UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

A TWO-PART STUDY IN THE EVALUATION OF QUANTITATIVE PRECIPITATION FORECASTS FROM CONVECTION-ALLOWING MODELS

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE IN METEOROLOGY

By

ESWAR R. IYER
Norman, Oklahoma
2017

A TWO-PART STUDY IN THE EVALUATION OF QUANTITATIVE
PRECIPITATION FORECASTS FROM CONVECTION-ALLOWING MODELS

A THESIS APPROVED FOR THE
SCHOOL OF METEOROLOGY

BY

____
Dr. Ming Xue, Chair

____
Dr. Adam Clark

____
Dr. Xuguang Wang

____
Dr. Harold Brooks

Acknowlegments

This work was made possible by a Presidential Early Career Award for Scientists and Engineers (PECASE). Additional support was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement #NA11OAR4320072, U.S. Department of Commerce. CAPS SSEF forecasts were supported by the NOAA Collaborative Science, Technology, and Applied Research (CSTAR) Program with supplementary support from NSF grant AGS-0802888, using computing resources at the National Science Foundation XSEDE National Institute of Computational Science (NICS) at the University of Tennessee.

I would also like to thank my adviser, Adam Clark as well as Ming Xue, Fanyou Kong, Burkely T. Gallo, and Christopher Melick. This thesis could not have been written without their contributions to the field.

Table of Contents

List of Tables

List of Figures

Abstract

Previous studies examining convection-allowing models (CAMs), as well as NOAA/Hazardous Weather Testbed Spring Forecasting Experiments (SFEs) have typically emphasized "day 1" (12-36 h) forecast guidance. These studies find a distinct advantage in CAMs relative to models that parameterize convection, especially for fields strongly tied to convection like precipitation. During the 2014 SFE, "day 2" (36-60 h) forecast products from a CAM ensemble provided by the Center for Analysis and Prediction of Storms at the University of Oklahoma were examined. Quantitative Precipitation Forecasts (QPFs) from the CAPS ensemble, known as the Storm Scale Ensemble Forecast (SSEF) system, are compared to NCEP's operational Short Range Ensemble Forecast (SREF) system, which provides lateral boundary conditions for the SSEF, to see if the CAM ensemble outperforms the SREF through forecast hours 36-60.

Equitable Threat Score (ETS) was computed for precipitation thresholds ranging from 0.10 in. to 0.75 in. for each SSEF and SREF member, as well as ensemble means, for 3 h accumulation periods. The ETS difference between the SSEF and SREF peaked during hours 36-42. Probabilistic forecasts were evaluated using the area under the receiver operating characteristic curve (ROC Area). The SSEF had higher values of ROC Area, especially at thresholds ≥ 0.50 in. Additionally, time-longitude diagrams of diurnally averaged rainfall were constructed for each SSEF/SREF ensemble member. Spatial correlation coefficients between forecasts and observations in time-longitude space indicated that the SSEF depicted the diurnal cycle much better than the SREF, which under-forecasted precipitation with a peak that had a 3 h phase lag. A minority of SREF members performed well.

One component of the 2016 NOAA/Hazardous Weather Testbed (HWT) Spring

Forecasting Experiment (SFE) examined the impact of radar data assimilation on convection-allowing model (CAM) ensemble forecasts using two similarly configured 10-member ensembles with (rad-ens) and without (norad-ens) radar data assimilation, which were provided by the Center of Analysis and Prediction of Storms and the National Severe Storms Laboratory, respectively. While previous works have quantified the impact of radar data assimilation on the skill of deterministic forecasts, the impact on ensemble forecast skill has yet to be examined.

Equitable threat scores (ETSs) were computed for precipitation thresholds ranging from 0.10 to 0.75 in., and neighborhood-based ETS (ETS_r) was computed for radii ranging from 8- to 60-km. The ETS difference between rad-ens and norad-ens peaked at forecast hour 3, but existed until forecast hours 12-15. As the radius in the ETS_r increased, the difference in ETS_r between the rad-ens and norad-ens increased, even out to hour 36. 3-h probabilistic QPFs were evaluated using the area under the ROC curve (ROC Area). Similar to ETS, ROC area results showed a difference in favor of the rad-ens out to forecast hour 15. The rad-ens had slightly greater variance and lower MSE out to hour 15, then both ensembles had nearly identical variance/MSE values from hours 15 to 36. Rank histograms were similar for each ensemble, and indicated over-forecasting. Most of the subjective evaluations filled out by 2016 SFE participants indicated that the positive effects of data assimilation lasted less than 12 h. A small minority of the respondents saw an advantage at lead times of 15-24 h.

Chapter 1: A Comparison of 36-60 Hour Forecasts from Convection-Allowing and Convection-Parameterizing Ensembles

i) Introduction

Historically, warm season quantitative precipitation forecasts (QPF) have been especially challenging for numerical weather prediction (NWP). While NWP forecasts for fields such as 500 hPa heights have improved, the skill of warm season QPF has exhibited little change over time (e.g., Fritsch et al. 1998). Improvements in warm season QPF would not only help with predictions of hazards such as flash floods (e.g., Vasiloff et al. 2007), which have accounted for roughly 2500 fatalities in the U.S. over the past half century (NOAA 2015), but would also benefit agriculture (e.g., through improved irrigation management), transportation industries, government, and emergency management (e.g., Sukovich et al. 2014). Recognizing the gap in skill for QPF relative to other variables, Roebber et al. (2004) discussed a number of avenues for closing this gap including increasing model forecasts to sufficient resolution to explicitly depict convection and utilizing ensembles to depict the high degree of forecast uncertainty often associated with convection. However, due to computational limitations, it has only been very recently that operational models with sufficient resolution to explicitly depict convection [hereafter referred to as convection-allowing models (CAMs)[1]] have become available, and assessing and improving their capabilities is a rich area of research.

Recent work has shown that CAMs provide advantages relative to models that parameterize convection for several aspects of QPF. The advantages include an improved depiction of the diurnal precipitation cycle (Clark et al. 2007, 2009; Weisman et al. 2008;

[1] It is generally believed that a maximum of 4-km grid-spacing is required for CAMs to adequately resolve the bulk circulations within organized convective systems (e.g., Weisman et al. 1997).

and Berenguer et al. 2012) and better representation of observed convective mode (e.g., Done et al. 2004; Kain et al. 2006). Additionally, Clark et al. (2009) found a distinct advantage using objective verification metrics in a small-membership CAM-based ensemble relative to a much larger convection-parameterizing ensemble. Furthermore, Roberts and Lean (2008), Schwartz et al. (2009), and Clark et al. (2010) also show improved precipitation forecasts in CAMs relative to coarser models, but illustrate that in some cases to see the improvements, spatial scales larger than the model grid-spacing need to be considered using neighborhood-based objective metrics.

Other recent work has examined how CAM guidance is perceived relative to convection-parameterizing guidance by forecasters in simulated operational forecasting environments. For example, during the 2010 and 2011 NOAA/Hazardous Weather Testbed (HWT) Spring Forecasting Experiments (SFEs), QPF products from a CAM-based ensemble were compared to guidance from the operational Short Range Ensemble Forecast (SREF) system by a group of participants led by Weather Prediction Center (WPC; formerly the Hydro-meteorological Prediction Center) forecasters. The WPC-led group found such an advantage in the CAM-based ensemble that they viewed the guidance as "transformational" to warm-season QPF (Clark et al. 2012). Evans et al. (2014) also conducted an experiment in a simulated forecasting environment; finding that forecasters felt CAM-based guidance added perceived value in QPF relative to operational models that parameterized convection for an extreme heavy rainfall event related to a tropical storm.

From the aforementioned studies, it has become clear that CAM-based guidance provides important gains in QPF relative to convection-parameterizing guidance, but little work has been done to extend these comparisons past the day 1 (12-36 h) forecast period. For example, until April 2014, NOAA/HWT SFEs have only focused on CAM-based guidance extending to 36 h. However, starting in 2014, CAPS ran the SSEF system to 60 h to test the performance through the day 2 period. At such long forecast lead times, infiltration of lateral boundary conditions (LBCs) begins to have a relatively large influence on the forecasts, thus, whether a CAM-based ensemble can still maintain its advantage warrants further investigation. Therefore, the main purpose of this study is to objectively analyze these forecasts to evaluate whether the advantages relative to convection-parameterizing guidance translate to these longer lead times. For this purpose, the 60 h SSEF system forecasts are compared to those from the 16 km grid-spacing SREF ensemble using a variety of objective metrics for evaluating deterministic and probabilistic forecasts.

The remainder of the study is organized as follows: Section 2 presents information on datasets and methods, Section 3 presents results along with graphs of these metrics, and Section 4 provides a summary and conclusions.

ii) Data and methodology

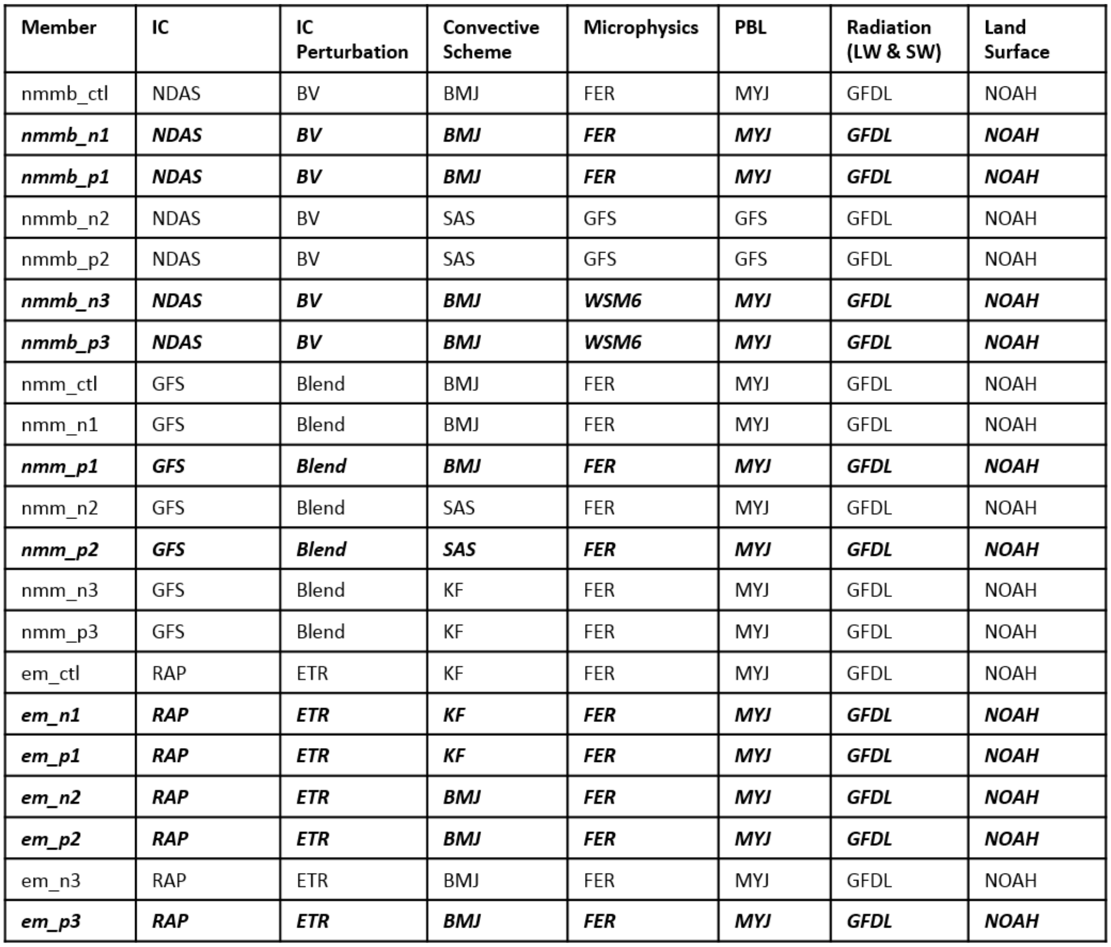
a) SSEF and SREF Ensemble Description

Forecast precipitation for 3 h periods was examined from the convection-allowing SSEF and the convection-parameterizing SREF, which had 4 km and 16 km grid spacing configurations, respectively. The SSEF system was provided to support the 2014 NOAA/HWT Spring Forecasting Experiment and consists of model integrations conducted from 21 April through 6 June. The 30 days of model integrations that were used for this study are as follows: 24-29 April, 1-2 May, 5-9 May, 12-16 May, 19-23 May, 26-30 May, and 2-3 June. The SREF system 2100 UTC initializations were used and the SSEF was initialized 3 hours later at 0000 UTC (Kong et al. 2014). Because CAPS did not run the SSEF ensemble members during most weekend days in April, May, and June, 30 days during the 2014 SFE period had the full datasets available from both ensembles. Observed precipitation data was derived from the NCEP Stage IV dataset (Baldwin and Mitchell 1997), which was on a 4 km grid.

The SREF ensemble data (Du et al. 2014) were available on a 32 km grid from NCEP's archives. At the time, the 21 member ensemble consisted of 7 members from the Non-hydrostatic Multiscale Model on the B grid (NMMB) (Janjić 2005, 2010; Janjić and Black 2007; Janjić et al. 2011; Janjić and Gall 2012), 7 members from the WRF Non-hydrostatic Multiscale Model (NMM) (Janjić 2003), and 7 members from the Advanced Research WRF model (ARW; Skamarock et al. 2008). Each set of 7 members had one control member, 3 positive perturbations, and 3 negative perturbations. To generate these perturbations, the NMMB members used a breeding cycle (e.g., Toth and Kalnay 1997) initialized at 2100 UTC to create perturbations, which are added and subtracted from the control, creating 6 different perturbed analyses. The ARW members used Ensemble Transform with rescaling (ETR; Ma et al. 2014) to generate perturbations, while the NMM members used a blend of ETR and breeding.

Physics parameterizations in the SREF system consisted of the MYJ planetary boundary layer scheme (Mellor and Yamada 1982; Janjić 1990, 2002), the MRF planetary boundary layer scheme (Troen and Mahrt 1986; Hong and Pan 1996), and the Noah land surface model (Ek et al. 2003). Surface layer parameterizations consisted of the MYJ surface layer scheme (Mellor and Yamada 1982; Janjić 2002) and the Monin/Obukhov scheme with the Janjić ETA model (Monin and Obukhov 1954; Paulson 1970; Dyer and Hicks 1970; Webb 1970; Janjić 1996, 2002). Radiation schemes consisted of the GFDL shortwave (Lacis and Hansen 1974) and GFDL longwave (Fels and Schwarzkopf 1975; Schwarzkopf and Fels 1991). Microphysical parameterizations consisted of the scheme used in the GFS (Zhou and Carr 1997), the Ferrier scheme (Ferrier et al. 2002), and the WSM6 scheme (Ferrier et al. 2002; Hong and Lim 2006). The convection parameterization schemes consisted of the KF (Kain and Fritsch 1998), the BMJ (Betts and Albrecht 1987; Janjić 2002), and the SAS (Arakawa 2004). Full specifications of the SREF are given in Table 1.

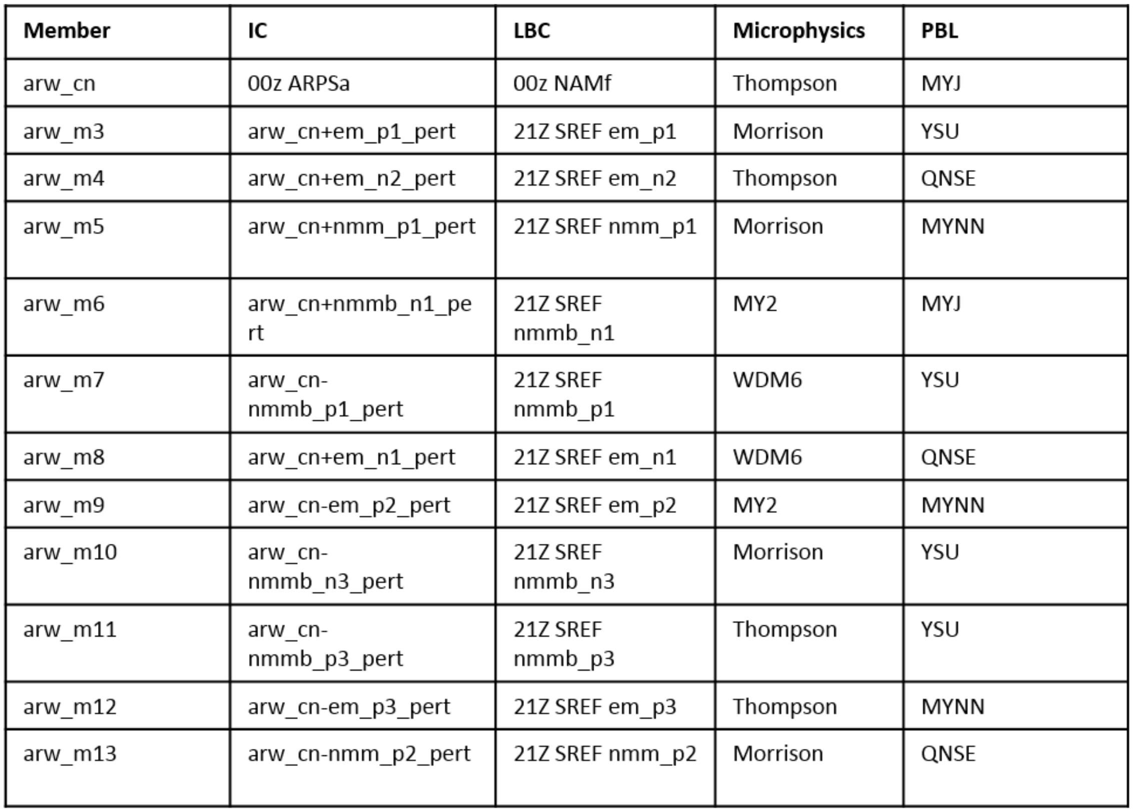
Table 1 Model specifications for all 21 operational SREF members, divided into 3 models (NMMB, NMM, ARW) of 7 members each (1 control, 6 perturbed). Bolded and italicized text indicates a member that supplied LBCs to the SSEF. NDAS refers to the NAM Data Assimilation System, GFS refers to the Global Forecast System, and RAP is the Rapid Refresh Model. BV stands for Breeding Vector, ETR stands for Ensemble Transform with Rescaling, and Blend refers to a blend of BV and ETR. BMJ refers to the Betts-Miller-Janjić convective scheme, SAS refers to the Simplified Arakawa-Schubert convective scheme, and KF refers to the Kain-Fritsch convective scheme. FER refers to the Ferrier microphysics scheme and WSM6 refers to the WRF Single-Moment 6 class scheme. MYJ refers to the Mellor-Yamada-Janjić planetary boundary layer scheme. GFDL refers to the Geophysical Fluid Dynamics Laboratory radiation scheme. References are provided in the text.



Member	IC	IC Perturbation	Convective Scheme	Microphysics	PBL	Radiation (LW & SW)	Land Surface
nmmb_ctl	NDAS	BV	BMJ	FER	MYJ	GFDL	NOAH
nmmb_n1	***NDAS***	***BV***	***BMJ***	***FER***	***MYJ***	***GFDL***	***NOAH***
nmmb_p1	***NDAS***	***BV***	***BMJ***	***FER***	***MYJ***	***GFDL***	***NOAH***
nmmb_n2	NDAS	BV	SAS	GFS	GFS	GFDL	NOAH
nmmb_p2	NDAS	BV	SAS	GFS	GFS	GFDL	NOAH
nmmb_n3	***NDAS***	***BV***	***BMJ***	***WSM6***	***MYJ***	***GFDL***	***NOAH***
nmmb_p3	***NDAS***	***BV***	***BMJ***	***WSM6***	***MYJ***	***GFDL***	***NOAH***
nmm_ctl	GFS	Blend	BMJ	FER	MYJ	GFDL	NOAH
nmm_n1	GFS	Blend	BMJ	FER	MYJ	GFDL	NOAH
nmm_p1	***GFS***	***Blend***	***BMJ***	***FER***	***MYJ***	***GFDL***	***NOAH***
nmm_n2	GFS	Blend	SAS	FER	MYJ	GFDL	NOAH
nmm_p2	***GFS***	***Blend***	***SAS***	***FER***	***MYJ***	***GFDL***	***NOAH***
nmm_n3	GFS	Blend	KF	FER	MYJ	GFDL	NOAH
nmm_p3	GFS	Blend	KF	FER	MYJ	GFDL	NOAH
em_ctl	RAP	ETR	KF	FER	MYJ	GFDL	NOAH
em_n1	***RAP***	***ETR***	***KF***	***FER***	***MYJ***	***GFDL***	***NOAH***
em_p1	***RAP***	***ETR***	***KF***	***FER***	***MYJ***	***GFDL***	***NOAH***
em_n2	***RAP***	***ETR***	***BMJ***	***FER***	***MYJ***	***GFDL***	***NOAH***
em_p2	***RAP***	***ETR***	***BMJ***	***FER***	***MYJ***	***GFDL***	***NOAH***
em_n3	RAP	ETR	BMJ	FER	MYJ	GFDL	NOAH
em_p3	***RAP***	***ETR***	***BMJ***	***FER***	***MYJ***	***GFDL***	***NOAH***

The SSEF system was generated using the WRF model (Skamarock et al. 2008) run by the Center for Analysis and Prediction of Storms (CAPS) for the 2014 NOAA/HWT Spring Forecasting Experiment (Kong et al. 2014). During 2014, the SSEF system had 20 members with 4 km grid-spacing that were initialized on weekdays at 0000 UTC and integrated 60 h over a CONUS domain from late April to the beginning of June. Initial condition (IC) analysis background and LBCs (3-h updates) for the control member were taken from the NAM analyses and forecasts, respectively. Radial velocity and reflectivity data from up to 140 Weather Surveillance Radar-1988 Doppler (WSR-88D) and other high-resolution observations were assimilated into the ICs using the ARPS three-dimensional variational data assimilation (3DVAR; Xue et al. 2003; Gao et al. 2004) data and cloud analysis system (Xue et al. 2003; Hu et al. 2006; Gao and Xue 2008). IC perturbations were derived from evolved (through 3 h) perturbations of 2100 UTC initialized members of the SREF system and added to the control member ICs. For each perturbed member, the forecast of the SREF member used for the IC perturbations was also used for the LBCs. For the purposes of this study, only the 12 members comprised of the control member and the 11 members with IC/LBC perturbations were utilized. The other 8 SSEF members were run with the same ICs/LBCs as the control with different physics parameterizations to study physics sensitivities.

Table 2 Configurations for 12 out of the 20 SSEF members. The abbreviations in the table are described in the text.



Member	IC	LBC	Microphysics	PBL
arw_cn	00z ARPSa	00z NAMf	Thompson	MYJ
arw_m3	arw_cn+em_p1_pert	21Z SREF em_p1	Morrison	YSU
arw_m4	arw_cn+em_n2_pert	21Z SREF em_n2	Thompson	QNSE
arw_m5	arw_cn+nmm_p1_pert	21Z SREF nmm_p1	Morrison	MYNN
arw_m6	arw_cn+nmmb_n1_pert	21Z SREF nmmb_n1	MY2	MYJ
arw_m7	arw_cn-nmmb_p1_pert	21Z SREF nmmb_p1	WDM6	YSU
arw_m8	arw_cn+em_n1_pert	21Z SREF em_n1	WDM6	QNSE
arw_m9	arw_cn-em_p2_pert	21Z SREF em_p2	MY2	MYNN
arw_m10	arw_cn-nmmb_n3_pert	21Z SREF nmmb_n3	Morrison	YSU
arw_m11	arw_cn-nmmb_p3_pert	21Z SREF nmmb_p3	Thompson	YSU
arw_m12	arw_cn-em_p3_pert	21Z SREF em_p3	Thompson	MYNN
arw_m13	arw_cn-nmm_p2_pert	21Z SREF nmm_p2	Morrison	QNSE

Because a subset of the 21 SREF member forecasts are used as LBCs for the SSEF members (except in the control), the two systems are inherently linked to each other. With the LBCs infiltrating into much of the domain interior, particularly by the latter half of the 60 h forecasts, we are essentially testing whether—given similar driving data (i.e., the LBCs)—the convection-allowing grid spacing can still provide an advantage. Table 2 shows detailed specifications of the 12 SSEF members used for this study. 9 out of the 12 members used the Noah land surface model (Ek et al. 2003) as was used in the SREF ensemble. PBL schemes include the MYJ, MYNN (Nakanishi 2000, 2001; Nakanishi and Niino 2004, 2006), YSU (Noh et al. 2003), and QNSE (Sukoriansky et al. 2005) schemes. The microphysical parameterization consisted of the Thompson scheme (Thompson et al. 2004), Morrison scheme (Morrison 2005), the WRF double moment 6-class scheme (Morrison et al. 2005; Lim and Hong 2010), and the double moment Milbrandt and Yau scheme (Milbrandt and Yau 2005).

Before any verification metrics were computed, the SREF, SSEF, and 3 h observed NCEP Stage IV precipitation data were interpolated to the 32-km grid of the SREF using a neighbor budget interpolation (e.g. Accadia et al. 2003). In addition, the SREF has a larger domain than the SSEF, which only includes the CONUS. So, a mask was used to consider only points within the SSEF domain, east of the Rocky Mountains over land, and only in the United States (Figure 1). This is due to the relative lack of reliable WSR-88D radar observations over the mountains, water, and outside the United States. The area of analysis is displayed in Figure 1.

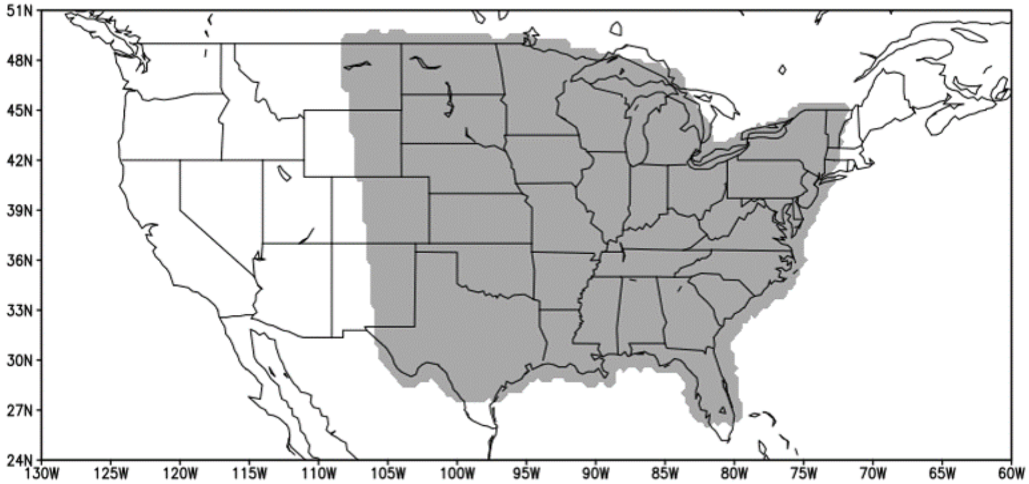


Figure 1 Analysis domain (in gray) for both of the ensembles.

b) Forecast Evaluation Metrics

The first metric that was used to evaluate the precipitation forecasts from each ensemble was the Equitable Threat Score (ETS; Schaefer 1990). ETS measures the fraction of observed and/or forecast events that were correctly predicted, adjusted for hits associated with random chance. The ETS was calculated using contingency table elements computed from every grid point in the 32 km grid spacing analysis domain for each ensemble member every 3 hours as follows: ETS = $(H - H_{cha})/(H + FA + M + H_{cha})$, where H represents a hit (model correctly forecasted precipitation to exceed a certain threshold), H_{cha} represents the number of hits expected by chance, FA represents the number of false alarms (model forecasted precipitation to exceed a certain threshold, but the observed precipitation did not exceed that threshold), and M represents a miss (model did not forecast precipitation to exceed a certain threshold, but the observed precipitation did exceed that threshold). An ETS score of 1 is perfect, and a score below zero represents no forecast skill.

In addition to computing ETS for individual ensemble members, ETS was also computed for ensemble mean precipitation forecasts. Ensemble means were computed using the probability matching technique (Ebert 2001). This technique assumes that the best spatial representation of the precipitation field is given by the ensemble mean, and that the best probability density function (PDF) of rain rates is given by the ensemble member QPFs for all *n* ensemble members. To compute the probability matched mean, the precipitation forecasts from the ensemble members for every grid point are ranked in order from largest to smallest, keeping every *n*th value. The precipitation forecasts from the ensemble mean forecast are similarly ranked from largest to smallest, keeping every value. Then, the grid point with the highest value in the ensemble mean QPF field is reassigned to the highest QPF value in the ensemble member QPF distribution. Then the grid point with the second highest value in the ensemble mean QPF field is reassigned to the second highest value in the ensemble member QPF distribution. This process is then repeated for all of the rankings, ending with the lowest. ETS scores were then calculated from these probability matched means for both the SSEF and SREF for forecast hours 3-60.

Finally, hypothesis testing was conducted to evaluate whether the SSEF forecasts were significantly more accurate than the SREF forecasts. The hypothesis testing was conducted for all 20 accumulation periods spanning 60 forecast hours using the resampling method of Hamill (1999). To apply this method, the test statistic used to look at the difference in accuracy of the 3 h precipitation forecast ending at hour hr (where hr is a given forecast hour) is (ETS_{SSEFhr} – ETS_{SREFhr}). The null hypothesis, H_o, is ETS_{SSEFhr} – ETS_{SREFhr} = 0.00. The alternative hypothesis, H_a, is $ETS_{SSEFhr} - ETS_{SREFhr} \neq 0.00$. The significance level used was $\alpha = 0.05$ and resampling was done 1000 times for each hypothesis test. In addition, the forecasts were corrected for bias before the resampling hypothesis tests were conducted. This was done for each threshold by calculating the average bias of the two ensemble means, and then finding the precipitation threshold at which the bias of each ensemble equals the average bias from the original precipitation threshold. More detailed information about the resampling method can be found in Hamill (1999).

In addition to the deterministic forecasts, probabilistic quantitative precipitation forecasts (PQPFs) for both the SSEF and SREF were generated for all five precipitation thresholds for 3 hour QPF. Probabilities were computed using the ratio of members that exceeded the specified threshold to the total number of members. The probabilistic forecasts were evaluated using the area under the Receiver Operating Characteristics Curve (ROC Area; Mason 1982), which measures the ability to discriminate between events (exceedances of specified threshold) and non-events (failure to exceed specified threshold). It is calculated by computing the area under a curve constructed by plotting the probability of detection (POD) versus the probability of false detection (POFD). The area under the curve is computed using the trapezoidal method (Wandishin et al. 2001) using the probabilities 0.05 to 0.95, in increments of 0.05. Values of ROC Area range from 0 to 1, with a value of 0.5 indicating no forecast skill and values above 0.5 indicating positive forecast skill. Similar to ETS, statistical significance was tested for the ROC areas using the resampling method (Hamill 1999).

To analyze the diurnal precipitation cycle, the 3 h QPF was averaged over each forecast hour for each ensemble member, the probability matched means, and the Stage IV observations. Latitudinal averages of forecast and observed 3-h precipitation were then computed and plotted in time-longitude space (i.e Hovmoeller Diagrams) for each ensemble member and the means. A Hovmoeller diagram depicting the difference between model forecast and observed precipitation was also constructed for each ensemble member. Spatial correlation coefficients between the forecasts and observations were computed for each 24 h forecast period (hours 12-36 and hours 36-60) for each ensemble member in order to quantify how well the model forecast precipitation corresponded to the observed diurnal cycle. This method is similar to that used in Clark et al. (2007, 2009).

iii) Results

a) ETSs

Figures 2a-d depict the ETSs from the 0.10 to the 0.75 in. threshold for each of the ensemble members as a function of forecast hour. Generally, ETSs were fairly low, with values above 0.2 only existing at the lower thresholds and at forecast hours 3-12, mainly in the SSEF. However, these low values are consistent with the results from previous work that focused on the day 1 period (e.g. Clark et al. 2007). The SSEF outperformed the SREF during the day 1 period up to hour 36, with differences in ETS around 0.05. ETS in SSEF members had a broad maximum near 1200 UTC (forecast hours 12 and 36) and a broad minimum around 0000 UTC (forecast hour 24). There was a pronounced diurnal cycle in the SSEF member ETSs, especially for thresholds ≥ 0.50 in., likely associated with morning mesoscale convective system (MCS) activity leading to the peaks in ETSs. These peaks were likely due to the SSEF system members being able to explicitly depict large organized convective systems and their associated precipitation. This diurnal cycle in the ETS was not as pronounced in the SREF ensemble, likely due to its inability to depict these types of convective systems.

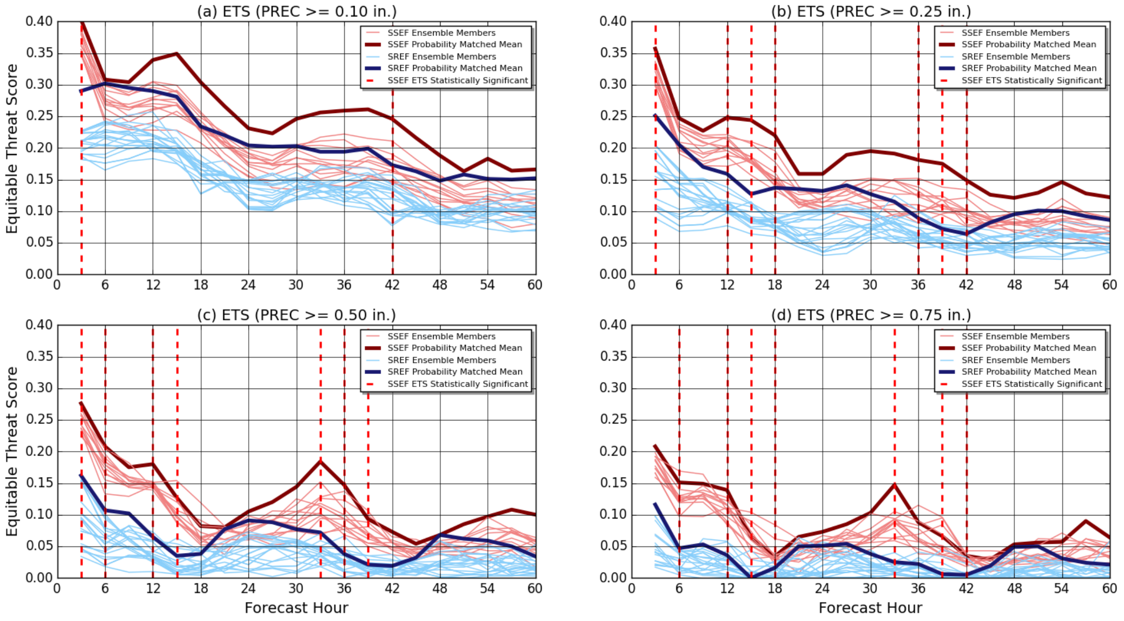


Figure 2 (a) 3 h ETS at each forecast hour for the 0.10 in. precipitation threshold. Hours with significant differences between ensemble means are indicated by red, dashed, vertical lines. (b), (c), and (d) are same as (a) except for 0.25, 0.50, 0.75 in. thresholds, respectively.

The SSEF continued to have ETSs of 0.02 to 0.05 points higher than the SREF in the day 2 forecast period from hours 36-60, with a pronounced diurnal cycle. The difference in ETS between the SSEF and SREF ensembles was greater in the 36-42 h forecast period when compared to the later periods. But, a definite benefit still exists all the way out to forecast hour 60, as the mean ETS from the SSEF is always higher than the mean ETS from the SREF.

Significant differences between the ETSs of the probability matched means (indicated by red, dashed, vertical lines in Figure 2) were more frequent at higher thresholds and generally occurred during the first 18 h of the forecasts and between hours 33 and 42. It is possible that a larger sample size or the consideration of larger spatial scales (e.g. Clark et al. 2010) would result in more times with significant differences.

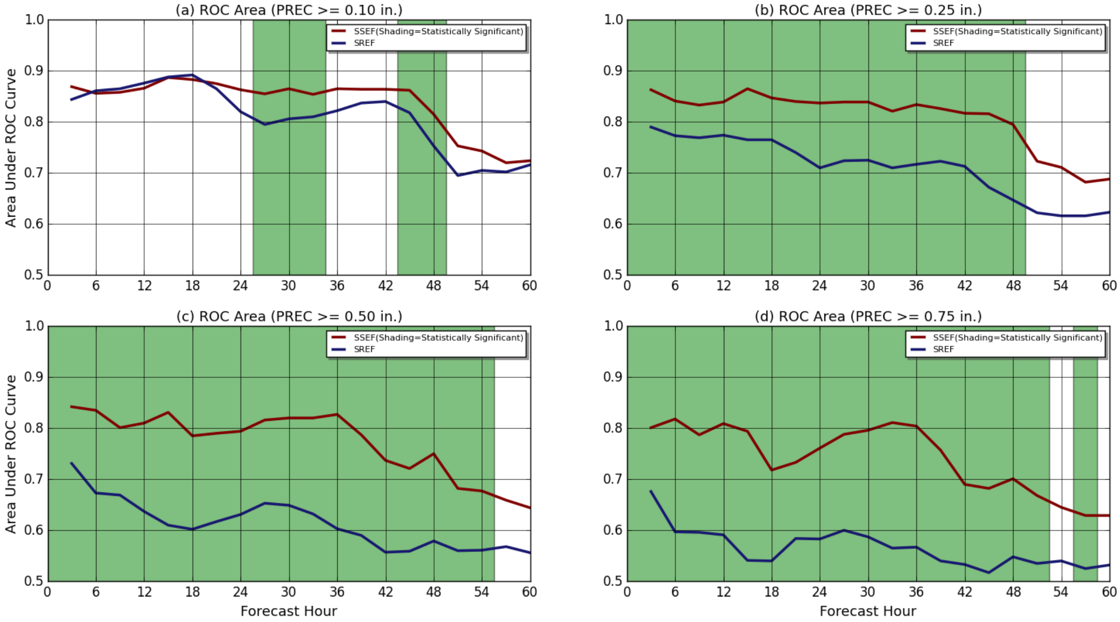


Figure 3 (a) 3 h ROC Area at each forecast hour for the 0.10 in. precipitation threshold. Hours with significant differences between ensemble PQPFs are indicated by green shading. (b), (c), and (d) same as (a) except for 0.25, 0.50, 0.75 in. thresholds respectively.

b) Area under the ROC curve

Figures 3a-d depict the area under the ROC curve for all 60 forecast hours for the 0.10 in., 0.25 in., 0.50 in., and 0.75 in. thresholds. The green shading represents statistically significant hours in favor of the SSEF. At the 0.10 in. threshold, both the SSEF and SREF have a similar amount of skill up to forecast hour 24. After that, the SSEF has a slightly greater ROC Area, even in the day 2 period, with the SSEF ROC Area being significantly higher at hours 42-48. At thresholds of 0.25 in. or greater, the SSEF outperforms the SREF by a much wider margin. The SSEF has significantly higher ROC Area values up to hour 48 for thresholds of ≥ 0.25 in. At the 0.25 in. threshold, both ensembles have positive skill, but the SSEF consistently has a ROC area of about 0.1 greater than the SREF, including during the entire day 2 period, although hours 48-60 are not significant. The gap widens further although the forecast skill starts to decrease at the 0.50 in. threshold. The SREF has almost no skill in the day 2 period, while the SSEF has some positive skill all the way out to forecast hour 60. During the day 2 period at the 0.75 in. threshold, the SREF has essentially no skill, while the SSEF still has a distinct positive amount of forecast skill, although the ROC Area values are lower than values at the lower thresholds. Similar to the ETS scores, ROC area values for all four thresholds show a distinct advantage for the SSEF during the day 2 period. Unlike the ETS results, there was not a more pronounced difference in ROC Area values for forecast hours 36-42 versus the remainder of the day 2 period. For both ensembles, the ROC Area values decreased sharply after hour 48, with only a fraction of those hours being statistically significant for the SSEF. When compared to the ETS results, there were many more hours with significant differences in favor of the SSEF using ROC Area. The lower differences for ROC Area likely occur because ETS evaluates single deterministic forecasts from the individual members or the ensemble mean and requires grid-point matches. Thus, if the forecast is wrong at the grid-point it is penalized. In contrast, ROC Area evaluates probabilistic forecasts that incorporate information from all ensemble members. Thus, if only 1 or 2 members of the ensemble have a correct forecast at a grid-point, the forecast is only partially penalized and receives some credit for being correct. Thus, because of the inherent uncertainty associated with longer range, high-resolution precipitation forecasts, it is easier for a superior ensemble system to receive more credit using probabilistic measures that account for forecast uncertainty.

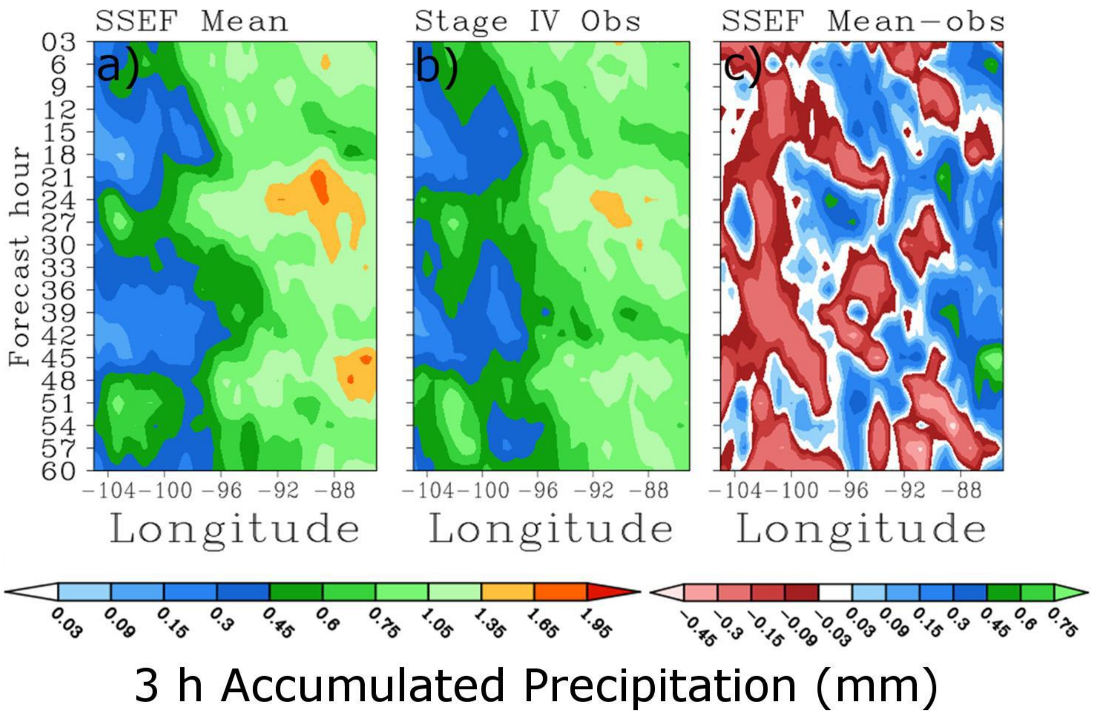


Figure 4 Latitudinal average Hovmoeller diagram of 3 h accumulated precipitation of the (a) SSEF probability matched ensemble mean, (b) Stage IV observed 3 h accumulated precipitation, and (c) difference between forecast and observed precipitation.

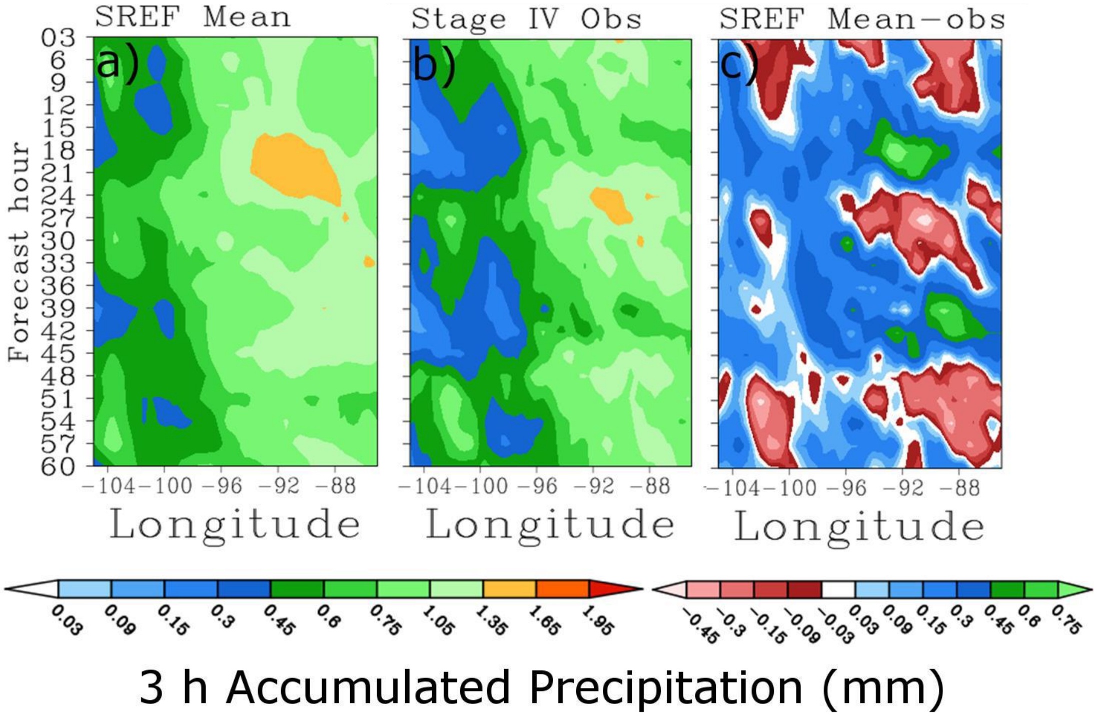


Figure 5 Latitudinal average Hovmoeller Diagram of 3 h accumulated precipitation of the (a) SREF probability matched ensemble mean, (b) Stage IV observed 3 h accumulated precipitation, and (c) difference between forecast and observed precipitation.

c) Hovmoeller Diagrams and Related Metrics

Hovmoeller diagrams were created for each SSEF and SREF ensemble member, but only the probability matched means and selected members are displayed in Figures 4-5. The SSEF ensemble mean is a representative depiction of the latitudinal average forecast precipitation field of all SSEF members, as variation between members was small (not shown). There was more member to member variability in the SREF ensemble, but Figure 5 is fairly representative of what the Hovmoeller diagrams of most of the SREF ensemble members looked like, with the exception of some NMM and NMMB members. These exceptions will be addressed later in this paper. In general, the SSEF better represented observed precipitation, although there was slight over-forecasting evident in the eastern areas. In the day 2 period, a diurnal cycle was clearly evident in the SSEF, with both forecast and observed precipitation maxima occurring around forecast hour 48. In the SREF, it can be seen that much of the precipitation is smoothed out, especially in the day 2 period. Furthermore, there is a 3 h phase lag relative to observations evident in the SREF mean Hovmoeller. This phase lag was observed with most of the SREF ensemble members, but not all of them. No phase lag was observed in any of the SSEF members.

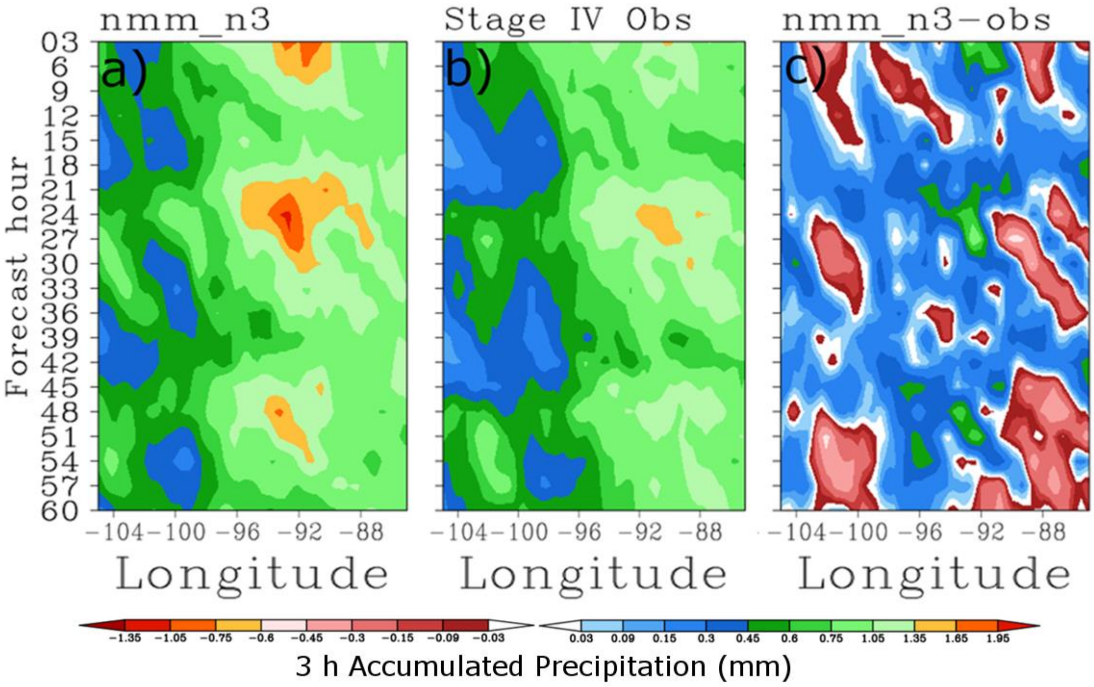


Figure 6 Latitudinal average Hovmoeller Diagram of 3 h accumulated precipitation of the (a) nmm_n3 SREF member, (b) Stage IV observed 3 h accumulated precipitation, and (c) difference between forecast and observed precipitation.

As noted earlier in this section, there were a few SREF members that performed noticeably better than the SSEF ensemble mean in both the day 1 and day 2 periods. In total, 4 members came from the NMM model and 2 were from the NMMB model. Figure 6 is a Hovmoeller Diagram of the NMM_n3, one of the 6 better performing SREF members. As can be seen, there is no phase lag with the NMM_n3 SREF member, and a coherent diurnal cycle is evident even in the day 2 period. The other 5 better performing SREF members had similar looking Hovmoeller diagrams, with well-defined diurnal cycles (not shown). These 6 better performing SREF members can be seen by looking at the results of the computed spatial correlation coefficients for the day 2 period, displayed in Figure 7, along with the other SSEF/SREF members. The SSEF members had much higher values than most of the SREF members, but the 6 SREF members indicated in the plot had slightly higher coefficients than the SSEF members. In addition, the better performing SREF members were able to accurately depict precipitation amounts across the area of analysis without a phase lag (not shown). Four of the 6 better performing SREF members used the SAS convective parameterization scheme, but it is still not certain exactly why these 6 SREF members performed better.

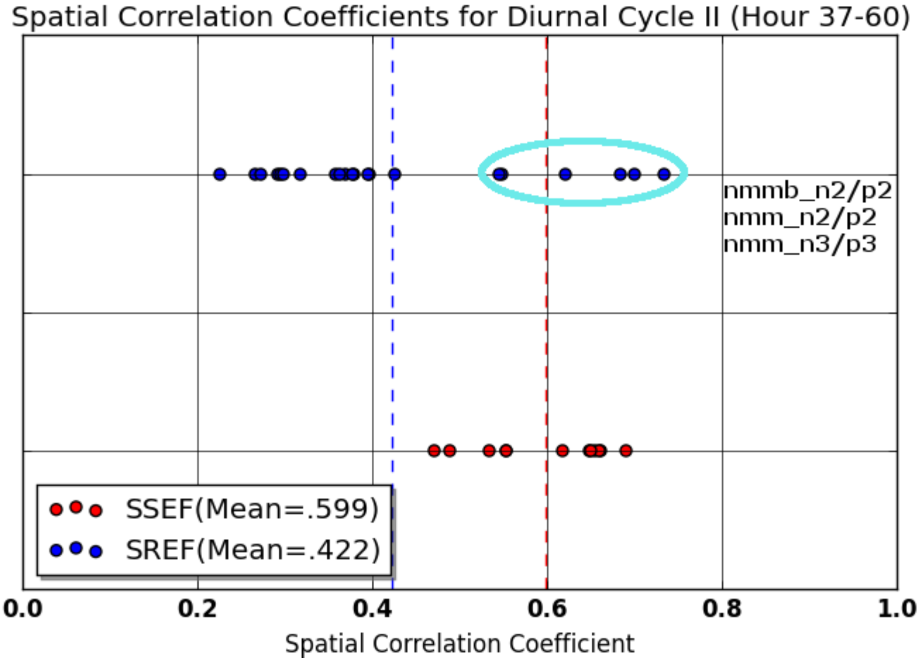


Figure 7 Spatial correlation coefficients computed for each SSEF and SREF ensemble member for the day 2 forecast period. Red dots denote individual SSEF members, while blue dots denote individual SREF members. The red and blue dashed lines denote the mean spatial correlation coefficients of the SSEF and SREF ensembles, respectively. The 6 blue dots that are circled represent the 6 better performing SREF members, which are listed in the plot.

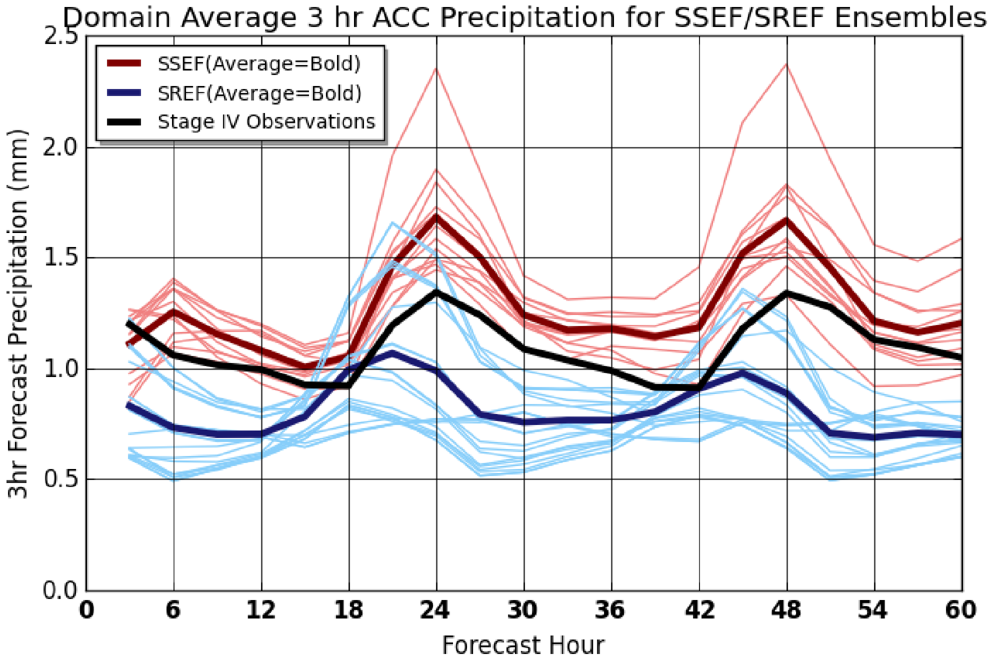


Figure 8 Graph of domain averaged 3 hour forecast precipitation from the SSEF members, SREF members, and Stage IV observations.

Domain averaged precipitation amounts are shown in Figure 8 for all of the SSEF and SREF members in red and blue, respectively, along with observations in black. The phase lag can easily be seen, as the maxima in the SREF forecast precipitation are at hour 21 and hour 45, while the maxima in the SSEF forecasts and Stage IV observations are both at hour 24 and hour 48. In addition, most of the SREF members greatly under forecasted the precipitation, in the both the day 1 and day 2 periods. Although the SSEF over-forecast throughout the 60 hours, the SSEF forecast and observed precipitation match up fairly well, even in the day 2 period.

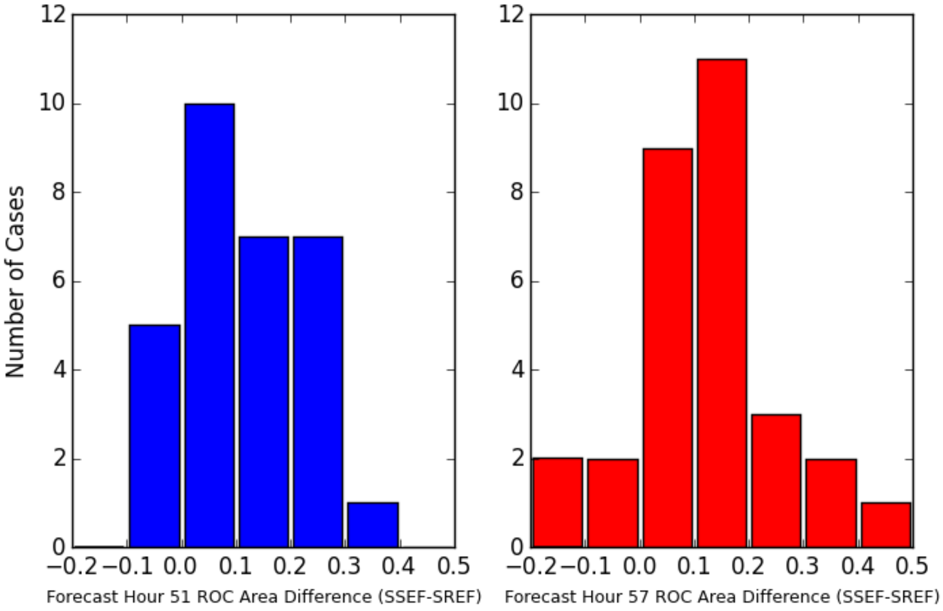


Figure 9 Differences in ROC Area for all of the cases for (a) forecast hour 51 and (b) forecast hour 57

d) Individual Cases

ROC Area for both ensembles was computed for each of the 30 cases to examine the distribution of ROC Area differences between the SSEF and SREF during the day 2 forecast period. Figure 9 shows this distribution of ROC Area Differences (SSEF-SREF) for forecast hour 51 and 57. For both forecast hours, the median difference is around 0.1, but during forecast hour 57 there is a larger spread, with differences ranging from -.164 to .498. For both forecast hours, in about 15% of the cases, the SREF performed slightly better in the day 2 period. In a few instances, especially when large convective systems were present, the SSEF far outperformed the SREF during the day 2 period. Selected forecasts and observations from two of these cases are displayed in Figure 10. Figure 10a and b show the 51 h forecast PQPF valid at 0300 UTC on 29 April for the 0.50 in. threshold (color shading), ROC Area, and 3 h observed precipitation greater than 0.50 in. (overlaid in red stippling on PQPF forecast maps). Figure 10c and d display the same parameters as Figure 10 a-b except it displays the 57 h forecast valid at 0900 UTC on 4 June.

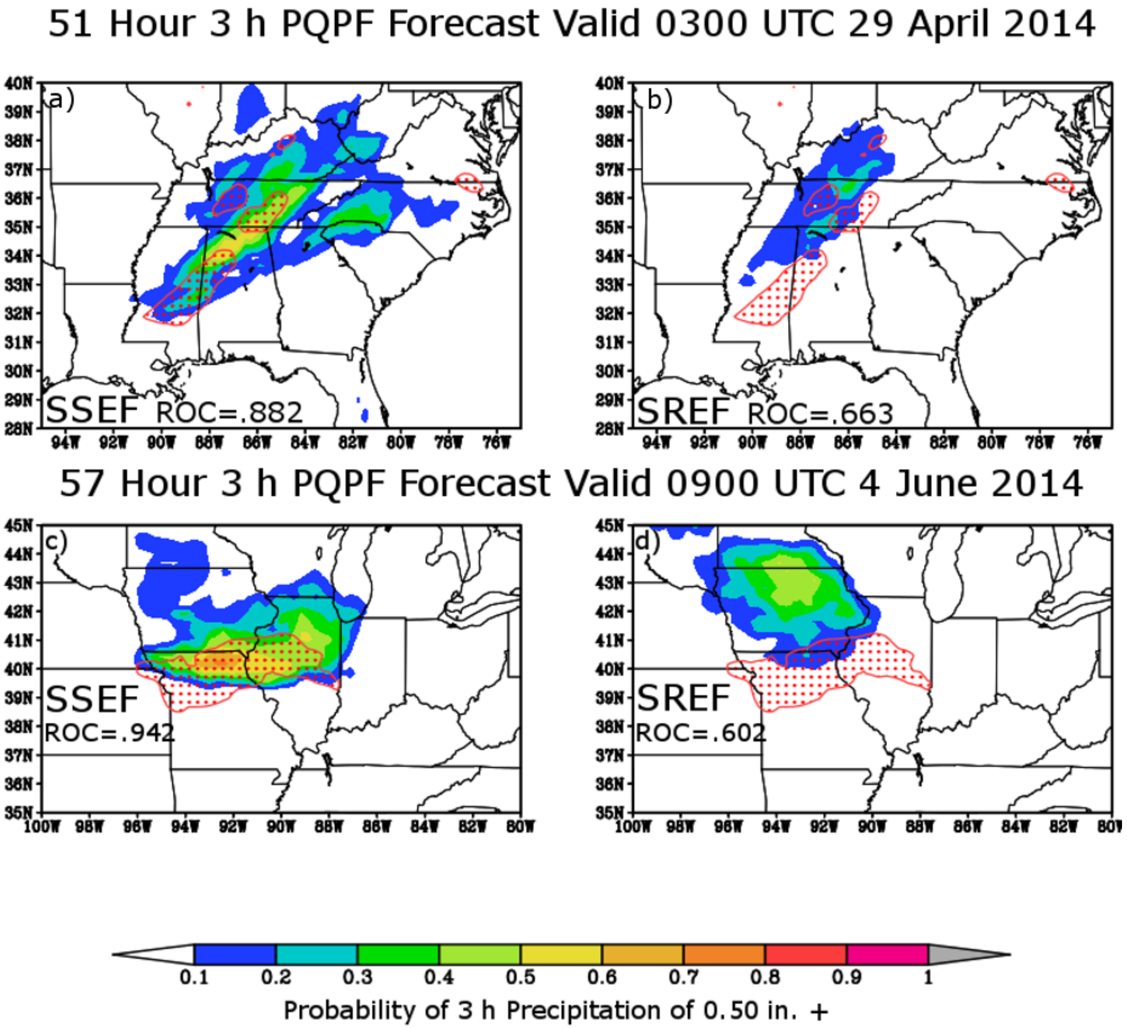


Figure 10 3 h PQPF forecasts for the 0.50 in. threshold valid from 0000-0300 UTC 29 April and 0600-0900 UTC 4 June in color overlaid with 3 h observed precipitation of ≥ 0.50 in. in red stippling for the (a) SSEF on 29 April, (b) SREF on 29 April, (c) SSEF on 4 June, and (d) SREF on 4 June.

There were a few differences in synoptic setup and severe weather frequencies between the two cases. The 28-29 April case featured a vertically stacked closed low from the surface to around 300 hPa over southeast Nebraska and 50-80 knot 500 hPa winds over the region of interest displayed in Figure 10a-b. At the surface, temperatures were only in the upper 70s and lower 80s (degrees F) across Alabama and Mississippi, with dew points in the mid to upper 60s (degrees F). A weak cold front was advancing from west to east across the southeastern United States. Over 100 tornadoes were reported during this severe weather outbreak across the southeast, although none were rated higher than EF2 (not shown). The SREF failed to depict where the heaviest precipitation would fall during the selected 3 h period on 29 April. None of the area with observed 3 h precipitation ≥ 0.50 in. had PQPFs greater than 0.2 for the 0.50 in. threshold. In addition, the SREF did not pick up on the southern part of the observed line of thunderstorms, having zero probabilities for a large area of Alabama and Mississippi at 0300 UTC on 29 April. On the other hand, the SSEF was able to predict the location of the main line of thunderstorms extending from eastern Mississippi to central Tennessee. Even the southern end of observed 3 h precipitation ≥ 0.50 in. corresponded fairly well to the southern end of the SSEF PQPFs for the forecast initialized 51 h before the event.

In contrast to the 29 April case, there were no well-defined upper level features during the June 3-4 case across the Plains and Midwestern United States. Zonal flow was observed in the upper levels, but there were 40-50 knot winds at 500 hPa over Nebraska, Iowa, and northern Missouri. Unlike the April 28-29 outbreak, a large temperature gradient existed across the area as a front extended from west to east across Nebraska and

Iowa. South of the front, temperatures rose to around 90 on 3 June and dew points were around 70 degrees F, while north of the front, it was only in the 60s. Thunderstorms developed during the day in Nebraska and Iowa on 3 June, then moved into northern Missouri during the early morning hours on 4 June. Again, the SREF did not accurately predict where the bulk of the precipitation fell. It did not pick up on the area of northern Missouri and central Illinois where 3 h precipitation amounts were ≥ 0.50 in. In addition, the SREF also had a large area of false alarm in Iowa where PQPFs of 0.4 to 0.5 were forecasted, but no 3 h precipitation observations of ≥ 0.50 in. were observed. Conversely, the SSEF was able to predict the spatial extent of precipitation associated with an MCS and had less false alarm in Iowa, as forecast probabilities were only in the 0.1 to 0.2 range. In the area where precipitation ≥ 0.50 in. was observed in the June case, the SSEF had probabilities of greater than 0.6 that verified. The SSEF was also able to accurately depict the eastern extension of the MCS into Illinois where the SREF had zero probability of 3 h precipitation ≥ 0.50 in.

In both cases, the SREF failed to depict where the heaviest precipitation would fall during the selected 3 h period as none of the area with observed 3 h precipitation ≥ 0.50 in. had PQPFs greater than 0.2 for the 0.50 in. threshold. In addition, the SREF did not pick up on the southern part of the observed line of thunderstorms in both the April and June cases, having zero probabilities for a large areas where precipitation amounts of ≥ 0.50 in. were observed. The ROC area differences of 0.20 to 0.35 during these cases along with the forecasts and observations clearly show that the CAM SSEF ensemble had a definite forecast advantage in the day 2 period, with the advantage most recognizable when large scale convection was present.

iv) Summary and discussion

Three h QPFs during the day 2 forecast period from the 32-km grid-spacing SREF ensemble were compared to QPFs from the 4-km grid-spacing CAM SSEF ensemble. The forecasts were initialized at 2100 and 0000 UTC on 30 days during the 2014 NOAA/HWT Spring Forecasting Experiment from 24 April to 3 June. Some of the SREF members provide lateral boundary conditions for the SSEF. The goal of this study was to see whether the SSEF CAM ensemble QPFs outperform those from the convection-parameterizing SREF ensemble during the day 2 (36-60 h) forecast period, as previous work has found a distinct advantage in QPFs in CAMs relative to models that parameterize convection during the day 1 (12-36 h) period. The analysis was done by computing ETSs for both ensembles, using spatial correlation coefficients and Hovmoeller diagrams to examine the diurnal precipitation cycle, evaluating PQPFs by computing ROC Area, implementing hypothesis testing on ETS and ROC Area, and examining two specific cases. Results are summarized below.

ETSs computed for the 0.10 in. to 0.75 in. thresholds were fairly low (≤ 0.2), but the ETSs from the SSEF were consistently 0.02 to 0.05 points higher during the day 2 period, with the advantage peaking during forecast hours 36-42, likely associated with morning MCS activity. Hypothesis testing supported this, as significant differences were observed during the 36-42 h forecast period at many of the thresholds. ROC Area showed an even more pronounced advantage in the SSEF, especially at the higher thresholds. While the ROC Area of the SSEF was only a few hundredths of a point higher at the 0.10 in. threshold, the average difference in ROC Area was about 0.10-0.15 points higher at the higher thresholds. In addition, forecast hours 36-48 had significant differences at all thresholds ≥ 0.25 in. Hovmoeller diagrams for each ensemble member of latitudinally averaged QPFs compared with a diagram of observed precipitation (in time-longitude space) showed that the SSEF over-forecasted precipitation, but it modeled the diurnal cycle well, as evidenced by much higher mean spatial correlation coefficients than the SREF. Most of the SREF members had a 3 h phase lag in QPF. However, there were 6 SREF members that modeled the diurnal cycle well and had higher spatial correlation coefficients than the SSEF members, which was an unexpected result. Four of these 6 better performing SREF members used the SAS convective parameterization scheme. However, there would need to be more investigation to determine why these 6 SREF members performed significantly better than the other SREF members. Day 2 PQPFs at the 0.50 in. threshold from two severe weather cases (28-29 April and 3-4 June) were examined in order to see how the SSEF and SREF forecasts would appear to an operational forecaster during a high impact event. The SSEF greatly outperformed the SREF during the two selected cases, with ROC Areas of 0.2-0.35 points higher for the selected forecast hours during these events.

After examining all of the results, it is clear that objective evaluation metrics, like the ETS, are in favor of the SSEF into the day 2 period. At all five thresholds examined, the difference in ETS is more pronounced in the early morning timeframe (when some hours have statistically significant differences), when larger scale convective systems tend to play more of a role in the forecast. The SSEF has the ability to explicitly depict these features, while the SREF cannot. Even out to hour 60, the ETS difference in favor of the SSEF ranges from about 0.02-0.05, with the larger differences at the higher thresholds. At thresholds higher than 0.50 in., the SREF showed essentially no skill in the day 2 period. Also, the ETS scores of the SSEF gradually decreased as the threshold increased. However, there was still some amount of skill even at the 1.00 in. threshold.

For ROC Area, similar to ETSs, the advantage of the SSEF was more pronounced as the precipitation threshold increased. This advantage was pronounced during the day 2 period as well, where the SREF showed no skill in terms of ROC Area at thresholds greater than 0.50 in. The main difference in the ROC Area results versus the ETS was that there was no pronounced maximum in performance during the early morning hours. Instead, there was a consistent difference between the SSEF and SREF during the day 2 period, with virtually all of the hours having statistically significant differences up to forecast hour 48 (at thresholds ≥ 0.25 in.).

In general, the SSEF depicts a more coherent diurnal cycle than the SREF over the domain of analysis. The SREF ensemble mean does not represent the diurnal cycle well, mainly due to its inability to explicitly depict convection, especially in the day 2 period. In addition, the SREF ensemble mean significantly under forecasted precipitation in the eastern part of the domain. The only result in favor of the SREF was those 6 NMM/NMMB members that slightly outperformed the SSEF in terms of spatial correlation coefficients. This was an unexpected result.

Future work should continue to examine the forecast range at which convection-allowing guidance demonstrates increased PQPF skill relative to coarser modeling systems. The NOAA/HWT Spring Forecasting Experiments serve as an ideal place for these tests, since both researchers and operational forecasters can be involved in the evaluations. The preliminary results from this study suggest that there would be at least noticeable benefits to extending an operational convection-allowing ensemble to at least

60 h.

Chapter 2: A Comparison of Precipitation Forecast Skill in Convection-Allowing Model Ensembles with and without Radar Data Assimilation

i) Introduction

Warm season quantitative precipitation forecasts (QPFs) have been studied extensively over the past few years as the number of convection-allowing models (CAMs) available to forecasters has grown. CAMs are models with a resolution fine enough to explicitly depict convection. It is widely acknowledged that a maximum of 4-km grid spacing is required for NWP models to explicitly depict convection (Weisman et al. 1997).

One thing that has been used to improve QPFs and forecasts for other parameters in CAMs is assimilating radial velocity and reflectivity data from the WSR-88D radar network as well as other high resolution observations [hereafter referred to as Radar Data Assimilation (Radar DA)] into the initial conditions (ICs) of a CAM. Some CAMs use Radar DA while others simply derive the ICs via a larger scale model such as the North American Mesoscale (NAM; Rogers et al. 2009) model or Global Forecast System (GFS), which is commonly referred to as a "cold start" initialization. Recent studies dating back to the mid-2000s have mixed results regarding the positive impact of Radar DA. Some studies indicate a benefit mainly in the first 3 to 6 hours (e.g. Xiao et al. 2007), with some positive impact all the way out to 12 hours (e.g. Zhu et al. 2015), and others show minimal benefits of Radar DA, even in very short term QPFs (e.g. Li et al. 2012). Other recent work has focused on how Radar DA has improved the prediction of location and intensity of convection in the United States. Part of Xiao et al. (2007b) examined Advanced Research Weather Research and Forecasting Model (ARW; Skamarock et al. 2008) 3-h QPFs for a squall line in Kansas and Oklahoma and found that adding Radar DA to the NWP model had a positive impact for forecasts with 6 and 9 h lead times for both light and heavy precipitation thresholds. Additionally, Zhao et al. (2008) noted that assimilating radar reflectivity improved forecasts of thunderstorm location and intensity for forecasts with 1 to 4-h lead times (the study did not look at longer lead times).

Some other recent work has focused on how CAM guidance using Radar DA is perceived versus CAM guidance with "cold start" forecasts using a variety of verification metrics. For example, during the 2008, 2009, and 2010 NOAA/Hazardous Weather Testbed (HWT) Spring Forecasting Experiments (e.g., Clark et al. 2012), QPF and forecast reflectivity products from a CAM version of the WRF model (Skamarock et al. 2008) that used Radar DA were compared to an identical WRF configuration that used a "cold start". SFE participants were asked to "define when the cold start forecasts appeared to catch up with the hot start[2] forecasts in terms of its degree of correspondence with reality" (Stratman et al. 2013). The consensus of the participants was that by forecast hour 3 and clearly by forecast hour 6, both forecasts had equal skill, although some objective metrics indicated that the version of the WRF that used Radar DA had at least some advantage out to forecast hour 12 (Kain et al. 2010). Examining the model configurations with and without radar DA from the 2009 and 2010 SFEs using spatial verification methods, Stratman et al. (2013) found that while both models have little skill at scales < 40 km, the run with radar DA had greater skill at larger scales up to 320 km.

While the studies that examined the subjective evaluations from the 2008-2010 NOAA/HWT SFEs examined deterministic comparisons of NWP models with and without Radar DA, there have not been any studies directly comparing two identical ensembles with and without Radar DA. During the 2016 NOAA/HWT SFE, the National

[2] A "hot start" is commonly used to refer to a model initialized with clouds and hydrometeors in the ICs.

Severe Storms Laboratory (NSSL) ran a 10 member CAM ensemble with a cold start that used the 12-km NAM Analysis for ICs [hereafter referred to as “norad-ens”] while the Center of Analysis and Prediction of Storms (CAPS) ran an identical 10 member CAM ensemble except that it had the addition of Radar DA into the ICs [hereafter referred to as “rad-ens”]. Both the norad-ens and rad-ens provided forecasts out to 36-h, and subjective evaluations were completed by SFE participants answering the question: “How long into the forecast does the assimilation of radar data have a positive impact on the [rad-ens] ensemble forecast [compared to the norad-ens]?” This provided an opportunity to examine identical ensembles with and without Radar DA, rather than just looking at deterministic forecasts, which has been done several times in the past. The three main ways that the ensemble forecasts will be compared are to 1) to see how far into the 36 h forecast period the positive effects of radar data assimilation extend, 2) examine the spread of each of the two ensembles, and 3) compare the subjective evaluations of the probabilistic reflectivity forecasts completed by the 2016 SFE participants to the objective evaluation metrics based on the reflectivity forecasts from the two ensembles. This study is somewhat similar to Kain et al. (2010) and Stratman et al. (2013) except for the fact that it focuses on ensemble rather than deterministic forecasts.

The remainder of the study is organized as follows: Section 2 presents information on datasets and methods, Section 3 presents results along with graphs of these metrics, and Section 4 provides a summary and conclusions.

ii) Data and methodology

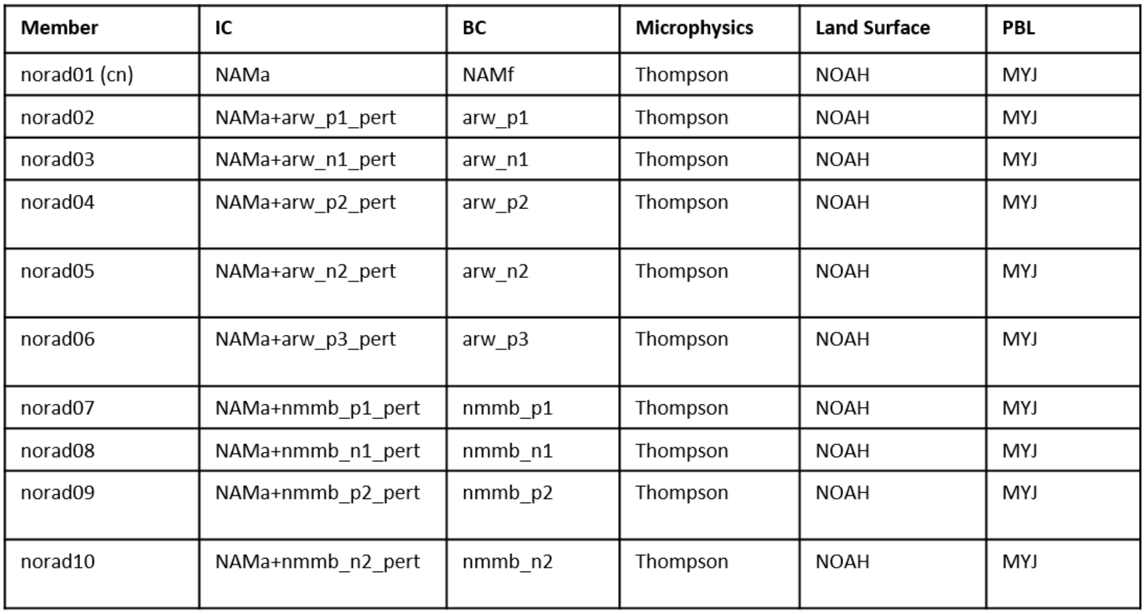
a) NSSL and CAPS ensemble description

Forecast precipitation for 3-h periods was examined from the rad-ens and norad-ens, which both had 3-km grid spacing configurations. The rad-ens and norad-ens were provided to support the 2016 NOAA/HWT SFE and consisted of model integrations conducted from 2 May through 3 June. The 22 days of model integrations that were used for this study are as follows: 2-5 May, 9-13 May, 16-20 May, 23-25 May, 27 May, 30-31 May, and 2-3 June. Both ensembles were initialized at 0000 UTC (Clark et al. 2016). Because NSSL and CAPS did not run the ensemble members during most weekend days in May and June (and there were technical issues that prevented several of the ensemble members from running during some weekdays), 22 days during the 2016 SFE period had full or nearly full datasets available from both ensembles. Observed precipitation data was derived from the NCEP Stage IV dataset (Baldwin and Mitchell 1997), which was on a 4-km grid.

The norad-ens was generated using the Advanced Research WRF (ARW) model (Skamarock et al. 2008) run by NSSL and CAPS for the 2016 NOAA/HWT SFE (Clark et al. 2016). During the 2016 SFE, the norad-ens had 10 members with 3-km grid spacing that were initialized on weekdays at 0000 UTC and integrated 36 h over a CONUS domain from the beginning of May to the beginning of June. ICs and boundary conditions (BCs) for the control member were taken from the NAM analyses and forecasts, respectively (NAMa and NAMf in Table 1). IC perturbations were derived from evolved (through 3 h) perturbations of 2100 UTC initialized members of the SREF system and added to the control member ICs. For each perturbed member, the forecast of the SREF member used for the IC perturbations was also used for the BCs. Note that the letters "p" and "n" in Table 1 signify whether each perturbation is positive or negative, respectively. 4 SREF members from the Advanced Research WRF (ARW; Skamarock et al. 2008) and

5 members from the Non-hydrostatic Multiscale Model on the B grid (NMMB) (Janjić 2005, 2010; Janjić and Black 2007; Janjić et al. 2011; Janjić and Gall 2012) provided ICs and BCs for the norad-ens members. All members used the Noah land surface model (Ek et al. 2003), Mellor-Yamada-Janjic planetary boundary layer scheme (MYJ; Mellor and Yamada 1982; Janjic 1990, 2002), and Thompson microphysics scheme (Thompson et al. 2004). Table 1 shows detailed specifications of all 10 norad-ens members.

Table 1 Model specifications for all 10 NSSL members. Abbreviations in the table are described in the text.

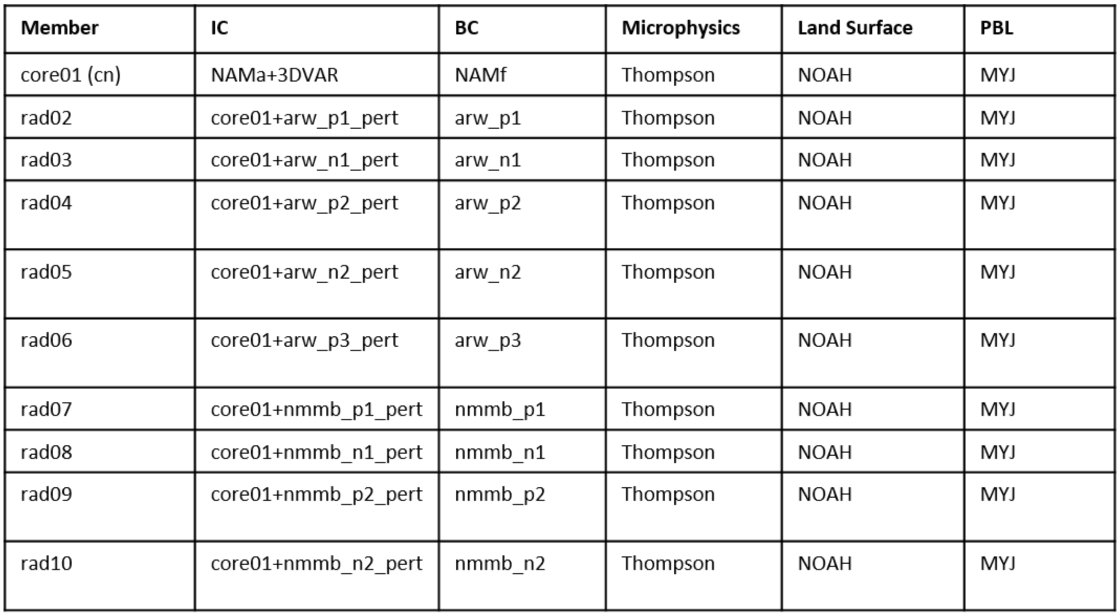


Member	IC	BC	Microphysics	Land Surface	PBL
norad01 (cn)	NAMa	NAMf	Thompson	NOAH	MYJ
norad02	NAMa+arw_p1_pert	arw_p1	Thompson	NOAH	MYJ
norad03	NAMa+arw_n1_pert	arw_n1	Thompson	NOAH	MYJ
norad04	NAMa+arw_p2_pert	arw_p2	Thompson	NOAH	MYJ
norad05	NAMa+arw_n2_pert	arw_n2	Thompson	NOAH	MYJ
norad06	NAMa+arw_p3_pert	arw_p3	Thompson	NOAH	MYJ
norad07	NAMa+nmmb_p1_pert	nmmb_p1	Thompson	NOAH	MYJ
norad08	NAMa+nmmb_n1_pert	nmmb_n1	Thompson	NOAH	MYJ
norad09	NAMa+nmmb_p2_pert	nmmb_p2	Thompson	NOAH	MYJ
norad10	NAMa+nmmb_n2_pert	nmmb_n2	Thompson	NOAH	MYJ

The rad-ens also had 10 members with a 3-km grid spacing that were initialized on weekdays at 0000 UTC and integrated 36 h over the same CONUS domain for the same dates as the NSSL ensemble. All of the ICs, BCs, and physics of the CAPS are the same as the NSSL, except the CAPS ensemble used Radar DA. Radial velocity and reflectivity data from Weather Surveillance Radar-1988 Doppler (WSR-88D) and other high-resolution observations were assimilated into the ICs using the ARPS three-dimensional variational data assimilation (3DVAR; Xue et al. 2003; Gao et al. 2004) data and cloud analysis system (Xue et al. 2003; Hu et al. 2006; Gao et al. 2008). Table 2 shows detailed specifications of all 10 CAPS members.

Before any verification metrics were computed, the rad-ens, norad-ens, and 3-h observed NCEP Stage IV precipitation data were interpolated to 4-km grid with dimensions of 1199 by 799 grid points using a neighbor budget interpolation (e.g. Accadia et al. 2003). A mask was used to consider only points within all model domains, east of the Rocky Mountains over land, and only in the United States. This is due to the relative lack of reliable WSR-88D radar observations over the mountains, water, and outside the United States. The area of analysis is displayed in Figure 1.

Table 2 Configurations for all 10 CAPS members. The abbreviations in the table are described in the text.



Member	IC	BC	Microphysics	Land Surface	PBL
core01 (cn)	NAMa+3DVAR	NAMf	Thompson	NOAH	MYJ
rad02	core01+arw_p1_pert	arw_p1	Thompson	NOAH	MYJ
rad03	core01+arw_n1_pert	arw_n1	Thompson	NOAH	MYJ
rad04	core01+arw_p2_pert	arw_p2	Thompson	NOAH	MYJ
rad05	core01+arw_n2_pert	arw_n2	Thompson	NOAH	MYJ
rad06	core01+arw_p3_pert	arw_p3	Thompson	NOAH	MYJ
rad07	core01+nmmb_p1_pert	nmmb_p1	Thompson	NOAH	MYJ
rad08	core01+nmmb_n1_pert	nmmb_n1	Thompson	NOAH	MYJ
rad09	core01+nmmb_p2_pert	nmmb_p2	Thompson	NOAH	MYJ
rad10	core01+nmmb_n2_pert	nmmb_n2	Thompson	NOAH	MYJ

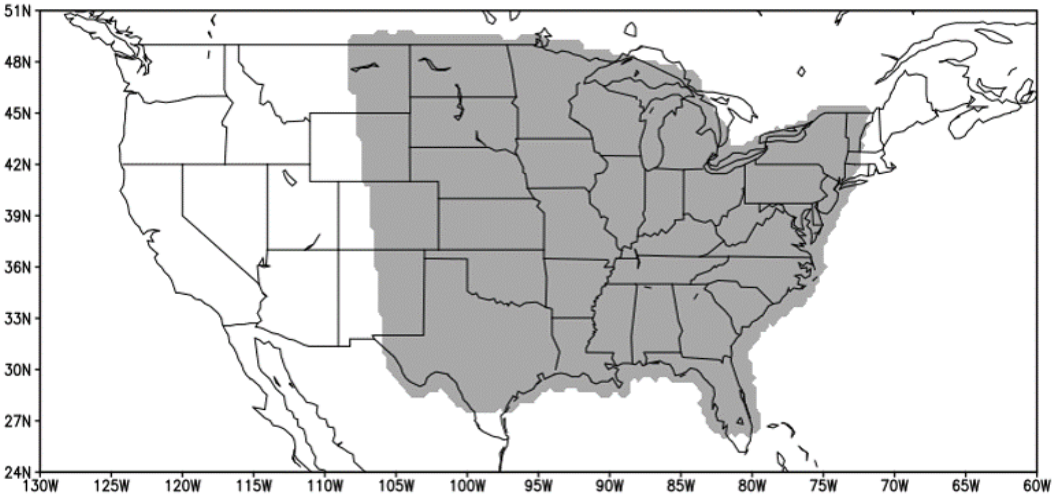


Figure 1 Analysis domain (in gray) for both of the ensembles.

b) Forecast Evaluation Metrics

The first metric that was used to evaluate the precipitation forecasts from each ensemble was the Equitable Threat Score (ETS; Schaefer 1990). ETS measures the fraction of observed and/or forecast events that were correctly predicted, adjusted for hits associated with random chance. The ETS was calculated using contingency table elements computed from every grid point in the 4-km grid spacing analysis domain for each ensemble member every 3 hours as follows: ETS = $(H - H_{cha})/(H + FA + M + H_{cha})$, where H represents a hit (model correctly forecasted precipitation to exceed a certain threshold), H_{cha} represents the number of hits expected by chance, FA represents the number of false alarms (model forecasted precipitation to exceed a certain threshold, but the observed precipitation did not exceed that threshold), and M represents a miss (model did not forecast precipitation to exceed a certain threshold, but the observed precipitation did exceed that threshold). An ETS score of 1 is perfect, and a score below zero represents no forecast skill.

In addition to computing ETS for individual ensemble members, ETS was also computed for ensemble mean precipitation forecasts. Ensemble means were computed using the probability matching technique (Ebert 2001). This technique assumes that the best spatial representation of the precipitation field is given by the ensemble mean, and that the best probability density function (PDF) of rain rates is given by the ensemble member QPFs for all *n* ensemble members. To compute the probability matched mean, the precipitation forecasts from the ensemble members for every grid point are ranked in order from largest to smallest, keeping every *n*th value. The precipitation forecasts from the ensemble mean forecast are similarly ranked from largest to smallest, keeping every value. Then, the grid point with the highest value in the ensemble mean QPF field is reassigned to the highest QPF value in the ensemble member QPF distribution. Then the grid point with the second highest value in the ensemble mean QPF field is reassigned to the second highest value in the ensemble member QPF distribution. This process is then repeated for all of the rankings, ending with the lowest. ETS scores were then calculated from these probability matched means for both ensembles for forecast hours 3-36 in 3-h intervals.

In addition to computing ETS, neighborhood-based ETS (ETS_r; Clark et al. 2010) was computed for all NSSL/CAPS ensemble members as well as ensemble means. The formula to calculate ETS_r is the same formula used for ETS, except that the criteria for a "hit" (and consequently a "miss" and "false alarm") is adjusted to include any number of grid points within a specified radius, r. That is, if an event is forecast at a specific grid point, and that event is observed in at least one grid point within a radius r of that specific grid point, then the forecast is considered a "hit." Thus, a "miss" is only assigned if an event is observed at a specified grid point and none of the grid points within the radius r forecast the event, while a "false alarm" is assigned only if an event is forecast at a specified grid point and not observed in any of the grid points within the radius r. For this study, radii r of 8, 30, and 60 km were used when calculating ETS_r.

Hypothesis testing was conducted to evaluate whether the rad-ens forecasts with 3DVAR were significantly more accurate than the norad-ens cold start forecasts. The hypothesis testing was conducted to compare the two ensemble means for ETS and ETS_r for r = 8, 30, and 60 km. In addition, the hypothesis testing was carried out to compare the ensemble control members. All 12 accumulation periods spanning 36 forecast hours were considered in each of the hypothesis tests conducted while using the resampling method of Hamill (1999). To apply this method, the test statistic used to look at the difference in accuracy of the 3-h precipitation forecast ending at hour hr (where hr is a given forecast hour) is ($ETS_{CAPShr} - ETS_{NSSLhr}$). The null hypothesis, H_o, is $ETS_{CAPShr} - ETS_{NSSLhr} = 0.00$. The alternative hypothesis, H_a, is $ETS_{CAPShr} - ETS_{NSSLhr} \neq 0.00$. The significance level used was $\alpha = 0.05$ and resampling was done 1000 times for each hypothesis test. More detailed information about the resampling method can be found in Hamill (1999).

In addition to the deterministic forecasts, probabilistic quantitative precipitation forecasts (PQPFs) for both of the ensembles were generated for all four precipitation thresholds for 3-h QPF. Probabilities were computed using the ratio of members that exceeded the specified threshold to the total number of members. The probabilistic forecasts were evaluated using the area under the Receiver Operating Characteristics Curve (ROC Area; Mason 1982), which measures the ability to discriminate between events (exceedances of specified threshold) and non-events (failure to exceed specified threshold). It is calculated by computing the area under a curve constructed by plotting the probability of detection (POD) versus the probability of false detection (POFD). The area under the curve is computed using the trapezoidal method (Wandishin et al. 2001) using the probabilities 0.05 to 0.95, in increments of 0.05. Values of ROC Area range from 0 to 1, with a value of 0.5 indicating no forecast skill and values above 0.5 indicating positive forecast skill. Similar to ETS, statistical significance was tested for the ROC areas using the resampling method (Hamill 1999).

In addition to the QPFs and PQPFs generated, 1-h probabilistic ensemble reflectivity forecasts (Probability of Reflectivity > 40 dBZ) were generated for both of the ensembles and evaluated subjectively and objectively throughout the 2016 NOAA/HWT SFE. Every weekday, (except for 30 May; Memorial Day) participants answered three questions about the previous day's NSSL and CAPS ensemble probabilistic reflectivity forecasts. The questions were: 1) Subjectively rate the [rad-ens] hourly ensemble reflectivity probabilities (>40 dBZ) during 00-23Z period, 2) Subjectively rate the [norad-ens] hourly ensemble reflectivity probabilities (>40 dBZ) during 00-23Z period, and 3) How long into the forecast does the assimilation of radar data have a positive impact on the ensemble forecast? All of these subjective evaluations were compiled into a database and results from them were compared to the ETS/ ROC Area evaluations from the rad-ens/norad-ens QPFs as well as Fractions Skill Score (FSS; Roberts and Lean 2008, Schwartz et al. 2010) from the probabilistic ensemble reflectivity forecasts. FSSs range from 0 to 1, with 0 indicating no skill and 1 indicating a perfect score. The method used to calculate FSS is described in Melick et al. (2012). Forecast reflectivity probabilities > 40 dBZ do not rely on fractional coverage of an event, but rather are determined by first obtaining the neighborhood maximum (with a radius of influence of 40-km) for the reflectivity at each grid point and determining if a 40-dBZ threshold has been surpassed (each ensemble member and observations). Then, the fraction of members (for each of the two ensembles) with one or more grid points meeting or exceeding the threshold (40 dBZ) within the radius of influence (40-km) is determined. A 2-D Gaussian kernel operator is then applied to smooth the ensemble probabilities and to create probabilities for the observations. The observed, hybridscan reflectivity comes from the operational Multiple Radar Multiple Sensor (MRMS) system.

The spread of both ensembles was examined by computing ensemble variance and by constructing rank histograms. The true ensemble mean, computed for all 22 cases, was simply computed by averaging the 3-h QPF value of all 10 ensemble members at each grid point for each forecast hour from 3 to 36. QPF values were then corrected for bias so that forecast precipitation amounts are reassigned using the corresponding distribution of observed precipitation amounts. This way, the bias corrected QPF values have the same spatial distribution as the original QPF values, but the rain rate distribution is the same as the rain rate distribution of the corresponding observed precipitation values. Variance (mm^2) was computed using the formula: $Var_e = \frac{1}{M}\sum_{m=1}^{M}[\frac{1}{n-1}\sum_{i=1}^{n}(e_{m,i} - e_{ave,m}^2)]$ (Eckel and Mass 2005). M represents the number of grid points, n is the number of ensemble members, $e_{m,i}$ is the i^{th} ensemble member's QPF value at grid point m, and $e_{ave,m}$ is the true ensemble mean (not probability matched mean) QPF value at grid point m. Variance was computed separately for each case at all 12 forecast hours. Then, the variance from each forecast hour was averaged over all 22 cases to compute the total ensemble variance over all cases. MSE of the true ensemble mean was computed using the formula: $MSE_{e,\ ave} = \frac{\frac{n}{n+1}}{M}\sum_{m=1}^{M}[e_{ave,m} - o_m]^2$ (Eckel and Mass 2005). M represents the number of grid points, n represents the number of ensemble members, $e_{ave,m}$ represents the true ensemble mean, and o_m represents the observed 3-h precipitation at grid point m. MSE of the true ensemble mean was averaged over all cases in the same way as variance. In addition, the correlation between ensemble variance and mean squared error (MSE) of the true ensemble mean was examined.

Rank Histograms (Hamill 2001) were computed for 3-h observed precipitation for both the rad-ens and norad-ens. For an ensemble with n members, observed precipitation values can fall into n+1 different bins. If there is a certain grid point in a 10 member ensemble where the observed precipitation value is greater than the QPF value of 6 ensemble members but less than the QPF value of the other 4 members, the “rank” of the observation would be 7 and thus one observation would be placed in bin number 7. This procedure of binning the observations by comparing their values relative to the QPF values of the 10 ensemble members was repeated and combined for all grid points and cases but separated by forecast hour, so in the end, there were 12 rank histograms for each ensemble. If the observed precipitation value was the exact same as the QPF value of one or more of the ensemble members, the bin was randomly selected between the possible bins based on the number of ties.

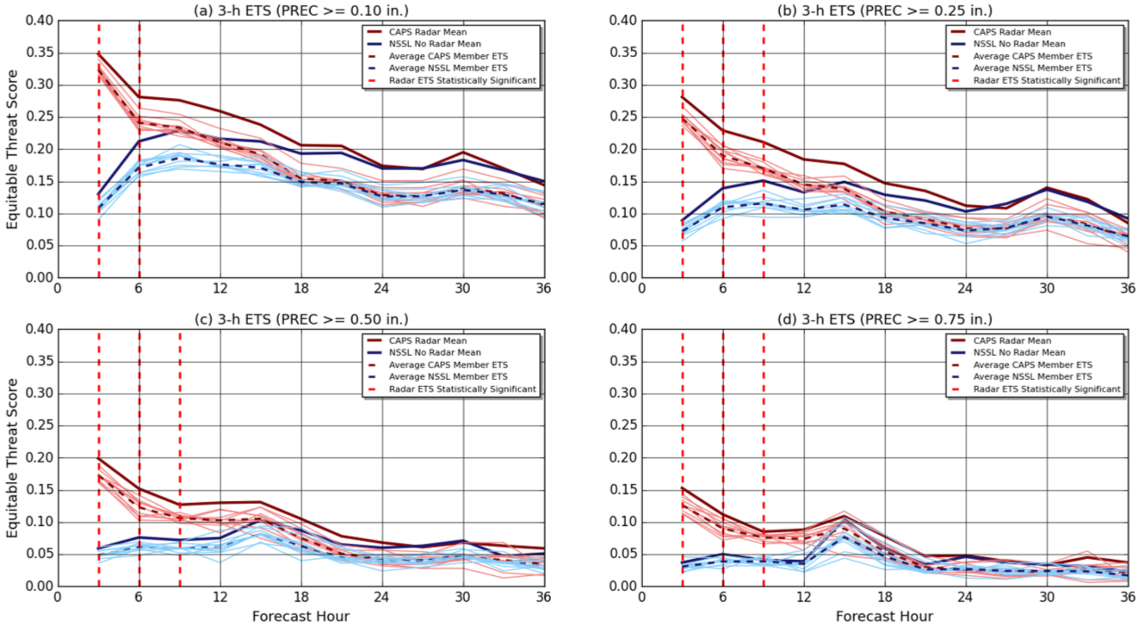


Figure 2 (a) 3 h ETS at each forecast hour for the 0.10 in. precipitation threshold. Hours with significant differences between ensemble means are indicated by red, dashed, vertical lines. Dashed horizontal lines indicate average member ETS. (b), (c), and (d) are same as (a) except for 0.25, 0.50, 0.75 in. thresholds, respectively.

iii) Results

a) ETSs

Figures 2a-d depict the 3-h ETSs from the 0.10- to 0.75-in. threshold for each of the ensemble members as well as the ensemble means for all 36 forecast hours in 3-h intervals. Hours that were significantly in favor of the rad-ens mean are denoted by red, dashed vertical lines. Generally, ETSs at the lowest two thresholds for the rad-ens mean tended to decrease steadily from hours 3-24, then more or less leveled off after that. For the norad-ens mean at all of the thresholds, there was not much of a change in ETS as time progressed from hour 3 to hour 36. At thresholds ≥ 0.50 in. for the rad-ens mean, there was not much of a decrease in ETS from hours 6 to 15, but a noticeable decrease from hours 15 to 21 before becoming steady through the end of the model forecast period. The rad-ens mean outperformed the norad-ens by at least 0.05 from hours 3 through 12 for all thresholds, and by around 0.03-0.04 for all thresholds except for the 0.75 in. threshold at forecast hour 15, where the rad-ens and norad-ens means had almost the same value. There was a small but noticeable (0.01-0.02) difference in the means out to hour 21 in favor of the rad-ens. So, the effect of Radar DA was greatest in the first 12 hours, but still there was a small difference in the means out to hour 21. It is important to note that the only statistically significant values in favor of the rad-ens mean were in hours 3, 6, and 9, despite the fact that the difference between the ensemble means was about as large in hour 12 as it was in hour 6 and 9 for all of the thresholds.

The individual ensemble members followed a similar pattern as the means, except that the differences between the individual members of the rad-ens and norad-ens are smaller, especially from hours 3-21. While there was a noticeable but very small difference in favor of the rad-ens out to hour 21 when looking at the ensemble means, there is essentially no difference in the average member ETS after hour 15 for all four thresholds. Figure 3 is very similar to Figure 2, but it is a plot of the control member ETS from each ensemble. The control members follow a trend similar to the ensemble means, but with slightly greater differences at the lower thresholds. For example, at the 0.10 in. threshold, the first 12 hours were significantly in favor of the rad-ens control member, but at the 0.75 in. threshold, only forecast hour 3 was significantly in favor of the rad-ens control member. When just looking at the control member ETSs side by side, there is a noticeable pattern that shows a slightly more positive effect of Radar DA at the lower thresholds.

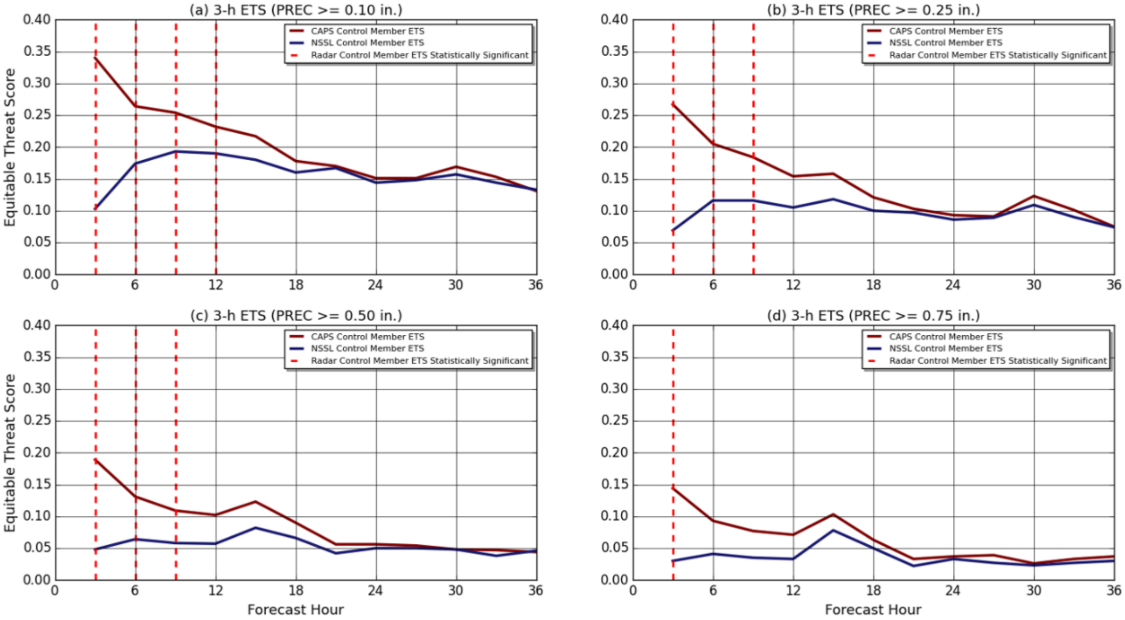


Figure 3 (a) 3 h ETS at each forecast hour for the 0.10 in. precipitation threshold for the control member of each ensemble. Hours with significant differences between ensemble control members are indicated by red, dashed, vertical lines. (b), (c), and (d) are same as (a) except for 0.25, 0.50, 0.75 in. thresholds, respectively.

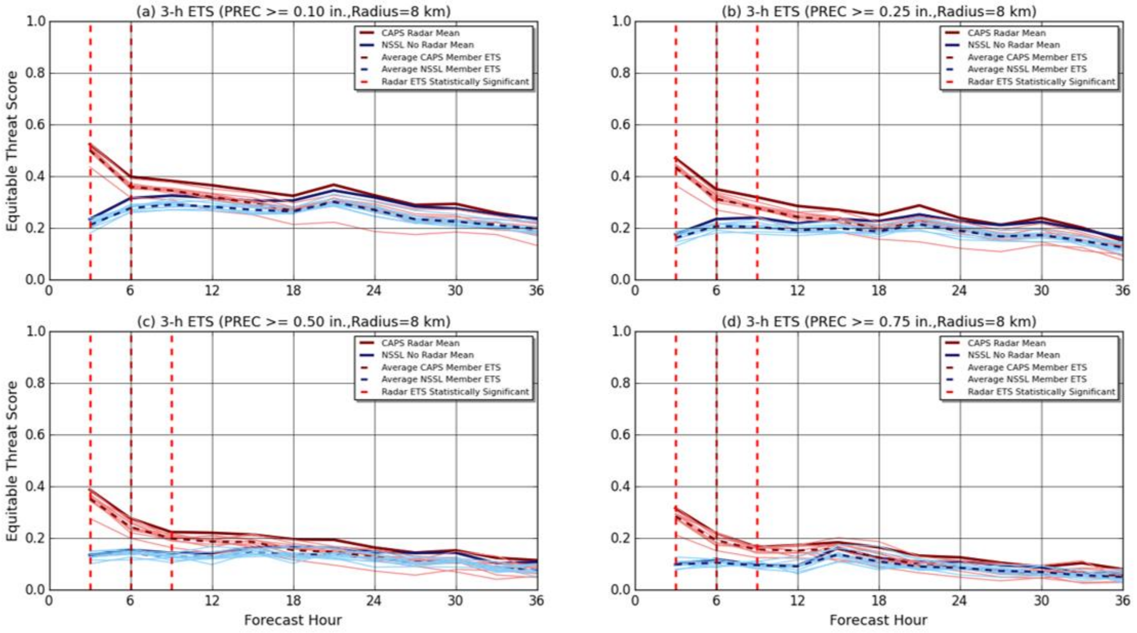


Figure 4 (a) 3 h ETS_r for r=8 km at each forecast hour for the 0.10 in. precipitation threshold. Hours with significant differences between ensemble means are indicated by red, dashed, vertical lines. Dashed horizontal lines indicate average member ETS_r. (b), (c), and (d) are same as (a) except for 0.25, 0.50, 0.75 in. thresholds, respectively.

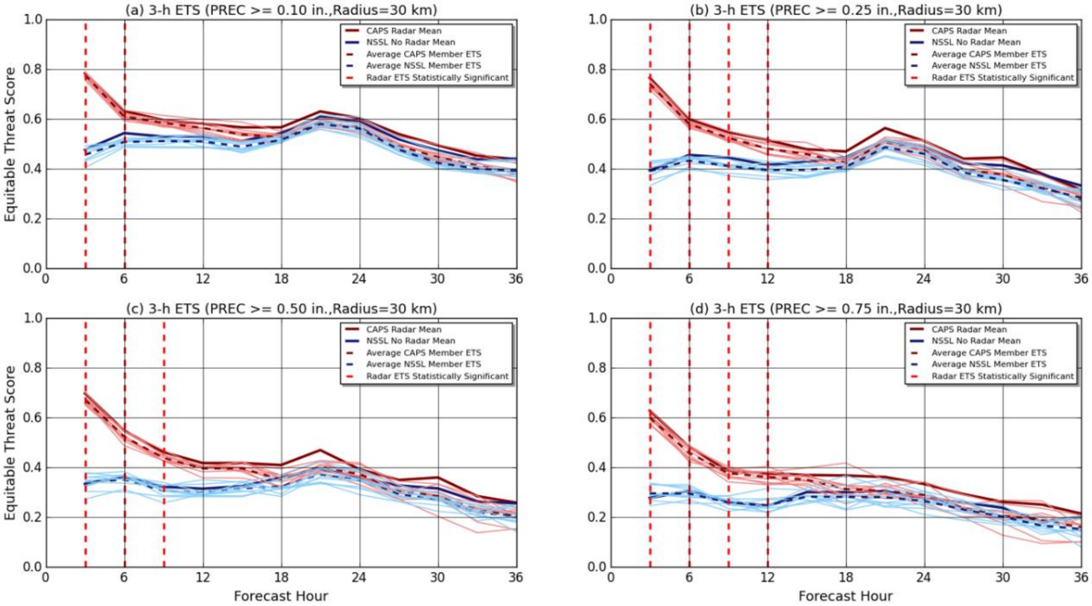


Figure 5 (a)-(d) Same as Figure 4 except for r=30 km.

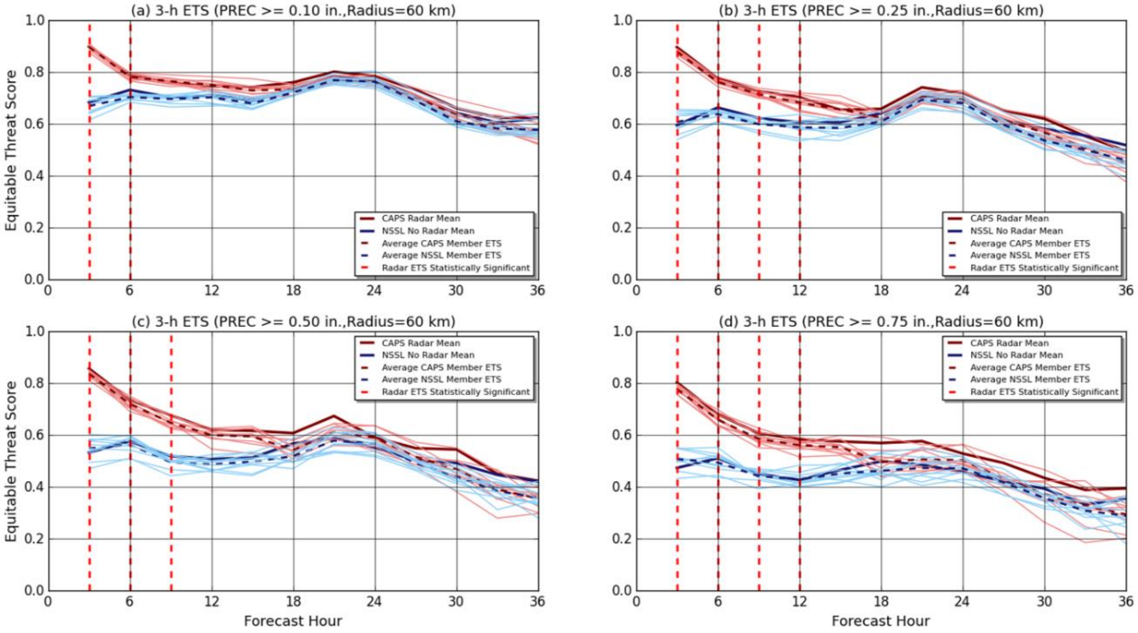


Figure 6 (a)-(d) Same as Figure 4 except for r=60 km.

Figures 4-6 are the same as Figure 2, except for ETS_r instead of ETS (Figures 4, 5, and 6 display ETS_r for radii of 8, 30, and 60 km, respectively). As expected and shown in numerous past studies (e.g., Clark et al. 2010) ETS_r increased as the radius was increased from 8 to 60 km. ETS_r for the rad-ens and norad-ens behaved similarly to ETS with the exact same hours of significance for the means which can be seen by comparing Figures 2 and 4. However, ETS_{30} and ETS_{60} tell a different story. Differences between the rad-ens and norad-ens ETS_r means widen as the radius is increased and also as the threshold is increased. It can be seen that there is a difference of at least 0.10 in ETS_{30} and ETS_{60} at all thresholds of 0.25 in. and greater out to forecast hour 12 in favor of the rad-ens mean over the norad-ens mean. Almost all of the hours from 3 to 12 exhibited statistical significance in favor of the rad-ens mean for ETS_{30} and ETS_{60}. The difference in ETS_{30} and ETS_{60} in favor of the rad-ens mean at the 0.75 in. thresholds is at least 0.05 from hour 15 all out to hour 36. This is very different than what was observed in Figures 2 and 3 for point-based ETS where there was virtually no difference between the means or control members after hour 24. Interestingly, although the differences in ensemble means seemed to increase as the radius and threshold increased from hours 15-36, the difference in average member ETS only increased slightly in favor of the rad-ens during 18-36, as there was a lot of variation between ensemble members. Some members even surpassed their respective ensemble means, especially at the higher radii and thresholds (see Figure 6 at the 0.50 and 0.75 in. thresholds). Overall, the benefits of Radar DA with respect to the ensemble means seem to increase as the neighborhood radius is increased and thus more weight is placed on the forecast having the correct precipitation amounts while being close in location rather than having to accurately predict the exact location of convection right down to the grid point. Also, since the ETS_r differences were more in favor of the rad-ens mean when thresholds were higher, this seems to suggest that there could be benefits of Radar DA all the way out to 36 hours in certain situations.

b) Area under the ROC curve

Figures 7a-d depict the 3-h ROC area for all 36 forecast hours for the 0.10, 0.25, 0.50, and 0.75 in. thresholds. The green shading represents statistically significant values in favor of the rad-ens. There is a noticeable, but steadily decreasing difference in ROC area between the two ensembles in forecast hours 3-15. Starting at forecast hour 18 through hour 36, there really is no difference in ROC area between the ensembles. The differences from hours 3-12 are significantly in favor of the rad-ens at all four thresholds, while the difference at hour 15 is not statistically significant at any threshold. This suggests that Radar DA does have an impact through the first 15 hours of a forecast, with the impact being greatest at the beginning of the forecast period.

It is widely accepted that a ROC area value of 0.7 is the threshold for a useful forecast. At the lowest two thresholds, all of the forecasts from hour 6 to 36 would be considered “useful” as per this measure. However, at both the 0.50 and 0.75 in. thresholds, the ROC area values of the rad-ens PQPF were above 0.7 from hours 3 through 12, while the ROC area values were below 0.7 for the rad-ens. Therefore, it can be said that at those higher thresholds, Radar DA made a difference between a forecast that is useful and not useful, according to the ROC area metric. So, much like what was observed for ETS_{30} and ETS_{60}, we see the largest positive impact of Radar DA at higher precipitation thresholds, which represent situations when thunderstorm complexes are present.

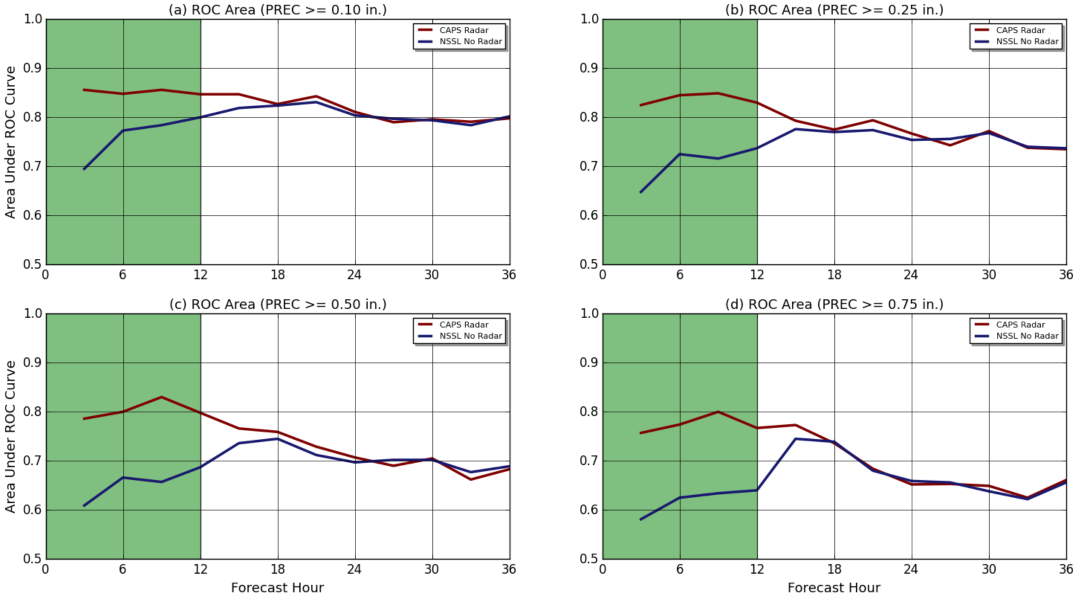


Figure 7 The 3-h ROC Area at each forecast hour for the (a) 0.10-, (b) 0.25-, (c) 0.50-, and (d) 0.75-in. precipitation thresholds. Hours with significant differences between ensemble PQPFs are indicated by green shading.

c) Ensemble Spread

Figure 8b depicts variance versus forecast hour for both ensembles. Overall, there is very little change in variance for both of the ensembles throughout the forecast period. There was a small relative maximum in variance during the afternoon and early evening hours (forecast hours 21-27) where diurnally driven single cell (and small multi-cell cluster) convection is responsible for a large amount of observed precipitation. While there is not a high amplitude diurnal cycle evident in terms of variance, but values do drop slightly during the morning hours (around forecast hours 12 and 36) when compared to the evening hours. When values of variance for each day and forecast hour are plotted versus MSE (Figures 8a and 8c) for each of the ensembles, MSE values are several (in some cases up to around 15) times larger than the variance values, pointing to the fact that there is indeed a lack of spread in these two ensembles. This was somewhat expected though, as all ensemble members for each of the ensembles used exactly the same physics parameterizations.

MSE (Figure 8d), on the other hand, does exhibit much larger changes versus forecast hour. The positive impact of Radar DA is evident in the first 12-15 hours, as the MSE for the rad-ens is about 0.2 to 0.5 mm^2 lower than that of the norad-ens. After hour 15, the MSE values of both ensembles are essentially equal. However, the MSE does not just steadily increase from hours 15-36 as was seen with ETS and ROC Area. Instead, MSE increases for both ensembles from hours 15-24, then it decreases from hours 24-36. The increases and decreases in MSE follow a diurnal cycle, with the peak at 0000 UTC (forecast hour 24) and the low point at 1200 UTC (forecast hours 12 and 36). Like what was seen with ensemble variance, the peak in MSE during the early evening likely occurs because there is a lot of diurnally driven single-cell convection during that time. So, if a single cell thunderstorm drops an inch of rain just 5 to 10 km east of the model forecast of that cell, then there will be large MSE value associated with both the forecast grid-point and the grid-point where the inch of rain was observed but none was forecast. Even a forecast that would be considered subjectively accurate by an operational forecaster could have a large MSE value.

Figure 9 consists of rank histograms valid at forecast hour 24 for each ensemble plotted on top of each other. It is clear from the figures that both ensembles have a high frequency of overestimating the amount of precipitation that fell. In around one-quarter of the cases, the observed 3-h precipitation value was lower than the QPF value of every ensemble member. This was somewhat of an expected result because CAMs can explicitly forecast individual convective elements. For example, if there were several days with diurnally-driven single-cell thunderstorms in the southeastern United States, the rank histograms would be extremely sensitive to the exact location where the ensemble members forecast the thunderstorm to occur. In other words, if the ensemble members fail to predict the exact locations where these single-cell thunderstorms occurred, then there will be a lot of grid-points where the observed precipitation values are lower than the forecast values (e.g. the members predicted a thunderstorm to form in city A, and in reality, it formed 25 miles to the east). The grid-points in and around city A would all see the observed QPF values fall into bin “1” on Figure 9. The converse to this may also be true in some cases. Taking the example with city A, the grid-points 25 miles east of the city where the storm *did* form would see observed QPF values fall into bin “11” on Figure 9. However, since we see a lot more grid-points fall into bin 1, we can conclude that both ensembles over-predicted precipitation as a whole, and also may have overestimated the amount of diurnally-driven single-cell thunderstorms that formed on days conducive to their development.

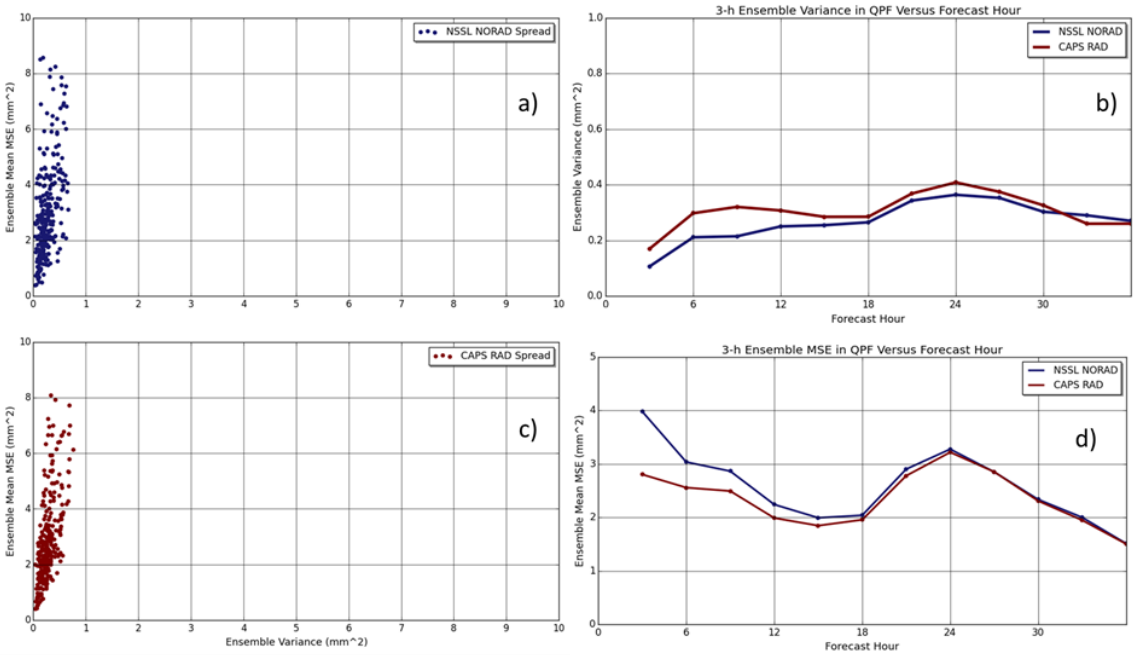


Figure 8 a) The 3-h variance at each forecast hour for the NSSL (norad-ens) ensemble (blue) and the CAPS ensemble (red), b) variance versus forecast hour for each ensemble, c) same as a) except for the CAPS (rad-ens) ensemble, and d) MSE versus forecast hour for each ensemble.

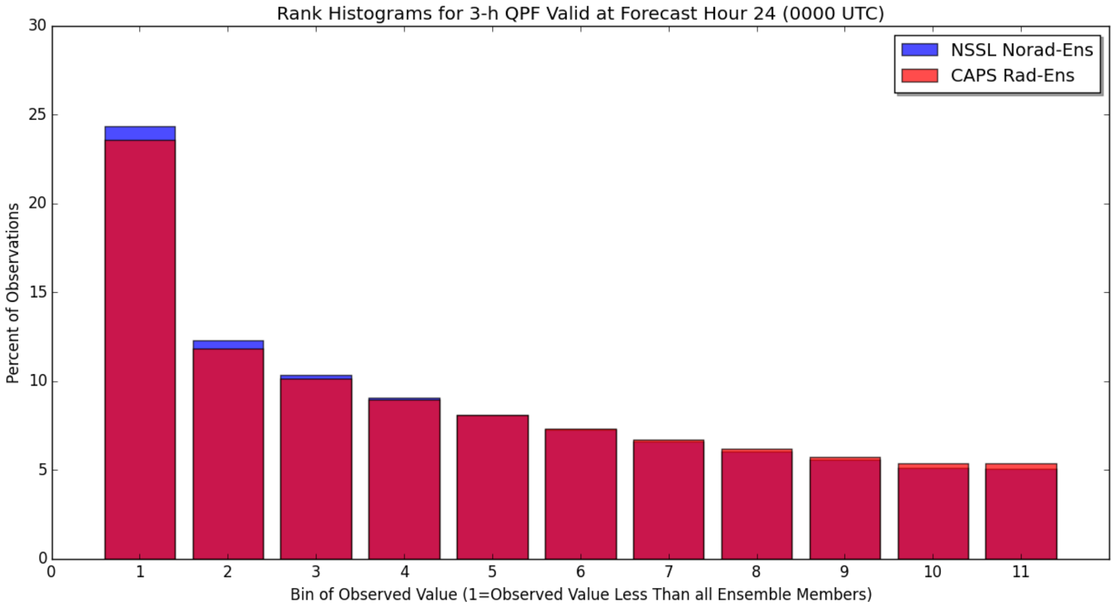


Figure 9 Rank Histograms overlaid for 3-h QPF valid at forecast hour 24 for the Norad-Ens (red) and Rad-Ens (blue).

d) Individual Cases

The ROC Area for both ensembles was calculated for each of the 22 cases in order to examine the differences in the ensembles for each individual case. Figure 10 shows the distribution of ROC Area differences (rad-ens minus norad-ens) for each of the ensembles at forecast hours 12 and 24. For both forecast hours, the median difference was between 0 and 0.1 in favor of the Radar ensemble. However, at forecast hour 12, almost one-third (7 out of 22) of the cases had ROC Area differences of greater than 0.1. At forecast hour 24, just one out of the twenty-two cases had a difference of greater than 0.1. In slightly more than half of the cases, norad-ens had a greater ROC Area value at hour 24. Norad-ens had a greater ROC Area than that of the Radar ensemble for only 4 out of the 22 cases at forecast hour 12. Two selected cases of 3-h QPF forecasts and observations at forecast hour 12 are displayed in Figure 11.

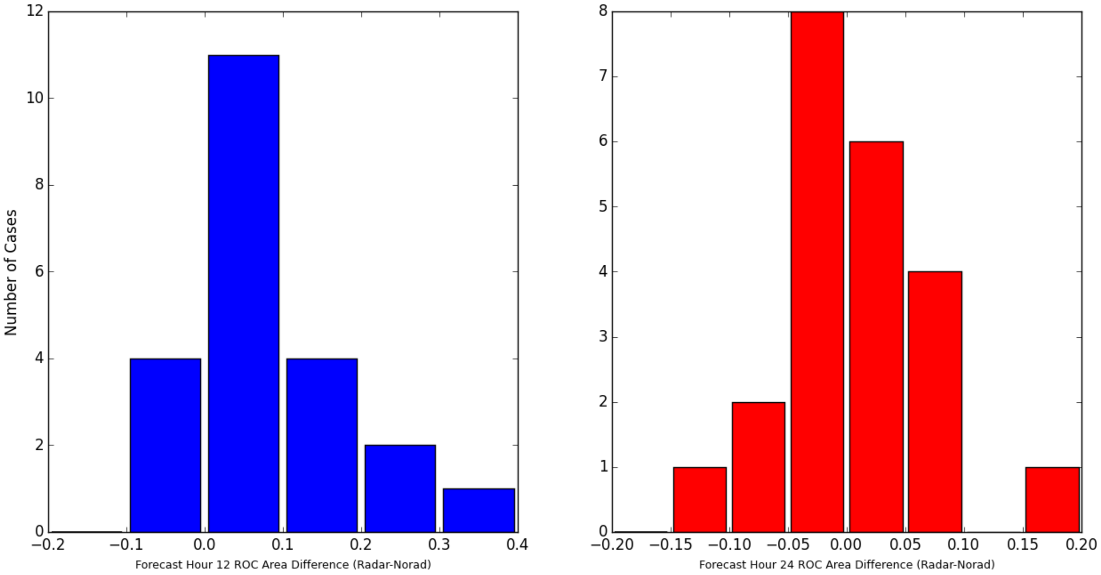


Figure 10 Histograms of ROC area difference (rad-ens minus norad-ens) for forecast hour 12 (blue) and forecast hour 24 (red).

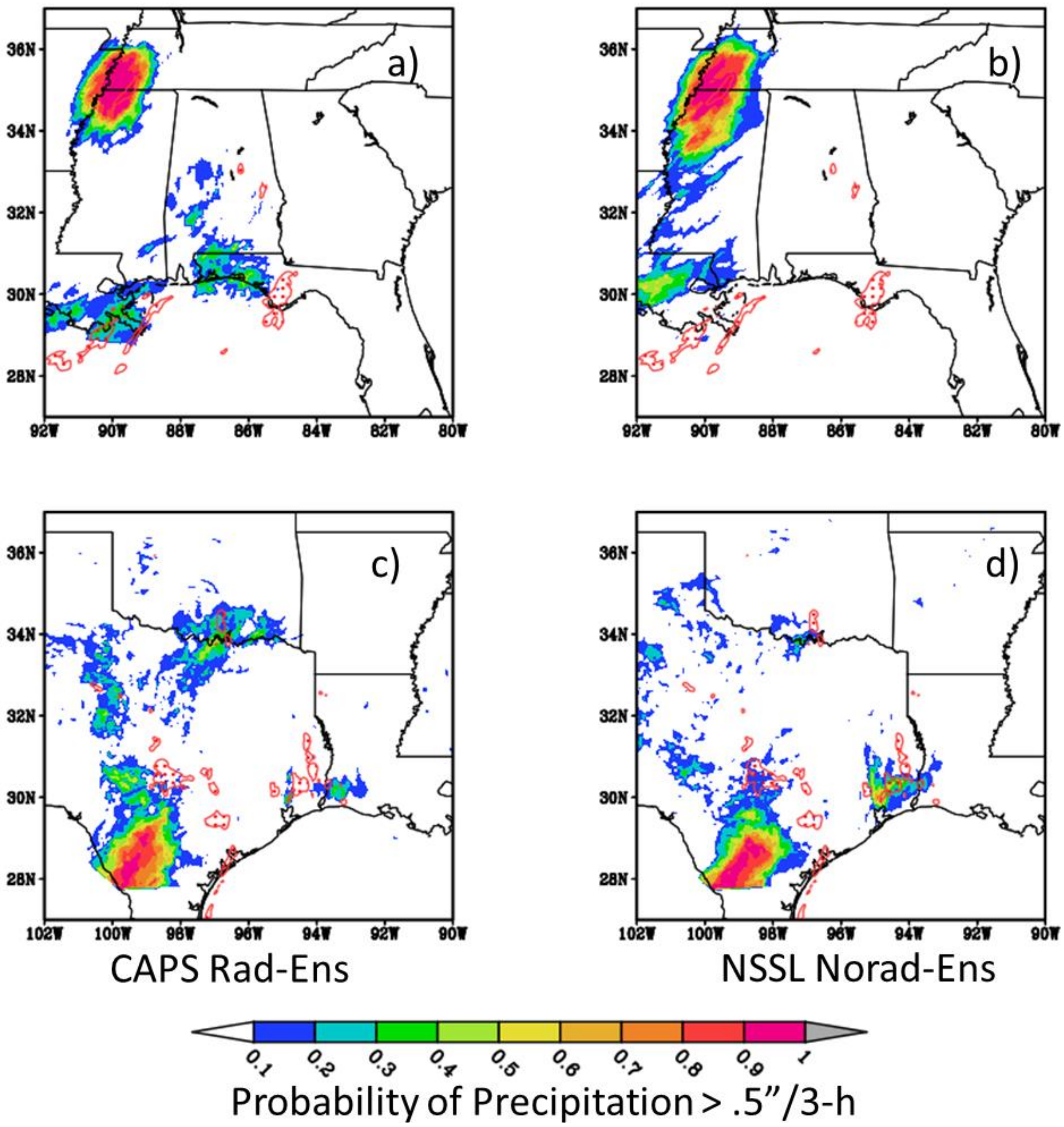


Figure 11 The 3-h PQPF forecasts for the 0.50-in. threshold valid from 0900 to 1200 UTC 20 May 2016 and 0600 to 0900 UTC 2 June 2016, in color, overlaid with 3-h observed precipitation of 0.50 in. in red stippling for the (a) rad-ens on May 20, (b) norad-ens on 20 May, (c) rad-ens on 2 June, and (d) norad-ens on 2 June.

Figures 11a and 11b show the PQPFs at forecast hour 12 on 20 May 2016, respectively, valid for the 0.50-in. threshold, ROC Area, and 3-h observed precipitation of 0.50 in. or greater. Figures 11c and 11d are the exact same as 11a and 11b except that they are valid ending at 1200 UTC on 2 June 2016. There were some significant differences in the synoptic pattern in the featured regions on 20 May and 2 June 2016. The 20 May case actually began on 19 May as an MCS formed over southeastern Texas during the afternoon of 19 May. The MCS tracked eastward along the Gulf Coast overnight and reached the Florida Panhandle at around 0900 UTC on 20 May. The flow at 500 hPa was around 30 kt throughout the southeastern United States, and there was a 500 hPa low present near Kansas City, several hundred miles north of the MCS, at 1200 UTC on May 20. At the same time, there was a weak 1010 hPa surface low in western Mississippi along with an associated warm front stretching eastward from Mississippi through South Carolina that was situated about 200 km north of the center of the MCS. In addition to producing heavy precipitation, there were several severe wind reports along the Gulf coast associated with the 20 May MCS. While both ensembles had a large area of false alarm in western Tennessee and northwestern Mississippi, the norad-ens failed to depict the eastern progression of the MCS into the Florida Panhandle. Instead, it only showed the eastern edge of the heavy precipitation entering south-central Louisiana by 1200 UTC. The rad-ens did fairly well in picking out the eastern extension of the MCS by 1200 UTC, as it had a line of higher probabilities from the Florida Panhandle northward through Alabama, close to where 3-h precipitation greater than 0.50 in. was observed. The rad-ens had an accurate forecast of heavy precipitation in southeastern Louisiana, while the norad-ens was about 100-150 km too far west with its forecast there.

This case where a large MCS dominated the precipitation pattern is just one example of when the rad-ens outperformed the norad-ens. There were other cases where MCSs formed and Radar DA provided substantial benefits.

However, there were a few cases (such as 2 June) where the NoRad had better ROC Area where mainly heavy, stratiform precipitation was observed. From 2 to 3 June 2016, a closed 500 hPa low slowly moved across Texas. By 1200 UTC on 2 June, the 500 hPa low was located about 200 km northwest of San Antonio, TX. There was an associated surface low about 150 km east of the 500 hPa low. While there were a few thunderstorms, the main focus was a large area of moderate to heavy rain rotating around the low pressure system. The intensity of the rain was heavy enough to produce rainfall amounts in excess of 0.50 in./3 h in spots. Figures 11c and 11d show 3-h QPFs with 12 h lead times for the rad-ens and norad-ens, respectively. Overall, there is not a large difference between the two forecasts and both forecasts were west of the observed 0.50 in. + precipitation amounts in central Texas. However, the norad-ens accurately depicted the heavy precipitation in southeastern Texas, as it had probabilities of over 50% in the area where 0.50 in. + rainfall was observed. In that same area, the rad-ens did not have forecast probabilities of 0.50 in. + precipitation higher than around 10%. So, even for this forecast with 12 h lead time, the norad-ens slightly outperformed the rad-ens.

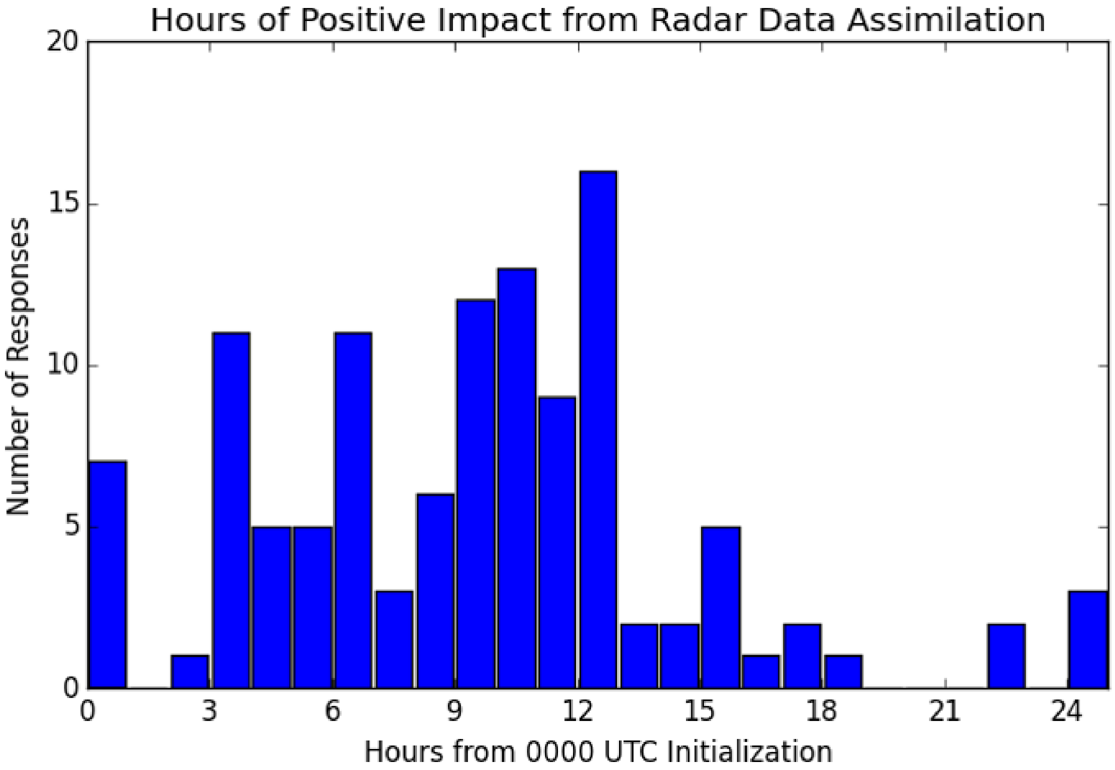


Figure 12 Participant responses to the question: “How long into the forecast does the assimilation of radar data have a positive impact on the [CAPS] ensemble forecast?” The y-axis indicated number of participant responses and the x-axis indicates hours after 0000 UTC initialization.

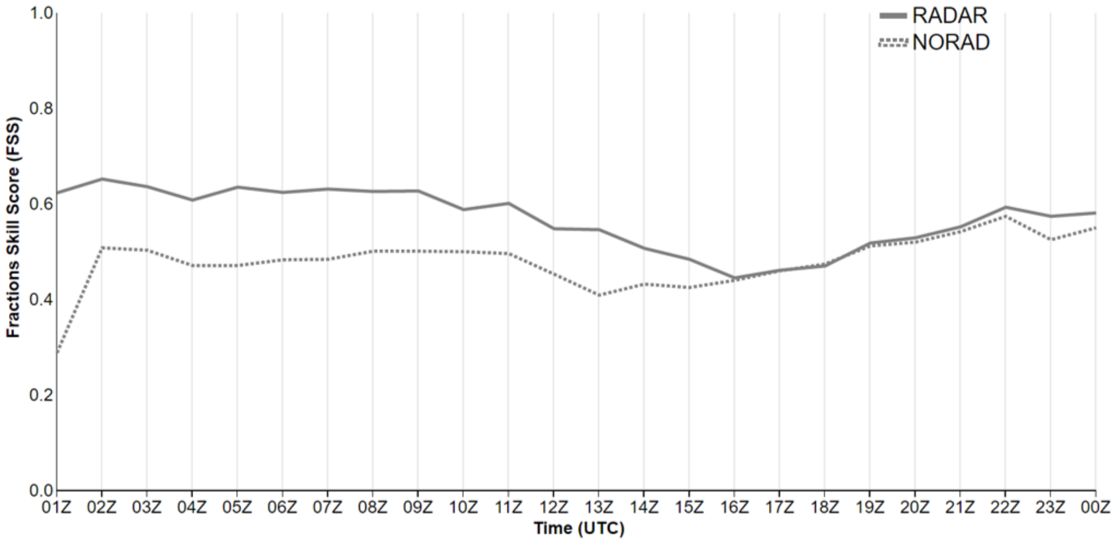


Figure 13 1-h FSS for ensemble probabilities of reflectivity > 40dBZ versus forecast hour for the CAPS ensemble (labeled RADAR) and the NSSL ensemble (labeled NORAD).

e) Subjective Versus Objective Evaluations of Ensemble Reflectivity Forecasts

Figure 12 is a bar graph of the number of responses (from 2016 SFE participants) versus hours after 0000 UTC initialization to the question: How long into the forecast does the assimilation of radar data have a positive impact on the ensemble forecast [for probability of reflectivity > 40 dBZ]?" Figure 13 is a plot of the 1-h FSS versus forecast hour for both the rad-ens (labed RADAR) and norad-ens (labeled NORAD) for the ensemble probabilities of reflectivity >40 dBZ. There is a clear difference in the FSS in favor of the rad-ens through forecast hour 13, and that difference steadily declines before the norad-ens catches up at forecast hour 16. After that, both ensembles have essentially the same FSS through forecast hour 24. On the other hand, the vast majority of the 2016 SFE participant responses indicated only between 3 and 12 hours of positive impact from Radar DA much less than the FSS indicated. However, a minority of responses stated that there were 13 to 18 hours of positive impact, and there were even a few that saw the positive impact of Radar DA out to 24 hours.

It is important to note that this is not quite an "apples to apples" comparison, as the FSS statistics are cumulative, while each participant response were based on just one day of forecasts, as there were around five or so participants completing the subjective evaluations daily. Perhaps this discrepancy between the subjective evaluations and FSSs exists because almost a third of the days evaluated did not have organized thunderstorm activity, so there was less positive benefit of Radar DA. However, on a few of the days, the participants noticed enhanced benefits of Radar DA past 12 hours and in some cases up to 24 hours. Radar DA benefits may have been enhanced during the 2016 SFE when there was organized convection, but more work would have to be done to test this inference further. It does make sense considering that some of the ETS/ROC Area results indicated a larger difference in favor of the rad-ens at the higher thresholds. And, a difference in ETS_{30} and ETS_{60} between the two ensembles was evident even out to and past forecast hour 24.

iv) Summary and discussion

3-h QPFs as well as 1-h probabilistic reflectivity forecasts were analyzed from the 3-km grid-spacing rad- and norad- ensembles. The forecasts were initialized at 0000 UTC on 22 days during the 2016 NOAA/HWT Spring Forecasting Experiment from 2 May to 3 June. Both ensembles have exactly the same configuration, except that the rad-ens includes 3DVAR Radar DA in its initial conditions. The goals of this study were to 1) to see how far into the 36 h forecast period the positive effects of radar data assimilation extend, and 2) to compare the subjective evaluations of the probabilistic reflectivity forecasts completed by the 2016 SFE participants to the objective evaluation metrics based on the reflectivity forecasts from both of the ensembles. The analysis to complete the goals outlined above was done by evaluating QPFs by computing ETSs and neighborhood ETSs (ETS_r) with radii ranging from 8 to 60 km for both ensembles, evaluating PQPFs by calculating ROC Area for both ensembles, and comparing subjective evaluations to FSSs computed from the probabilistic reflectivity forecasts from both ensembles. Hypothesis testing was carried out on the ETS, ETS_r, and ROC Area. Results are summarized below.

ETSs computed from the QPFs the CAPS and NSSL ensemble means were below 0.25 for the vast majority of forecast hours and thresholds. The difference between the rad-ens and norad-ens means (and thus the positive impact of Radar DA) peaked at hour 3. However, there was a difference of at least 0.05 at all thresholds from hours 6-12 and a slightly smaller difference at hour 15. After that, the ETS from the rad-ens mean was only 0.01-0.02 points higher through forecast hour 21, and both ensemble means had essentially the same ETS from hours 24-36. Statistical significance in favor of the rad-ens mean was present at hours 3 and 6 for the 0.10 in. threshold and at hours 3, 6, and 9 at the 0.25 to 0.75 in. thresholds. The ETSs of individual ensemble members followed a similar pattern as the ETSs from the means, except that the average member ETS difference between the ensembles was smaller at all hours from 3-36, and virtually nonexistent after forecast hour 15. ETSs from the control members of each ensemble were more in favor of the rad-ens at the lower thresholds, with the first 12 hours exhibiting statistical significance at the 0.10 in. threshold, but only hour 3 attained statistical significance. The results for ETS_r were highly dependent upon radius. As past studies such as Clark et al. (2010) have shown, scores increase as radius is increased. A large increase in scores was observed as the neighborhood radius was increased from 8 to 30 to 60 km. ETS_8 behaved very similarly to ETS, with the exact same hours of significance observed in favor of the rad-ens mean over the norad-ens. At ETS_{30} and ETS_{60}, as the general increase in score was observed for both ensembles as well as the means, the difference between the rad-ens and norad-ens also widened. In addition, the largest differences were seen at the 0.50 and 0.75 in. thresholds. While only the first 12 hours were statistically significant in favor of the rad-ens mean at the 0,75 in, threshold at ETS_{30} and ETS_{60}, there is a difference of at least 0.05 in favor of the rad-ens from hours 15-36. This was quite different from what was observed with ETS, where there was almost no difference in the ensemble means after hour 15.

The results from calculating ROC area from PQPFs of the two ensembles behaved similarly to the ETS_{30} and ETS_{60} for the first 15 hours. At all four thresholds, hours 3, 6, 9, and 12 had statistically significant differences in favor of the rad-ens. From hours 18-36, there were virtually no difference in ROC area between the rad-ens and norad-ens. At both the 0.50 and 0.75 in. thresholds from hours 3 through 12, ROC area values of the rad-ens ensemble were above the widely used 0.7 "usefulness" threshold for ROC area. During those same hours at the 0.50 and 0.75 in. thresholds, ROC area values were below 0.7 for the norad-ens. So, Radar DA was the difference between useful and not useful forecasts as per ROC area for those hours. All ROC area values for hours from 6-36 were above 0.7 for the 0.10 and 0.25 in. threshold for both ensembles.

Despite the differences in ETS and ROC area between the rad- and norad-ens, there was not a large difference in ensemble spread between the two. Both ensembles exhibited a low-amplitude diurnal cycle in variance (combined for all cases), with a peak around forecast hour 24. Additionally, values of variance throughout the forecast period were quite low (no greater than 0.4 mm^2). Even when the values of variance are broken out by day *and* forecast hour, there are no values above around 0.7 mm^2. When looking at MSE for both ensembles, the positive impact of radar DA was evident out to our 15, as the rad-ens had lower values. From hours 15 to 36, MSE values of both ensembles were essentially equal. There was a well-defined diurnal cycle present in the MSE values throughout the forecast period, with values peaking around hour 24. However, unlike ETS and ROC area, MSE values did not follow a declining trend from hours 3 to 36. Instead, the values stayed about the same and only varied based on the time of day. Constructing rank histograms revealed that both ensembles had a tendency to over-forecast precipitation. The rank histograms for the rad- and norad-ens looked nearly identical, with the observed precipitation value falling in bins 1 to 4 in over half of the cases. In nearly 25 percent of the cases, the observed 3-h precipitation value was less than the 3-h QPF forecast from *all* ensemble members of both the rad- and norad-ens. In contrast, the observed precipitation fell in the highest bin in only ~5% of the cases.

Subjective daily evaluations of 1-h probabilistic reflectivity > 40 dBZ forecasts from the NSSL/CAPS ensembles completed by 2016 NOAA/HWT SFE participants were examined and compared to 2016 SFE 1-h aggregate FSSs computed from those same forecasts for both of the ensembles. While the FSS pointed to a clear advantage for the rad-ens through hour 13 (this was consistent with the results obtained for ETS and ROC area), the daily participant responses answering the question: "How long into the forecast does the assimilation of radar data have a positive impact on the ensemble?" were between 3 and 12 h, with a mean and median of 9 h. A minority of responses were between 13 and 24 h. This discrepancy between the participant responses and the FSSs may indicate that while aggregate statistics from the 2016 SFE point to a benefit of Radar DA out to 12-15 h, there were many days where participants only saw a benefit out to 3-9 h. On a few of the days, the participants could see positive benefits into the 12-15 h range and even up to 24 h.

After examining all of the results, it is clear that all objective metrics computed from QPFs are in favor of the rad-ens at from the start of the forecast period to around forecast hour 12-15. At all four thresholds examined, the ETS of the rad-ens mean was at least 0.04 to 0.05 points greater than that of the NSSL mean out to hour 12. This difference also held true when examining the control members of each ensemble side by side. As expected, the greatest differences in the metrics between the ensembles occurred at forecast hour 3, when the positive effects of Radar DA were the greatest. The greatest differences were observed at the lower thresholds for point-based ETS.

For ROC area, similar to the ETSs, the advantage of the rad-ens peaked at forecast hour 3. However, ROC area indicated a larger positive effect of Radar DA than ETS did, as all forecast hours from 3 to 12 at all four thresholds had statistically significant differences in favor of the rad-ens. Unlike the ETS, the largest differences were observed at higher thresholds, as there was around a 0.05 point difference at ROC area out to hour 15 at the 0.50 and 0.75 in. thresholds. Neighborhood based ETS with > 30 km radii calculated from the QPFs of the two ensembles produced similar results to what was observed with ROC area for forecast hours 3 to 15. However, at the 0.75 in. threshold for ETS_{30} and ETS_{60}, there was at least a 0.05 point difference in ensemble mean ETS_r all the way out to forecast hour 36. This may indicate that Radar DA may provide some limited positive benefits in pinpointing the approximate location of larger-scale convection with a day or more of lead time. More work will have to be done on this in the future to support or reject this hypothesis.

Aggregate FSSs calculated from 1-h probabilistic reflectivity forecasts from the 2016 SFE produced results similar to those seen with ETS and ROC area with a clear advantage to the rad-ens from initialization to forecast hour 13-15. Despite the fact that all of the objective metrics based on QPFs and the FSSs calculated from the reflectivity forecasts pointed to a clear advantage for the rad-ens out to 12-15 h, participants saw less than 12 h of positive impact from Radar DA around 75% of the time. Therefore, the objective metrics calculated in this study clearly make a case for 12 to 15 h of positive impact from Radar DA and some limited positive impact in certain situations out to 24 h.

However, humans were not able to see positive impacts from Radar DA past 9 hours on half of the days although there were a select few respondents that were able to see positive impact out to 21-24 h.

Future work should continue to examine the forecast range at which Radar data assimilation can provide positive benefits to precipitation and/or radar reflectivity forecasts. This includes the need to conduct studies that are longer term than the duration of one Spring Forecasting Experiment in addition to more closely examining the possible benefits of Radar DA in forecasts with up to 24 h lead time. Since operational forecasters are ultimately the end users of these forecast products, it will also be important to include more subjective evaluations and input from these operational forecasters as well as researchers. The NOAA/HWT Spring Forecasting Experiments serve as an ideal place for these tests, since both researchers and operational forecasters can be involved in the evaluations. The preliminary results from this study suggest that there are indeed at least some benefits of including Radar DA in forecasts up to 15 h (in a few cases up to 24 h) of lead time.

References

Accadia C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2003: Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Wea. Forecasting*, **18**, 918–932. Arakawa A., 2004: The cumulus parameterization problem: past, present, and future. *J. Climate*, **17**, 2493–2525.

Arakawa A., 2004: The cumulus parameterization problem: past, present, and future. *J. Climate*, **17**, 2493–2525.

Baldwin, M., and K. Mitchell, 1997: The NCEP hourly multi-sensor U. S. precipitation analysis for operations and GCIP research, *13th AMS Conference on Hydrology, Am. Meteorol. Soc., Long Beach, Calif.*

Berenguer M., M. Surcel, I. Zawadzki, M. Xue, and F. Kong, 2012: The diurnal cycle of precipitation from continental radar mosaics and numerical weather prediction models. part ii: intercomparison among numerical models and with nowcasting. *Mon. Wea. Rev*., **140**, 2689–2705.

Betts A. K. and B. A. Albrecht, 1987: Conserved variable analysis of the convective boundary layer thermodynamic structure over the tropical oceans. *J. Atmos. Sci.*, **44**, 83–99.

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

Clark A. J., W. A. Gallus Jr., M. Xue, and F. Kong, 2009: A comparison of precipitation forecast skill between small convection-allowing and large convection-parameterizing ensembles. *Wea. Forecasting*, **24**, 1121–1140.

Clark A. J., W. A. Gallus Jr., M. Xue, and F. Kong, 2010: Convection-allowing and convection-parameterizing ensemble forecasts of a mesoscale convective vortex and associated severe weather environment. *Wea. Forecasting*, **25**, 1052–1081.

Clark A. J., W. A. Gallus Jr., M. Xue, and M. Weisman, 2010: Neighborhood-Based Verification of Precipitation Forecasts from Convection-Allowing NCAR WRF Model Simulations and the Operational NAM. *Wea. Forecasting*, **25**, 1495–1509.

Clark, A. J., and Coauthors, 2012: An overview of the 2010 Hazardous Weather Testbed Experimental Forecast Program Spring Experiment. *Bull. Amer. Meteor. Soc*., **93**, 55-74.

Clark A. J., and Coauthors, 2016: Spring Forecasting Experiment 2016 Conducted by the Experimental Forecast Program of the NOAA/Hazardous Weather Testbed: Program Overview and Operations Plan. 30 pp. [Available online at https://hwt.nssl.noaa.gov/Spring_2016/HWT_SFE2016_operations_plan_final.pdf]

Coniglio, M. C., 2009: Next-Day Convection-Allowing WRF Model Guidance: A Second Look at 2-km versus 4-km Grid Spacing. *Mon. Wea. Rev.*, **137**, 3351–3372.

Done, J., C. A. Davis, and M. Weisman 2004: The next generation of NWP: explicit forecasts of convection using the weather research and forecasting (WRF) model. Atmosph. Sci. Lett., 5: 110–117.

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

Dyer, A. J. and B. B. Hicks, 1970: Flux-gradient relationships in the constant flux layer. *Q.J.R. Meteorol. Soc.*, **96**, 715–721.

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

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

Evans C., D. F. Van Dyke, and T. Lericos, 2014: How do forecasters utilize output from a convection-permitting ensemble forecast system? case study of a high-impact precipitation event. *Wea. Forecasting*, **29**, 466–486.

Fels S. B. and M. D. Schwarzkopf, 1975: The simplified exchange approximation: a new method for radiative transfer calculations. *J. Atmos. Sci.*, **32**, 1475–1488.

Ferrier, B. S., Y. Lin, T. Black, E. Rogers, and G. DiMego, 2002: Implementation of a new grid-scale cloud and precipitation scheme in the NCEP Eta model. Preprints, *15th Conf. on Numerical Weather Prediction, San Antonio, TX, Amer. Meteor. Soc.*, 280-283.

Fritsch J. M., R. A. Houze Jr., R. Adler, H. Bluestein, L. Bosart, J. Brown, F. Carr, C. Davis, R. H. Johnson, N. Junker, Y-H. Kuo, S. Rutledge, J. Smith, Z. Toth, J. W. Wilson, E. Zipser, and D. Zrnic, 1998: Quantitative precipitation forecasting: report of the eighth prospectus development team, U.S. weather research program. *Bull. Amer. Meteor. Soc.*, **79**, 285–299.

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

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

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

Hong S.Y. and H. L. Pan, 1996: Nonlocal boundary layer vertical diffusion in a medium-range forecast model. *Mon. Wea. Rev.*, **124**, 2322–2339.

Hong S. Y. and J. J. Lim, 2006: The WRF Single Moment 6-Class Microphysics Scheme (WSM6). *J. Korean Meteorological Society*, **42**, 129-151.

Hu M., M. Xue, and K. Brewster, 2006: 3DVAR and cloud analysis with WSR-88D Level-ii data for the prediction of the Fort Worth, Texas, tornadic thunderstorms. part i: cloud analysis and its impact. *Mon. Wea. Rev.*, **134**, 675–698.

Janjić Z. I., 1990: The step-mountain coordinate: physical package. *Mon. Wea. Rev.*, **118**, 1429–1443.

Janjić, Z. I., 2002: Nonsingular implementation of the Mellor-Yamada Level 2.5 Scheme in the NCEP Meso model. NCEP Office Note No. 437, NOAA/NWS, 61 pp.

Janjić, Z. I., 2003: A nonhydrostatic model based on a new approach. *Meteor. Atmos. Phys.*, **82**, 271–285.

Janjić, Z. I., 2005: A unified model approach from meso to global scales. Geophysical Research Abstracts, Vol. 7, Abstract 05582. [Available online at http://www.cosis.net/abstracts/EGU05/ 05582/EGU05-J-05582.pdf

Janjić, Z. I. 2010: Recent advances in global nonhydrostatic modeling at NCEP. Proc. Workshop on Non-hydrostatic Modelling, ECMWF, Reading, United Kingdom. [Available online at http://nwmstest.ecmwf.int/newsevents/meetings/workshops/2010/Non_hydrostatic_Modelling/presentations/ Janjić.pdf

Janjić Z. I. and T. Black, 2007: An ESMF unified model for a broad range of spatial and temporal scales. Geophysical Research Abstracts, Vol. 9, Abstract 05025. [Available online at http:// meetings.copernicus.org/www.cosis.net/abstracts/EGU2007/ 05025/EGU2007-J-05025.pdf

Janjić Z. I. and R. Gall, 2012: Scientific documentation of the NCEP Nonhydrostatic Multiscale Model on the B Grid (NMMB). Part 1: Dynamics. NCAR/TN-4891STR, 75 pp. [Available online at http://nldr.library.ucar.edu/repository/assets/technotes/ TECH-NOTE-000-000-000-857.pdf

Janjić Z. I., T. Janjić, and R. Vasic, 2011: A class of conservative fourthorder advection schemes and impact of enhanced formal accuracy on extended-range forecasts. Mon. Wea. Rev., 139, 1556–1568.

Kain J. S. and J. M. Fritsch, 1998: Multiscale convective overturning in mesoscale convective systems: reconciling observations, simulations, and theory. *Mon. Wea. Rev.*, **126**, 2254–2273.

Kain, J. S., S. J. Weiss, J. J. Levit, M. E. Baldwin, and D. R. Bright, 2006: Examination of Convection-Allowing Configurations of the WRF Model for the Prediction of Severe

Convective Weather: The SPC/NSSL Spring Program 2004. *Wea. Forecasting*, **21**, 167–181.

Kain J.S. and Coauthors, 2010: Assessing advances in the assimilation of radar data within a collaborative forecasting–research environment. Wea. Forecasting, 25, 1510–1521Li, Y., X. Wang, and M. Xue, 2012: Assimilation of radar radial velocity data with the WRF hybrid ensemble–3DVAR system for the prediction of Hurricane Ike (2008). *Mon. Wea. Rev.*, 140, 3507–3524.

Kong, F., and Coauthors, 2014: An Overview of CAPS Storm-Scale Ensemble Forecast for the 2014 NOAA HWT Spring Forecasting Experiment. 27th Conference on Severe Local Storms, AMS, Madison WI, Paper 43.

Lacis, A.A., and J.E. Hansen, 1974: A parameterization for the absorption of solar radiation in the Earth's atmosphere. *J. Atmos. Sci.*, **31**, 118-133.

Lim K. S. and S. Y. Hong, 2010: Development of an effective double-moment cloud microphysics scheme with prognostic cloud condensation nuclei (ccn) for weather and climate models. *Mon. Wea. Rev.*, **138**, 1587–1612.

Ma J., Y. Zhu, D. Hou, X. Zhou, and M. Peña, 2014: Ensemble transform with 3d rescaling initialization method. *Mon. Wea. Rev.*, **142**, 4053–4073.

Melick C. J., I. L. Jirak, A. R. Dean, J. Correia Jr, and S. J. Weiss, 2012: Real Time Objective Verification of Convective Forecasts: 2012 HWT Spring Forecast Experiment. Preprints, *37th Natl. Wea. Assoc. Annual Meeting, Madison, WI.*

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

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

Milbrandt J. A. and M. K. Yau, 2005: A multimoment bulk microphysics parameterization. part i: analysis of the role of the spectral shape parameter. *J. Atmos. Sci.*, **62**, 3051–3064.

Monin, A.S and A. M. Obukhov, 1954: Basic laws of turbulent mixing in the surface layer of the atmosphere. *Tr. Akad. Nauk SSSR Geofiz. Inst* **24**, 163–187.

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. *J. Atmos. Sci.*, **62**, 1665–1677,

Nakanishi, M., 2000: Large-eddy simulation of radiation fog. *Bound.-Layer Meteor.*, **94**, 461–493.

Nakanishi M., 2001: Improvement of the Mellor–Yamada turbulence closure model based on large-eddy simulation data. *Bound.- Layer Meteor.*, **99**, 349–378.

Nakanishi M. and H. Niino, 2004: An improved Mellor–Yamada level-3 model with condensation physics: Its design and verifi-cation. *Bound.-Layer Meteor.*, **112**, 1–31.

Nakanishi M. and H. Niino, 2006: An improved Mellor–Yamada level-3 model: Its numerical stability and application to a regional prediction of advection fog. Bound.-Layer Meteor., 119, 397–407.

NOAA, 2015: 74 year list of severe weather fatalities. http://www.nws.noaa.gov/om/hazstats/resources/weather_fatalities.pdf

Noh Y., W. G. Cheon, and S. Raasch, 2003: The role of preconditioning in the evolution of open-ocean deep convection. *J. Phys. Oceanogr.*, **33**, 1145–1166.

Paulson, C. A., 1970: The mathematical representation of wind speed and temperature profiles in the unstable atmospheric surface layer. *J. Appl. Meteor.*, **9**, 857–861.

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

Roebber P. J., D. M. Schultz, B. A. Colle, and D. J. Stensrud, 2004: Toward improved prediction: high-resolution and ensemble modeling systems in operations. *Wea. Forecasting*, **19**, 936–949.

Rogers, E., and Coauthors, 2009: The NCEP North American Mesoscale modeling system: Recent changes and future plans. 23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction, Omaha, NE, Amer. Meteor. Soc., 2A.4. [Available online at https://ams. confex.com/ams/23WAF19NWP/techprogram/paper_ 154114.htm.]

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting*, 5, 570–575.

Schwartz, C. S., J. S. Kain, S. J. Weiss, M. Xue, D. R. Bright, F. Kong, K. W. Thomas, J. J. Levit, and Schwarzkopf, M. D., and S. B. Fels, 1991: The simplified exchange method revisited: An accurate, rapid method for computation of infrared cooling rates and fluxes, *J. Geophys. Res.*, **96**, 9075 – 9096.

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

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF Version 3, NCAR Tech Note, NCAR/TN-475+STR, 113 pp. [Available at: http://www.mmm.ucar.edu/wrf/users/docs/arw_v3.pdf.]

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

Sukoriansky, S., B. Galperian, and V. Perov, 2005: Application of a new spectral theory of stable stratified turbulence to the atmospheric boundary layer over sea ice. *Bound.-Layer Meteor.,* **117**, 231–257.

Sukovich E. M., F. M. Ralph, F. E. Barthold, D. W. Reynolds, and D. R. Novak, 2014: Extreme quantitative precipitation forecast performance at the weather prediction center from 2001 to 2011. *Wea. Forecasting*, **29**, 894–911.

Thompson G., R. M. Rasmussen, and K. Manning, 2004: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. part i: description and sensitivity analysis. *Mon. Wea. Rev.*, **132**, 519–542.

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

Troen, I. B. and L. Mahrt, 1986: A Simple Model of the Atmospheric Boundary layer; Sensitivity to Surface Evaporation. *Boundary-Layer Meteorology*, **37**, 129-148.

Vasiloff S. V., K. W. Howard, R. M. Rabin, H. E. Brooks, D. J. Seo, J. Zhang, D. H. Kitzmiller, M. G. Mullusky, W. F. Krajewski, E. A. Brandes, B. G. Brown, D. S. Berkowitz, J. A. McGinley, and R. J. Kuligowski, 2007: Improving qpe and very short term qpf: an initiative for a community-wide integrated approach. *Bull. Amer. Meteor. Soc.*, **88**, 1899–1911.

Weisman, M. L., W. C. Skamarock, and J. B. Klemp, 1997: The Resolution Dependence of Explicitly Modeled Convective Systems. *Mon. Wea. Rev.*, **125,** 527-548.

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

Weisman M. L., C. Davis, W. Wang, K. W. Manning, and J. B. Klemp, 2008: Experiences with 0–36-h explicit convective forecasts with the WRF-ARW model. *Wea. Forecasting*, **23**, 407–437.

Webb, E. K., 1970. Profile relationships: The log-linear range, and extension to strong stability. *Q.J.R. Meteorol. Soc.,* **96**: 67–90.

Xiao, Q., Y. Kuo, J. Sun, W. Lee, D. M. Barker, and L. Eunha, 2007: An approach of radar reflectivity data assimilation and its assessment with the inland QPF of Typhoon Rusa (2002) at landfall. *J. Appl. Meteor. Climatol.,* 46, 14–22.

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

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

Zhao, Q.-Y., and F. H. Carr, 1997: A prognostic cloud scheme for NWP models. *Mon. Wea. Rev.*, **125**, 1931-1953.

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

Zhu L. and Coauthors, 2015: Prediction and predictability of high-impact Western Pacific landfalling Tropical Cyclone Vicente (2012) through convection-permitting ensemble assimilation of doppler radar velocity. *Mon. Wea. Rev.*, **144**, 21–43.