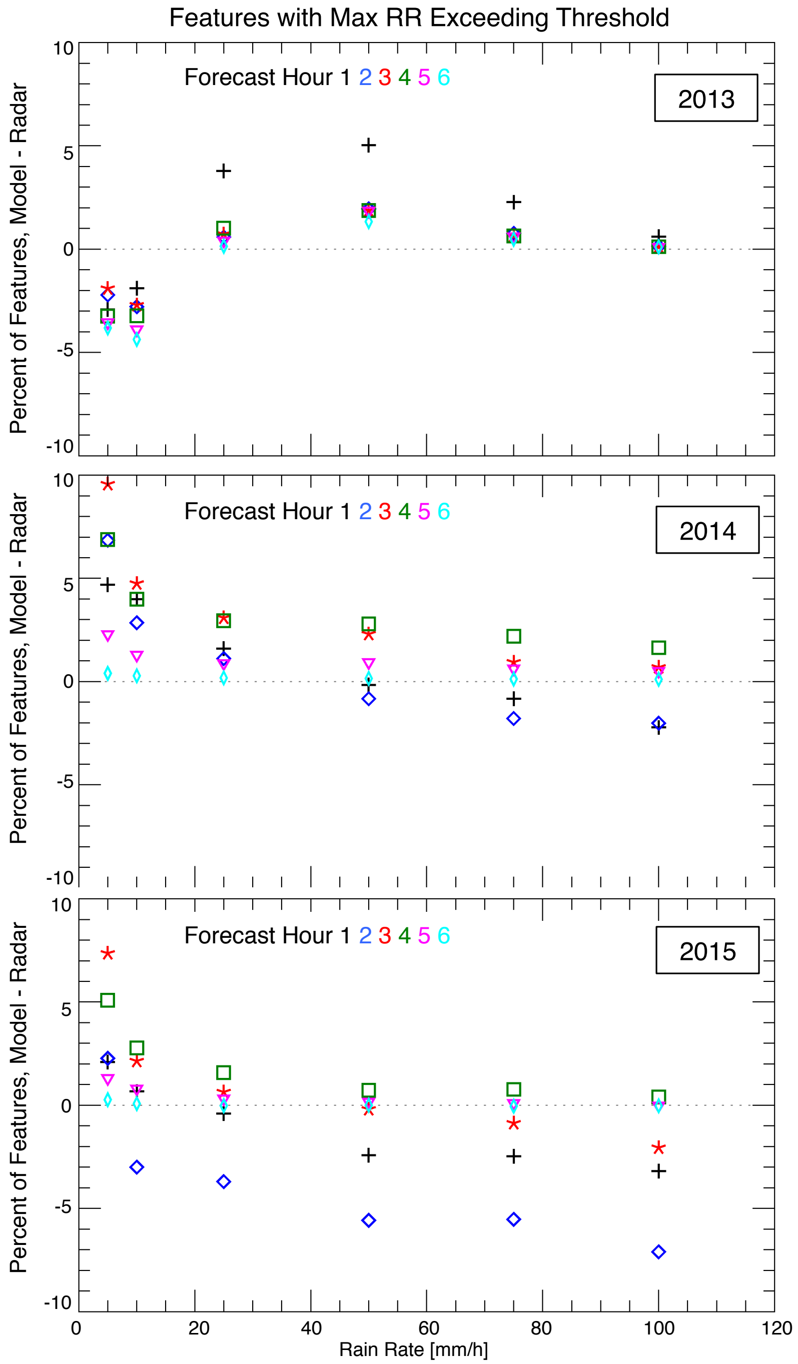

Figure 4.12. Difference in the fraction of identified features from the model and observations with maximum rain rate exceeding given thresholds in the first 6 forecast hours in the eastern domain for (top) 2013, (middle) 2014, and (bottom) 2015. Colored symbols indicate the hour of the forecast period.

forecast. DA changes were more likely responsible for the decreasing over-prediction of heavy rainfall with increasing thresholds in the first 3 forecast hours, and under-prediction of features exceeding 50 mm/h at FH1 and 2.

By 2015, only the fraction of features with maximum rainfall exceeding 5 mm/h are over-forecast at all of the first six forecast hours, while the tendency to under-predict the fraction of features with very heavy rainfall in the early forecast hours that became evident in 2014 is now apparent through the first three hours of the forecast, particularly for those features with maximum rainfall exceeding 75 mm/h. The largest under-prediction is seen at forecast hour 2 for all features with maximum rainfall exceeding 10 mm/h. Clearly the 2014 changes made to the DA combined with 2015's increased boundary layer stability have acted to dampen the production of convective rainfall early in the forecast, perhaps somewhat too much. It is hypothesized that, with the increased stability afforded by the physics and assimilation upgrades, the model is exhausting much of the energy needed to initiate and sustain these storms early in the forecast, resulting in under-forecast heavy rainfall in the next hour, with recovery taking place thereafter. Updates to the physics continue to produce improved forecasts at longer forecast hours, as biases for forecast hours 5-6 are nearly zero for features exceeding 25 mm/h, and over-prediction of features with heavy rainfall at FH4 is significantly reduced from 2014.

4.4.5 Onset of convection

While not studied in the original BK15 paper, an evaluation of how the HRRR handles the onset of convection is of interest to better understand model processes and how model upgrades have affected forecasts of convective precipitation. We already have some indication from figure 4.10 that the 2013 HRRR initiated convection around 17 UTC, and that initiation started closer to 18 UTC in the 2014 version of the model. Cai and Dumais (2015) found that the

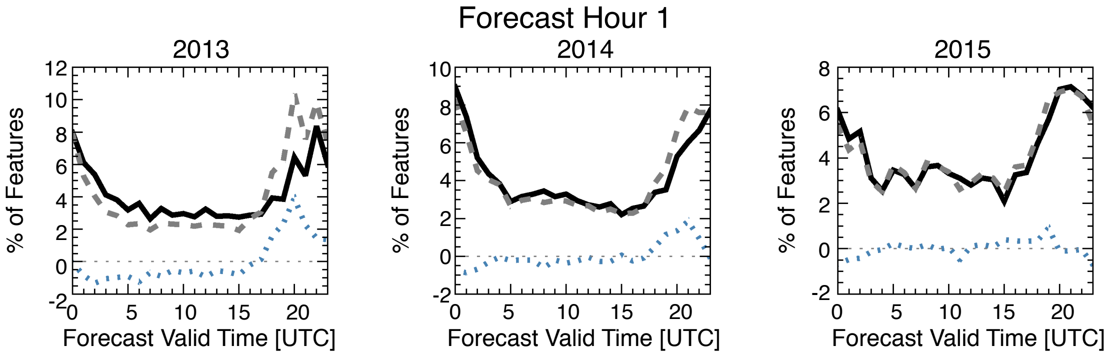

Figure 4.13. The fraction of storms produced by the HRRR at forecast hour 1 in each forecast valid time compared to the fraction of observed storms identified in each hour. Black solid lines represent the observations, grey dashed lines the model, and dotted blue shows model-radar.

2010 version of the HRRR tended to initiate convection somewhat later than observed with a four-hour lead time. Figure 4.13 shows the fraction of features identified at each valid forecast hour in the model (dashed gray), radar (solid black), and the difference between them (blue dotted), for all three years at FH1. Figure 4.14 displays the same for FH3.

In forecast hour 1, the 2013 HRRR starts increasing the fraction of predicted features at approximately 17 UTC, consistent with the results from figure 4.10, but, contrary to the results of Cai and Dumais (2015), approximately two hours before the increase in the fraction of observed features. The conflicting results are likely due to the Cai and Dumais (2015) study examining a different version of the model, different lead time, and different domain, as well as the previous study looking at absolute number of identified features rather than the fraction of total features identified at each forecast valid time. The 2013 HRRR also dissipates existing convection too quickly, or fails to produce convection in the overnight and morning hours. This is not too surprising: without the forcing and instability caused by daytime heating, long-lived convective features that were not already in existence at sunset would be difficult to produce.

By 2014, the model still initiates convection too early, although it appears to be slightly more in line with the radar. Overall, the fractional difference between the model and the radar is

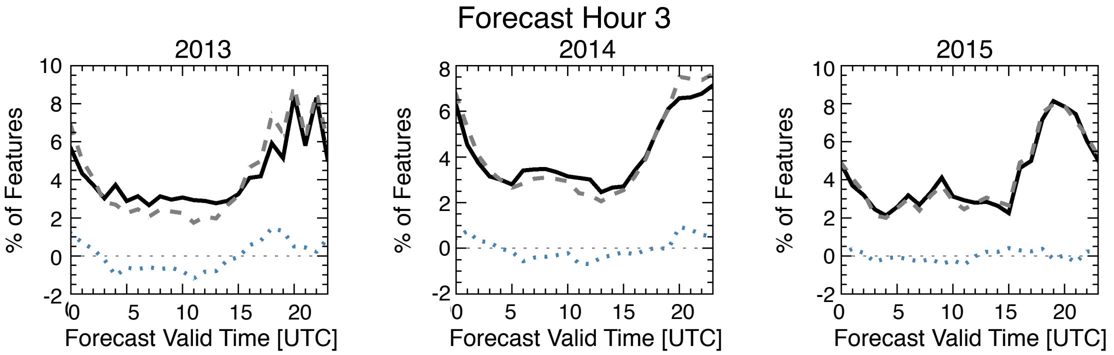

Figure 4.14. Same as figure 4.13 for forecast hour 3.

smaller than in 2013, and the overnight to early morning hours (04 UTC – 16 UTC) are very closely in line with the observations. While the tendency to dissipate convection too early or to fail to form evening convection remains, it is limited to 22 – 03 UTC. Given that these changes are found in FH1, they are likely a result of the DA changes. Whereas the 2013 version rapidly intensified assimilated storms within the first forecast hour, the 2014 version seems to have developed assimilated rainfall into convective storms more slowly.

The 2015 HRRR shows significant improvements in capturing the onset of afternoon convection. The tendency to dissipate or under-produce convection in the evening/overnight hours is also reduced. Since major changes to the DA were made in the 2014 version and physics changes were primarily made in the 2015 version, the continued improvement in convective onset is likely attributable to the improved representation of boundary layer stability and sub-grid cloud processes.

As with the southward bias, figure 4.14 indicates that the issues with convective onset are generally more evident in FH1, further indicating a continued need for model spin-up. By FH3, the 2013 HRRR still ramps up convective production a bit early, but falls more in line with the observations by 20 UTC. The lack of overnight convection in the absence of strong forcing is still evident in the same magnitude as at FH1. The 2014 HRRR slightly over-predicts the fraction of storms occurring between 19 and 03 UTC at FH3, while convective onset appears well represented at 3-hour lead time in the 2015 version of the model.

In addition to the desire to gain improved understanding of how the HRRR represented convective initiation, there was also some anecdotal evidence that the timing of storm production was dependent on the size of the features of interest. In particular, it was posited that smaller storms were generally produced too early, while larger, more mature storms were too slow to form. Figure 4.15 shows the fraction of features by forecast valid time at FH1 for features smaller (larger) than 750 km^2 (5000km^2). These size thresholds represent approximately the smallest and largest 25% of features.

Figure 4.15 shows that the smallest ~25% of forecast features do in fact form well before the smallest observed features in 2013 and 2014. In 2015, the timing is much improved. Despite the improved timing, all versions of the HRRR create a much larger fraction of its smallest identified features in the afternoon hours than are observed. With respect to the largest 25% of features, figure 4.15 shows that the 2013 HRRR produces these more mature features about an hour too early. By 2014 and 2015, the model times the initial onset of larger features more appropriately, but is not creating a large enough fraction of them. The 2014 HRRR also allowed a higher fraction of the largest features to remain overnight and into the early morning hours than was observed. This signal is still present in the 2015 results, but decreases in comparison to the 2014 version.

One hypothesis as to why the model initiates a large number of small features but lags behind in the larger features relates to the spin-up discussed previously. Many mesoscale convective features in the US start out as a number of smaller features that later coalesce into a

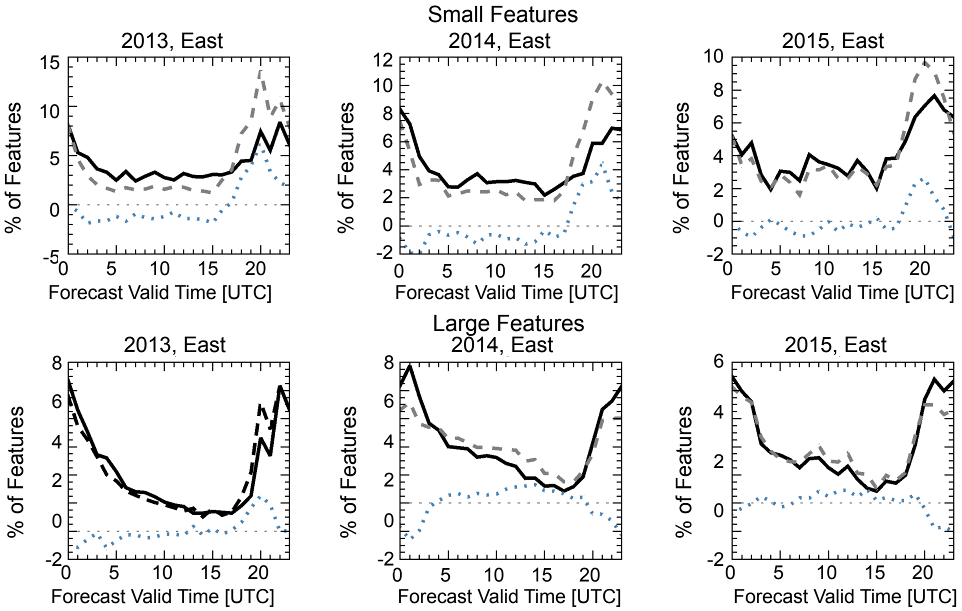

Figure 4.15. Same as figure 4.13 for the (top) smallest ~25% and (bottom) largest ~25% of features.

larger system. Current model capabilities are not able to represent this process in the one hour between initialization and the first forecast, though there has been improvement through updated DA and physics. Figure 4.16 shows the difference in the number of forecast and observed features identified at various size thresholds for forecasts valid at 18 UTC at a variety of lead times. As in figure 4.7, the 2013 HRRR significantly over-predicts the smallest features at all forecast lead times, but most appreciably at FH1. The over-estimation of the number of features decreases in each increasing size category, with overestimation of the number of features at 3 and 5-hour lead times surpassing that at FH1 in larger size bins. In the 2014 and 2015 versions of the model, the level of over-prediction of the smallest category of features decreases significantly in the first forecast hour. In 2014, all size bins show improvement, with some under-estimation of the number of features in the 500-1000km^2 bin at FH3. This is possibly an indication that the model is now growing small features into larger features at too rapid a pace.

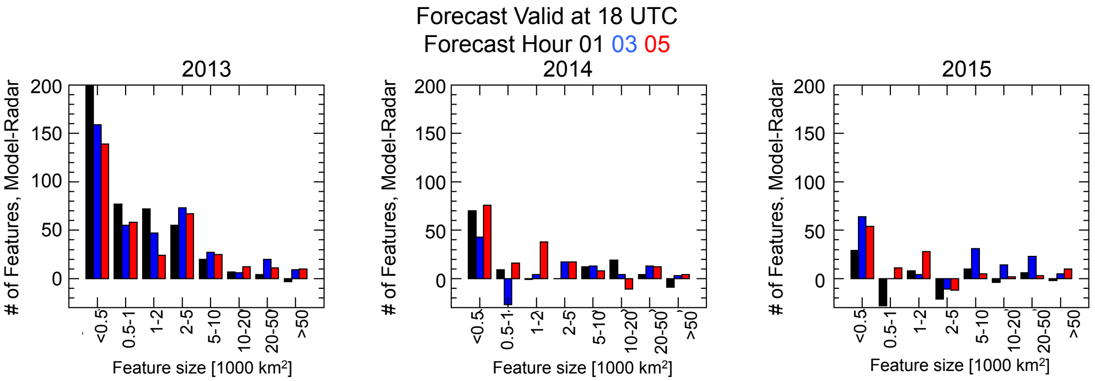

Figure 4.16. Difference in the number of features identified by the model and radar in various size bins at forecast lead times of 1, 3, and 5 hours for forecasts valid at 18 UTC.

In 2015, the tendency to over-forecast the number of small features at FH1 continues to decrease, with the model now under-estimating the number of features in the 500-1000km^2 and 2000-5000km^2 bins. In the 500-1000 km^2 bin, the one hour lead time now appears to have approximately the same level of under-prediction as the 2014 HRRR at FH3, which could potentially be related to the slight decrease in spin-up discussed in section 4.4.2.

4.5 Conclusions

The HRRR has undergone several years of development at ESRL, with the goal of consistently providing improved forecasts. Several changes made to the model physics, including to the representation of the surface and boundary layer, had the effect of reducing instability, particularly in the southern US where boundary layer forcing tends to produce more warm season convection than synoptically forced disturbances. Changes to the DA also improved initial state representation, while reducing the model's ability to produce volatile convection early in the forecast. Overall, changes to the HRRR model have resulted in environments that are less conducive to premature convective initiation, and less conducive to creating explosive convection, as indicated by improved initiation times and a reduced tendency to over-forecast very heavy rainfall in early forecast hours. As a result, bias distributions for mean, maximum, and total rainfall, as well as areal extent, have improved overall during the three years of development examined here, with the most significant improvements related to biases in the maximum and mean rainfall intensity.

The results discussed herein show that the southward bias discovered in BK15 was dependent on the time of the forecast, and not an issue with precipitation placement overall. The southward bias was found to correspond with the onset of convection, beginning at around 17UTC near the Gulf Coast, and migrating northward over the next several hours. This tendency for the HRRR to position convective rainfall south of its observed position was strongest in the 2013 version of the model and in the first hour of the forecast, with improvements such that the 2015 version of the model actually appears to have a bit of a northward bias at forecast hour 1.

The median biases of raining area and intensity indicate continued need for model spin-up, although the forecast representation of storm sizes indicates significantly reduced spin-up times in 2014 and 2015. The continuing need for spin-up is also evident in the tendency for the bias in forecast fraction of heavy rainfall to jump from positive to negative between forecast hours 1 and 2 in 2014 and 2015 (fig 12), the better representation of centroid placement and convective onset seen in the third forecast hour (figures 11 and 14), and the reduced difference between the number of predicted features in various size categories with increasing lead time (fig 16).

CHAPTER 5: CONCLUSIONS

The studies within this document seek to provide the first in depth examination of the quality of HRRR convective precipitation forecasts. A features-based methodology is employed in order to evaluate how well the model represents individual storms in two regions of the US. While grid-based statistical validation methodologies allow model performance to be evaluated over larger areas in a more general sense, the features-based methodology allows for a detailed examination into not only the forecast quality, but also model behavior within an individual storm, and thus provides the opportunity to link a successful or unsuccessful forecast to model processes. A features-based validation methodology also allows for the tracking of precipitation features through time, in turn allowing for an examination of how the forecast quality evolves through the period of the forecast.

The spatial heterogeneity and rapidly varying nature of precipitation, and in particular the warm season convective precipitation studied herein, requires a high spatial and temporal resolution reference dataset for use in validation of high resolution QPFs. For that reason, remote sensing datasets were used as references in the validation step of the assessment process. The Stage IV multisensor precipitation product provides a quality controlled estimate of precipitation at a spatial and temporal resolution that is similar to HRRR forecast output, however its reliance on radar data makes it somewhat less reliable in areas of complex terrain. Space-borne sensors, such as those aboard the GPM core satellite can fill in where surface radar data is less reliable, however due to the low satellite sampling frequency they currently lack the capabilities to assess feature performance through the forecast period.

The value of the assessment methodology introduced in chapter 2 is evident given the continued improvement seen in precipitation forecasts through three years of model development. This methodology not only brings attention to QPF biases, but also draws attention to the model processes and assumptions that may cause them. For example, the 2013 version of the HRRR represented the first US NWP model assimilating radar data at high spatial and temporal resolution. The addition of radar data to the assimilation process was expected to provide better initialized model states and therefore better forecasts of precipitation. Comparisons with larger-scale global models that lack radar data in the initialized state show that the HRRR represents an improvement in QPF forecast quality, but the more detailed narrative afforded by the features-based assessment showed that, particularly in this early version of the model, aggressive latent heating perturbations in the DA step combined with the threshold at which reflectivity was assimilated resulted in significant high biases in precipitation intensity and forecast storms that were smaller than the observed. As a result, the assimilation of radar reflectivity was altered in order to broaden its areal coverage, while the latent heat perturbation was weakened to avoid explosive convection, resulting in the improved representation of precipitation intensity discussed in chapter 4. While not studied here, model behavior could also be analyzed by season or as a function of forcing strength.

The assessment methodology also serves as a bridge between the modeling and observation communities. An example of this capability is given in chapter 3's examination of the inconsistencies between the satellite observations and the simulated satellite products created with the model output. The simulated satellite products indicated inconsistencies between observed storms and those produced by the model, and reproduced a phenomenon (excessive ice production) that has been widely reported among microphysical models. Chapter 3's results were found in only a fraction of the studied storms, and an in depth look at the environmental circumstances that produce these conditions is a worthwhile endeavor for future research.

The ability to link forecast biases with model processes is necessary in order to continue to develop NWP models and improve forecasts, as was demonstrated in chapter 4. As discussed, the concentration of heavy precipitation in smaller-than-observed storms early in the forecast found in chapter 2 was linked to the reflectivity DA, and significantly improved in subsequent versions of the model. The assessment process is also constructive in the opposite direction, that is, testing for an expected outcome from a model change. For example, chapter 2 described a southward bias in the placement of convection, but it was in-house work by the developers that linked it boundary layer processes. Adjustments to the boundary layer physics and the addition of sub-grid scale clouds were expected to remove this bias and also improve the representation of convective onset, both outcomes which were seen in the results of chapter 4.

Sharing the outcomes of model assessments with the model developers is an essential step to successfully put the assessment results to use. Validation of the HRRR has been performed both by the developers themselves and by several independent researchers, with results from different studies often supporting the results of others. In combination, the results of these assessments can help the model developers cultivate a full picture of what may need to be changed in the model. The results in chapter 4 demonstrate that a thorough understanding of the processes behind a bias pattern can lead to improved model forecasts. Continued improvement will come from continued understanding of the model processes causing remaining biases, and the assessment methodology will become more critical as the reasons for biases become harder to flush out after all the low-hanging fruit model upgrades have been made.

REFERENCES

Aligo, E. A., 2011: An evaluation of fall speed characteristics in bin and bulk microphysical schemes and use of bin fall speeds to improve forecasts of warm-season rainfall. Dissertation, Department of Geological and Atmospheric Sciences, Iowa State University, Ames, IA. 200pp.

AghaKouchak, A., N. Nasrollahi, J. Li, B. Imam, and S. Sorooshian, 2011: Geometrical characterization of precipitation patterns. *Journal of Hydrometeorology*. **12**, 274-285, doi: 10.1175/2010JHM1298.1.

Benjamin, S., C. Alexander, M. Hu, S. Weygandt, P. Hofmann, J. Brown, J. Olson, K. Brundage, and B. Jamison, 2013: Data assimilation and model updates in the 2013 Rapid Refresh (RAP) and High-Resolution Rapid Refresh (HRRR) analysis and forecast systems, *NCEP/EMC meeting*. [Available online at http://ruc.noaa.gov/pdf/NCEP_HRRR_RAPv2_6jun2013-Benj-noglob.pdf].

Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Monthly Weather Review*, **144**, 1669-1694, doi: 10.1175/MWR-D-15-0242.1.

Bryan, G. H. and H. Morrison, 2012: Sensitivity of a simulated squall line to horizontal resolution and parameterization of microphysics. *Monthly Weather Review*, **140**, 202-225, doi: 10.1175/MWR-D-11-00046.1.

Bryan, G. H., J. C. Wyngaard, and J. M. Fritsch, 2003: Resolution requirements for the simulation of deep moist convection. *Monthly Weather Review*, **131**, 2394-2416, doi: 10.1175/1520-0493(2003)131<2394:RRFTSO>2.0.CO;2.

Burghardt, B. J., C. Evans, and P. J. Roebber, Assessing the predictability of convection initiation in the high plains using an object-based approach. *Weather and Forecasting*, **29**,403-416, doi: 10.1175/WAF-D-13-00089.1.

Bytheway, J. L. and C. D. Kummerow, 2015: Toward an object-based assessment of high-resolution forecasts of long lived convective precipitation in the central US. *Journal of Advances in Modeling Earth Systems*, **7**, 1248-1264, doi: 10.1002/2015MS000497.

Cai, H. and R. E. Dumais Jr., 2015: Object-based evaluation of a numerical weather prediction model's performance through forecast storm characteristic analysis. *Weather and Forecasting*, **30**, 1451-1468, doi:10.1175/WAF-D-15-0008.1.

Craig, G. C., C. Keil, and D. Leuenberger, 2012: Constraints on the impact of radar rainfall data assimilation on forecasts of cumulus convection. *Quarterly Journal of the Royal Meteorological Society*, **138**, 340-352, doi: 10.1002/qj.929.

Davis, C., B. Brown, and R. Bullock, 2006: Object-based verification of precipitation forecasts. Part 1: Methodology and application to mesoscale rain areas. *Monthly Weather Review*, **134**, 1772-1784, doi: 10.1175/MWR3145.1.

Davis, C. A., B. G. Brown, R. Bullock, and J. Halley-Gotway, 2009: The method for object-based diagnostic evaluation (MODE) applied to numerical forecasts from the 2005 NSSL/SPC spring program. *Weather and Forecasting*, **24**, 1252-1267, doi: 10.1175/2009WAF2222241.1.

Dixon, M. and G. Wiener, 1993: TITAN: Thunderstorm identification, tracking, analysis, and nowcasting – A radar-based methodology. *Journal of Atmospheric and Oceanic Technology*, **10**, 785-797, doi: 10.1175/1520-0426(1993)010<0785:TTITAA>2.0.CO;2.

Ebert, E. E. and W. A. Gallus Jr., 2009: Toward better understanding of the Contiguous Rain Area (CRA) method for spatial forecast verification. *Weather and Forecasting*, **24**, 1401-1415, doi: 10.1175/2009WAF2222252.1.

Ebert, E. E., J. E. Janowiak, and C. Kidd, 2007: Comparison of near-real-time precipitation estimates from satellite observations and numerical models. *Bulletin of the American Meteorological Society*, **88**, 47-64, doi: 10.1175/BAMS-88-1-47.

Ebert, E. E. and J. L. McBride, 2000: Verification of precipitation in weather systems: determination of systematic errors. *Journal of Hydrology*, **239**, 179-202, doi: 10.1016/S0022-1694(00)00343-7.

Errico, R. M., P. Bauer, and J.-F. Mahfouf, 2007: Issues regarding the assimilation of cloud and precipitation data. *Journal of the Atmospheric Sciences*, **64**, 3785-3798, doi:10.1175/2006JAS2044.1.

ESRL, 2015: High-Resolution Rapid Refresh (HRRR). [Available at http://rapidrefreh.noaa.gov/hrrr].

Fritsch, J. M. and R. E. Carbone, 2004: Improving quantitative precipitation forecasts in the warm season. *Bulletin of the American Meteorological Society*, **85**, 955-965, doi:10.1175/BAMS-85-7-955.

Fritsch, J. M., R. J. Kane and C. R. Chelius, 1986: The contribution of mesoscale convective weather systems to the warm-season precipitation in the United States. *Journal of Climate and Applied Meteorology*, **25**, 1333-1345, doi: /10.1175/1520-0450(1986)025<1333:TCOMCW>2.0.CO;2.

Gallus, W. A. Jr. and M. Pfeifer, 2008: Intercomparison of simulations using 5 WRF microphysical schemes with dual-Polarization data for a German squall line. *Advances in Geosciences*, **16**, 109-116, doi: 10.5194/adgeo-16-109-2008.

Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Weather and Forecasting*, **24**, 1416-1430, doi: 10.1175/2009WAF2222269.1.

Gilmore, M. S., J. M. Straka, and E. N. Rasmussen, 2004: Precipitation uncertainty due to variations in precipitation particle parameters within a simple microphysics scheme. *Monthly Weather Review*, **132**, 2610-2627, doi: 10.1175/MWR2810.1.

Han, M., S. A. Braun, T. Matsui, and C. R. Williams, 2013: Evaluation of cloud microphysics schemes in simulations of a winter storm using radar and radiometer measurements. *Journal of Geophysical Research*, **118**, 1401-1419, doi: 10.1002/jgrd.50115.

Haynes, J. M., R. T. Marchand, Z. Luo, A. Bodas-Salcedo, and G. L. Stephens, 2007: A multipurpose radar simulation package: QuickBeam. *Bulletin of the American Meteorological Society*, **88**, 1723-1727, doi: 10.1175/BAMS-88-11-1723.

Hou, A. Y. and S. Q. Zhang, 2007: Assimilation of precipitation information using column model physics as a weak constraint. *Journal of the Atmospheric Sciences*, **64**, 3865-3878, doi: 10.1175/2006JAS2028.1.

Knievel, J. C., D. A. Ahijevych, and K. W. Manning, 2004: Using temporal modes of rainfall to evaluate the performance of a numerical weather prediction model. *Monthly Weather Review*, **132**, 2995-309, doi; 10.1175/MWR2828.1

Kummerow, C. 1993: On the accuracy of the Eddington Approximation for radiative transfer in the microwave frequencies. *Journal of Geophysical Research*, **98**, 2757-2765, doi: 10.1029/92JD02472.

Kummerow, C., Y. Hong, W. S. Olson, S. Yang, R. F. Adler, J. McCollum, R. Ferraro, G. Petty, D.-B. Shin, and T. T. Wilheit, 2001: The evolution of the Goddard Profiling Algorithm (GPROF) for rainfall estimation from passive microwave sensors. *Journal of Applied Meteorology*, **40**, 1801-1820, doi: 10.1175/1520-0450(2001)040<1801:TEOTGP>2.0.CO;2.

Kummerow, C., W. S. Olson, and L. Giglio, 1996: A simplified scheme for obtaining precipitation and vertical hydrometeor profiles from passive microwave sensors. *IEEE Transactions on Geoscience and Remote Sensing*, **54**, 1213-1232, doi: 10.1109/36.536538.

Kummerow, C., D. L. Randel, M. Kulie, N.-Y. Wang, R. Ferraro, S. J. Munchak, and V. Petkovic, 2015: The evolution of the Goddard Profiling Algorithm to a fully parametric scheme. *Journal of Atmospheric and Oceanic Technology*, **32**, 2265-2280, doi: 10.1175/JTECH-D-15-0039.1.

Lack, S. A., G. L. Limpert, and N. I. Fox, 2010: An object-oriented multiscale verification scheme. *Weather and Forecasting*, **25**, 79-92, doi: 10.1175/2009WAF2222245.1.

Lean, H. W., P. A. Clark, M. Dixon, N. M. Roberts, A. Fitch, R. Forbes, and C. Halliwell, 2008: Characteristics of high-resolution versions of the Met Office Unified Model for forecasting convection over the United Kingdom. *Monthly Weather Review*, **136**, 3408-3424, doi: 10.1175/2008MWR2332.1.

Lin, C., S. Vasić, A. Kilambi, B. Turner, and I. Zawadzki, 2005: Precipitation forecast skill of numerical weather prediction models and radar nowcasts. *Geophysical Research Letters*, **32**, L14801, doi: 10.1029/2005GL023451.

Lin, Y., B. A. Colle, C. Woods, and B. F. Smull, (2006), Verification of WRF for the 4-5 December 2001 IMPROVE-2 event over the central Oregon Cascades. Preprints, *7th WRF Users' Workshop*, Boulder, CO, National Center for Atmospheric Research, P2-2.

Lin, Y., and K. E. Mitchell, 2005: The NCEP StageII/IV hourly precipitation analyses: development and applications. Preprints, *19th Conf. on Hydrology*, American Meteorological Society, San Diego, CA, 9-13 January 2005, Paper 1.2.

Lopez, P., 2007: Cloud and precipitation parameterizations in modeling and variational data assimilation: A review. *Journal of the Atmospheric Sciences*. **64**, 3766-3784, doi: 10.1175/2006JAS2030.1.

Luo, Z. J., J. Jeyaratnam, S. Iwasaki, H. Takahashi, and R. Anderson, 2014: Convective vertical velocity and cloud internal vertical structure: An A-Train perspective. *Geophysical Research Letters*, **41**, 1-7, doi: 10.1002/2013GL058922.

Maddox, R. A., J. Zhang, J. J. Gourley, and K. W. Howard, 2002: Weather radar coverage over the contiguous United States. *Weather and Forecasting*, **17**, 927-934, doi: 10.1175/1520-0434(2002)017<0927:WRCOTC>2.0.CO;2.

Marzban, C., S. Sandgathe, H. Lyons, and N. Lederer, 2009: Three spatial verification techniques: cluster analysis, variogram, and optical flow. *Weather and Forecasting*, **24**, 1457-1471, doi: 10.1175/2009WAF2222261.1.

Morrison, H., J. A. Curry, and V. I. Khvorostyanov 2005: A new double moment microphysics parameterization for application in cloud and climate models. Part I: Description. *Journal of Atmospheric Science*, **62**, 1665–1677, doi:10.1175/JAS3446.1.

Morrison, H., G. Thompson, and V. Tatarskii 2009: Impact of cloud microphysics on the development of trailing stratiform precipitation in a simulated squall line: Comparison of one- and two- moment schemes, *Monthly Weather Review*, **137**, 991–1007, doi:10.1175/2008MWR2556.1.

Morrison, H. and J. Milbrandt, 2011: Comparison of two-moment bulk microphysics schemes in idealized supercell thunderstorm simulations. *Monthly Weather Review*, **139**, 1103-1130, doi:10.1175/2010MWR3433.1.

Nelson, B. R., O. P. Prat, D.-J. Seo, and E. Habib, 2016: Assessment and implications of NCEP Stage IV quantitative precipitation estimates for product intercomparisons. *Weather and Forecasting*, **31**, 371-394, doi:10.1175/WAF-D-14-00112.1.

Olson, W. S., and H. Masunaga 2016: GPM combined radar-radiometer precipitation algorithm theoretical basis document (Version 4), [Available at https://pps.gsfc.nasa.gov/Documents/Combined_algorithm_ATBD.V04.rev.pdf].

Prat, O. P. and B. R. Nelson, 2015: Evaluation of precipitation estimates over CONUS derived from satellite, radar, and rain gauge data sets at daily to annual scales (2002-2012). *Hydrology and Earth System Sciences*, **19**, 2037-2056, doi: 10.5194/hess-19-2037-2015.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, **136**, 78-97, doi: 10.1175/2007MWR2123.1.

Rogers, R. F., J. M. Fritsch, and W. C. Lambert, 2000: A simple technique for using radar data in the dynamic initialization of a mesoscale model. *Monthly Weather Review*, **128**, 2560-2574, doi: 10.1175/1520-0493(2000)128<2560:ASTFUR>2.0.CO;2.

Sheng, C., S. Gao, and M. Xue, 2006: Short-range prediction of a heavy precipitation event by assimilating Chinese CINRAD-SA radar reflectivity data using complex cloud analysis. *Meteorology and Atmospheric Physics*, **94**, 167-183, doi: 10.1007/s00703-005-0177-0.

Shrestha, D. L., D. E. Robertson, Q. J. Wang, T. C. Pagano, and H. A. P. Hapuarachchi, 2013: Evaluation of numerical weather prediction model precipitation forecasts for short-term streamflow forecasting purpose. *Hydrology and Earth System Sciences*, **17**, 1913-1931, doi: 10.5194/hess-17-1913-2013.

Smalley, M., T. L'Ecuyer, M. Lebsock, and J. Haynes, 2014: A comparison of precipitation occurrence from the NCEP Stage IV QPE product and the CloudSat cloud profiling radar. *Journal of Hydrometeorology*, **15**, 444-458, doi: 10.1175/JHM-D-13-048.1.

Stensrud, D. J., and Coauthors, 2009: Convective-scale warn-on-forecast system, a vision for 2020. *Bulletin of the American Meteorological Society*, **90**, 1487-1499, doi: 10.1175/2009BAMS2795.1.

Stevenson, S. N. and R. S. Schumacher, 2014: A 10-year survey of extreme rainfall events in the central and eastern United States using gridded multisensory precipitation analyses. *Monthly Weather Review*, **142**, 3147-3162, doi:10.1175/MWR-D-13-00345.1.

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. *Weather and Forecasting*, **28**, 119-138. doi:10.1175/WAF-D-12-00022.1.

Sugimoto, S., N. A. Crook, J. Sun, Q. Xiao, and D. M. Barker, 2009: An examination of WRF 3DVAR radar data assimilation on its capability in retrieving unobserved variables and forecasting precipitation through observing system simulation experiments. *Monthly Weather Review*, **137**, 4011-4029, doi: 10.1175/2009MWR2839.1.

Sun, J., and Coauthors, 2014: Use of NWP for nowcasting convective precipitation: Recent progress and challenges. *Bulletin of the American Meteorological Society*, **95**, 409-426, doi: 10.1175/BAMS-D-11-00263.1.

Szoke, E. J., E. J. Zipser, and D. P. Jorgensen, 1986: A radar study of convective cells in mesoscale systems in GATE. Part I: Vertical profile statistics and comparison with hurricanes. *Journal of the Atmospheric Sciences*, **43**, 182-198, doi: 10.1175/1520-0469(1986)043<0182:ARSOCC>2.0.CO;2.

Tao, W.-K., and J. Simpson 1993: Goddard cumulus ensemble model. Part I: Model description. *Terrestrial Atmospheric and Oceanic Sciences*, **4**, 35–72.

Tao, W.-K., and Coauthors, 2003: Microphysics, radiation and surface processes in the Goddard Cumulus Ensemble (GCE) model, *Meteorology and Atmospheric Physics*, **82**, 97–137, doi:10.1007/s00703-001-0594-7.

Thompson, G. and T. Eidhammer, 2014: A study of aerosol impacts on clouds and precipitation development in a large winter cyclone. *Journal of the Atmospheric Sciences*, **71**, 3636-3658, doi: 10.1175/JAS-D-13-0305.1.

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. *Monthly Weather Review*, **136**, 5095-5115, doi: 10.1175/2008MWR2387.1.

Warner, T. T., 2011: Quality assurance in atmospheric modeling. *Bulletin of the American Meteorological Society*, **92**, 1601-1610, doi: 10.1175/BAMS-D-11-00054.1.

Weisman, M. L., W. C. Skamarock, and J. B. Klemp, 1997: The resolution dependence of explicitly modeled convective systems. *Monthly Weather Review*, **125**, 527-548, doi: 10.1175/1520-0493(1997)125<0527:TRDOEM>2.0.CO;2.

Wernli, H., C. Hofmann, and M. Zimmer, 2009: Spatial forecast verification methods intercomparison project: Application of the SAL technique. *Weather and Forecasting*, **24**, 1472-1484, doi: 10.1175/2009WAF2222271.1.

Wernli, H., M. Paulat, M. Hagen, and C. Frei, 2008: SAL – A novel quality measure for the verification of quantitative precipitation forecasts. *Monthly Weather Review*, **136**, 4470-4487, doi:10.1175/2008MWR2415.1.

Weygandt, S. and S. Benjamin 2007: Radar reflectivity-based initialization of precipitation systems using a diabatic digital filter within the Rapid Update Cycle. *Preprints: 18th AMS Conference on Numerical Weather Prediction*, Park City, UT.

Weygandt, S. S., S. G. Benjamin, T. G. Smirnova, and J. M. Brown 2008: Assimilation of radar reflectivity data using a diabatic digital filter within the Rapid Update Cycle. *Preprints, AMS 12th Conf. IOAS-AOLS*, New Orleans, LA.

Wu, D., X. Dong, B. Xi, Z. Feng, A. Kennedy, G. Mullendore, M. Gilmore, and W.-K. Tao 2013: Impacts of microphysical scheme on convective and stratiform characteristics in two high precipitation squall line events, *Journal of Geophysical Research*, **118**, 11,119–11,135, doi:10.1002/jgrd.50798.

Xiao, Q., and J. Sun, 2007: Multiple-radar data assimilation and short-range quantitative precipitation forecasting of a squall line observed during IHOP_2002. *Monthly Weather Review*, **135**, 3381-3404, doi: 10.1175/MWR3471.1.

Zipser, E. J., and K. R. Lutz, 1994: The vertical profile of radar reflectivity of convective cells: A strong indicator of storm intensity and lightning probability. *Monthly Weather Review*, **122**, 1751-1759, doi: /10.1175/1520-0469(1986)043<0182:ARSOCC>2.0.CO;2.

APPENDIX A:

LIST OF ACRONYMS

2BCMB	Level 2B Combined
AMB	Assimilation and Modeling Branch
ARW	Advanced Research WRF
CAPE	Convective Available Potential Energy
CDF	Cumulative Distribution Function
CIN	Convective Inhibition
CONUS	Contiguous United States
CRA	Contiguous Raining Area
CRM	Cloud Resolving Model
CSI	Critical Success Index
DA	Data Assimilation
DPR	Dual-frequency Precipitation Radar
DSD	Drop size distribution
EnKF	Ensemble Kalman Filter
FAR	False Alarm Ratio
FH	Forecast Hour
FSS	Fractions Skill Score
GFS	Global Forecast System
GHz	Gigahertz
GMI	GPM Microwave Imager

GPM	Global Precipitation Measurement
GPROF	Goddard Profiling Algorithm
HRAP	Hydrologic Rainfall Analysis Project
HRRR	High Resolution Rapid Refresh
HSS	Heidke Skill Score
IQR	Interquartile Range
IR	Infrared
JAXA	Japan Aerospace Exploration Agency
KaPR	Ka band Precipitation Radar
KuPR	Ku band Precipitation Radar
MCS	Mesoscale Convective System
MODE	Method of Object-Based Deterministic Evaluation
NASA	National Aeronautics and Space Agency
NCAR	National Center for Atmospheric Research
NCEP	National Center for Environmental Prediction
NDVAR	N-Dimensional Variational
NOAA	National Oceanic and Atmospheric Administration
NOMADS	National Operational Model Archive Distribution System
NWP	Numerical Weather Prediction
NWS	National Weather Service
PDF	Probability Distribution Function
PMM	Precipitation Measurements Mission
POD	Probability of Detection

PSD	Particle Size Distribution
QPF	Quantitative Precipitation Forecast
RFC	River Forecast Center
RMSE	Root Mean Square Error
SAL	Structure Amplitude Location
Tbs	Brightness Temperatures
TI	Total Interest
TITAN	Thunderstorm Identification, Tracking, Analysis and Nowcasting
TPW	Total Precipitable Water
US	United States
UTC	Coordinated Universal Time
WRF	Weather Research and Forecasting
WSR-88D	Weather Surveillance Radar