THE INFLUENCE OF PBL PARAMETERIZATION ON THE PRACTICAL PREDICTABILITY OF CONVECTION INITIATION DURING THE MESOSCALE PREDICTABILITY EXPERIMENT (MPEX)

by

Bryan Burlingame

A Thesis Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Master of Science

in Mathematics

at

The University of Wisconsin-Milwaukee

May 2016

ABSTRACT

THE INFLUENCE OF PBL PARAMETERIZATION ON THE PRACTICAL PREDICTABILITY OF CONVECTION INITIATION DURING THE MESOSCALE PREDICTABILITY EXPERIMENT (MPEX)

by

Bryan Burlingame

The University of Wisconsin-Milwaukee, 2016
Under the Supervision of Professors Clark Evans and Paul Roebber

This study evaluates the influence of planetary boundary layer (PBL) parameterizations on short-range (0-15 h) forecasts of convection initiation (CI) within convection-allowing ensembles that utilize sub-synoptic-scale observations collected during the Mesoscale Predictability Experiment (MPEX). Running five thirty-member ensembles with the Advanced Research Weather Research and Forecasting Model (WRF-ARW) with each differing only in the chosen PBL parameterization, forecast skill, PBL sensitivity on the environment in which CI occurred, and the variability within are examined. Three MPEX cases, 19-20 May 2013, 31 May-1 June 2013, and 8-9 June 2013 are considered, each characterized by a different large-scale flow pattern to analyze a wider spectrum of events. Using an object-based method to verify and analyze the forecasts of CI, it was found that none of the Five PBL schemes analyzed significantly improved the forecast skill. The non-local mixing PBL schemes, MYJ and QNSE, had in all cases higher probability of detection (POD) but consequently had a higher false alarm ratio (FAR) resulting from the models overproducing the number of CI objects, with all PBLs, and thus resulting in relative high bias scores as well. The CSI showed only subtle changes between PBL schemes suggesting no one PBL scheme drastically outperforms the other. The temporal distribution of errors associated with the “hits” in the CI object matching showed an

approximate normal distribution around a mean of 0-s suggesting little systematic timing bias. While the spatial distribution of errors yielded skewed distributions with on average a mean (median) distance error of just over 44-km (28-km). Analysis of cumulative distribution functions (CDFs) of the "hits" highlighted limits to increased forecast skill beyond temporal and spatial thresholds of 60-min and 100-km. Mean error (ME) plots computed for surface features as well as vertical profiles in pre-convective environments highlighted biases in both the initial conditions as well as between ensembles. In agreement with previous studies, it was found that non-local mixing PBL schemes tend to produce PBLs that are too shallow, cool, and moist while local mixing schemes tend to be deeper, warmer, and drier as a function of the stronger (weaker) vertical mixing within the local (non-local) PBL schemes. Relative to the analysis of the vertical profiles, it was seen that the model has an inherent inability to accurately represent strong capping inversions in models across all PBL schemes suggesting an issue with the handling of vertical diffusion within the PBL and the implicit damping associated with the discretization schemes used within WRF.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. Introduction

The Mesoscale Predictability Experiment (MPEX; Weisman et al. 2015) tested the hypothesis that targeted sub-synoptic-scale observations in the upstream, pre-convective environment across the Intermountain West and their subsequent assimilation into convection-allowing numerical weather prediction (NWP) forecasts would lead to significant improvements in forecasts of convection timing, location, mode, and downstream evolution. MPEX launched its field campaign from 15 May 2013 to 15 June 2013 with motivation from studies (e.g., Roebber et al. 2002; Gold and Nielsen-Gammon 2008; Shafer et al. 2009; Schumann and Roebber 2010; Mercer et al. 2012) which demonstrated that there exist influences from the synoptic- and meso-α scales such as moisture source region, potential instability, and large-scale ascent from various features on deep moist convection (DMC). Resulting from MPEX, Schumacher (2015) found that for the 31 May-1 June 2013 supercell, increasing the resolution of the grid spacing (x) resulted in degradation of the forecast as characteristics of initiation changed substantially most likely as manifest from the chosen planetary boundary layer (PBL) parameterization scheme. Trier et al. (2015), using a convection-allowing 10-member ensemble, examined convection initiation (CI) along a daytime surface dryline to find that the minimum buoyancy of PBL air parcels must be small for successful deep CI to occur and that the maintenance of CI is affected by the grid-resolved forcing and forcing from temperature and moisture tendencies in the PBL scheme. Given the above, this study seeks to utilize ensembles of initial conditions that assimilated MPEX targeted observations to illuminate the control of PBL parameterization on the practical predictability of CI in convection-allowing ensembles.

Convection initiation is the culmination of a series of physical processes by which air parcels are brought to their level of free convection (LFC) by means of a triggering mechanism

and subsequently remain positively buoyant over a great vertical depth (Doswell 1987; Markowski and Richardson 2010). The initiation and maintenance of DMC and its transition into severe weather is vital to understand as it leads to the development of potentially life threatening severe weather that produces damaging winds, hail, tornadoes, and flash flooding that cause an estimated $4 billion in annual damage across the United States (Weiss et al. 2011). CI can result from many types of convergence both surface based and elevated, including but not limited to frontal boundaries, drylines, gust fronts, undular bores, orographic circulations, and elevated convergence zones as summarized by Jorgensen and Weckwerth (2003) and Weckwerth and Parsons (2006). CI however is not always successful as the environment in which it occurs is not always capable of sustaining a convective updraft and is sensitive to variability in near-surface temperature and moisture (e.g., Crook 1996) and lapse rate (Houston and Niyogi 2007).

The short-range predictability of CI has been examined previously for two categories of simulations, one in which convection is explicitly resolved and one where convection is parameterized. For convection-allowing studies the horizontal grid resolution is found to be $\Delta x <$ 4 km as this is the scale necessary to represent the evolution of convectively-driven mesoscale phenomena (Weisman et al. 1997; Bryan et al. 2003; Kain et al. 2008; Weisman et al. 2008). Convection-parameterizing simulations utilize grid spacings of $\Delta x > 10$ km, while the scales 4-km $< \Delta x <$ 10-km fall within a "gray zone" where some convection can be explicitly resolved and some cannot. Previous studies (e.g., Fowle and Roebber 2003; Done et al. 2004; Kain et al. 2008; Weisman et al. 2008) have illuminated that deterministic convection-permitting numerical solutions result in increased forecast skill with respect to precipitation, CI, and convective structure and evolution as compared to the convective-parameterizing simulations. Further, Clark et al. (2010) conducted a comparison of ten-member convection-allowing (Δx=4 km) and thirty-

member convection-parameterizing (Δx=20 km) ensembles in the forecast of a mesoscale convective vortex finding that the convection-allowing forecasts performed much better in the forecasts of the MCV track and associated environmental conditions (e.g., midlevel winds, low-level wind shear, and instability). Similarly, in comparing precipitation forecasts from a smaller (5-member) convection-allowing ensemble and a larger (15-member) convection-parameterizing ensemble, Clark et al. (2009) showed that various skill metrics evaluating precipitation forecasts reveal that the smaller convection-allowing ensemble providing more accurate precipitation forecasts at lead times of 9–21 h than its convection-parameterizing counterpart. Additionally, using convection-allowing resolutions has been found to outperform convection-parameterizing simulations in predicting both the diurnal precipitation cycle (Clark et al. 2007; Weisman et al. 2008) and MCS frequency and convective system mode (Done et al. 2004; Weisman et al. 2008). Further, both Weisman et al. (2008) and Kain et al. (2006) have suggested that convection-permitting forecasts exhibit greater skill in the forecast of CI timing and location than do the operational Δx = 12 km Eta numerical forecasts. Work by Duda and Gallus (2013), in which the predictability of warm-season CI in convection-permitting numerical forecasts is evaluated, found that there exists approximately a 100 km displacement error and no temporal bias in the forecast of CI. Burghardt et al. Additionally, Burghardt et al. (2013) examined the verification of CI using an object based approach within very high-resolution deterministic forecasts. They found there to exist an early overproduction of CI objects in the models as compared to the observed when object matching was done using a spatiotemporal error metric. By matching objects within only 100-km and 1-h, they also found on average there was a 38.4-km distance error in the forecast, and the model forecasted the 23 events with a probability of detection (POD) of 0.506. As for the capabilities of ensemble forecasts of CI within convection permitting

frameworks, Kain et al. (2013) have shown that simulations with Δx of ~ 4 km explicitly represented many key aspects of CI and did so with a relatively high POD and minimal systematic bias in the timing of the first CI event. Conversely, Kain et al. (2013) argues that CI forecasts, because of the relatively lenient verification thresholds they used for hazardous weather forecasts, are not particularly valuable and "were often inadequate indicators of impending hazardous or disruptive weather." For example the study points out that the CI identification algorithms often flagged initiation related to weak, elevated convection and that these flags were often false alarms for the CI of more severe, surfaced-based events.

The accurate prediction of CI is a problem which spans across the synoptic-, meso-α, meso-β, meso-γ, and micro- scales and is highly sensitive to the interactions between the surface and free atmosphere as manifest by the PBL (Stensrud 2007). This in part is due to situations involving moist dynamics become increasingly more complicated as manifest through upscale error growth resulting from the non-linearity of the mass and momentum fields that experience changes during moist convection (Baxter 2011; Zhang et al. 2002, 2003). The PBL structure reflects variability in low-level moisture, temperature, and winds which in turn affects thermodynamic instability, convective inhibition (Roebber et al. 2004; Hu et al. 2010; Coniglio et al 2013; Clark et al. 2015a; Cohen et al. 2015), and by extension CI. PBL parameterization schemes in NWP account for the sub-grid horizontal variability on the meso-β environment, the sub-grid vertical transport of turbulent kinetic energy (TKE), and the moistening of the meso-γ environment (Ziegler et al. 1997; Murphey et al. 2008; Markowski and Hannon 2006; Xue and Martin 2006; Weckwerth et al. 2008). The parameterization of vertical turbulent heat, moisture, and momentum fluxes is currently required in operational NWP as the horizontal grid spacing is too coarse to explicitly resolve the spatial and temporal scales of turbulence, or the irregular

fluctuations that occur in the fluids of the atmosphere (Bryan et al. 2003; Wyngaard 2004). Transport within the PBL is parameterized by different means of vertical mixing within the PBL scheme: local and non-local. Local schemes are characterized by the diagnosis of unknown atmospheric variables via only the adjacent model vertical levels to quantify the TKE whereas the non-local schemes attempt to simulate the effects of large eddies in the convective PBL using many vertical levels in the lower PBL model to estimate model fields. Readers are referred to section 2b of Coniglio et al. (2013) for additional analysis of the various PBL mixing characteristics and how they vary from one another.

Previous studies have examined the influences of PBL parameterization upon specific forecast elements within convection-allowing NWP forecasts. Hu et al. (2010), examining warm-season 36 h forecasts in Texas, analyzed the evolution of the PBL using the Asymmetric Cloud Model version 2 (ACM2), Mellor-Yamada-Janic (MYJ), and Yonsei University (YSU; all PBL schemes explained in section 2c) in high resolution Weather Research and Forecasting (WRF; Skamarock et al. 2008) simulations that ACM2 and YSU often were too vigorous in handling the mixing of the PBL resulting in PBLs too deep and dry. Coniglio et al. (2013) examined the 36 h performance of five different PBL schemes [ACM2, MYJ, MYNN (Mellor-Yamada-Nakanishi-Niino 2.5 closure), QNSE (Quasi-Normal Scale Elimination), and YSU] within a NWP forecast over the eastern ⅔'s of the contiguous United States (CONUS) to accurately forecast the warm-season pre-convective environment and found that local (non-local) PBL parameterization schemes produce PBLs that are generally too shallow and moist (deep and dry). For the same five PBL parameterizations, Clark et al. (2015a) found that all PBL schemes contained some east bias with respect to the dryline position 24 h warm-season forecasts, but comparatively vigorous mixing associated with non-local PBL schemes resulted in their mixing the dryline further east

than the local PBL schemes. Coniglio et al. (2013) goes on to highlight that because one PBL scheme may perform better for PBL structure over any given set of observations considered (e.g., MYNN outperformed the other PBL schemes in their study), it may not translate to better performance for specific forecast elements. Clark et al. (2015a) demonstrated this in the context of the aforementioned MYNN parameterization as it had the best representation of the pre-convective PBL structure in Coniglio et al. (2013) but had the worst dryline position errors in the Clark et al. (2015a) study. It stands to follow that for a specific forecast element such as CI, care must be taken when interpreting the findings of the study within. Cohen et al. (2015) examined the sensitivity of the same five PBL schemes, among others, of Coniglio et al. (2013) and Clark et al. (2015a) on the forecast of cold-season southeast US severe storm environments and found the tendency for local schemes to stunt the growth of the daytime PBL relative to non-local schemes. Markowski and Richardson (2010) argue that CI forecasting skill has advanced at a slower rate relative to our ability to forecast convective storm type, organization, and accompanying severe weather threats. Similarly, Lilly (1990) suggests that "the most difficult prediction task is to foresee development of the first convective storm in an area." Additionally, with all vertical diffusion in the WRF model being handled by the PBL parameterization used, the influence of vertical diffusion on model error is reflected by the influence of PBL parameterization on model error. Thus the need for further study and understanding of the limitations and biases in NWP on the forecast of CI and the associated influence of the PBL parameterization is present.

Given the above, this study focuses upon analyzing the sensitivities of the PBL parameterization on the forecast skill of warm-season CI at 0-15 h lead times utilizing convection-allowing ensembles initialized with initial analyses updated by the observations

gathered during the MPEX field campaign. The remainder of this thesis is structured as follows. The methodology, including the cases studied, model configuration, and CI verification techniques are presented in section 2. The results are discussed in section 3 and a summary and conclusions are provided in section 4.

2. Data and Methodology

a. Event Descriptions

Three cases from the set of all MPEX research flights, based on their large-scale flow regime so as to quantify the flow-dependence of PBL parameterization control upon the predictability of CI, are considered: 19 May, 31 May, and 8 June 2013. These cases comprise the fourth, tenth, and twelfth research flights, respectively, and are hereafter referred to as RF4, RF10, and RF12. RF4 was characterized by a deep mid-level trough over the Intermountain West of the US, resulting in southwesterly flow into the Central Plains states. The sensitivity of ~36 h numerical forecasts to the ~24 h forecast fields in these locations (as a proxy for subsequent initial conditions) promoted a targeted area for additional observations over Colorado, eastern Utah, northern New Mexico, Texas and Oklahoma panhandles, and far Western Kansas. Using this sensitivity analysis, the additional observations gathered via dropsondes within MPEX were done following the flight path seen in Figure 1a With a surface low approaching off the Colorado Front Range, this event produced widespread convection across Oklahoma, Kansas, and Nebraska with a dryline triggering powerful discrete supercells in central Oklahoma producing multiple tornadoes including the EF4 (on the Enhanced Fujita scale) rated Norman-Shawnee, OK tornado and a cold front initiating a squall line throughout eastern Kansas. Figure 1b depicts the shallow trough and mid-level low over the Dakotas for RF10 as well as the westerly winds aloft

pushing into the central plains. Similarly, Figure 1b highlights the resultant flight track for RF10 acquiring additional observations in Colorado, New Mexico, eastern Texas and Oklahoma, and southwestern Kansas. This event was equally as damaging as RF4 as a stalled front was draped from southwest to northeast through central Oklahoma with a dryline mixing east toward Oklahoma City. CI was triggered all along the stalled front with discrete supercells forming along the dryline near the triple point (just west of Oklahoma City) producing the infamous 2.6 mile wide long track El Reno, OK EF3 tornado. Last, RF12 shows the ridging in the Pacific Northwest which results in weak northwesterly mid-level flow into the plains states (Figure 1c) and the additional observations recorded in Wyoming, northern Colorado, and northeastern Utah. RF12 was less organized than RF4 and RF10, in that it produced less coherent convective modes and overall weaker storms. With a surface low tracking to the northeast through Nebraska, a trailing cold front produced a line of weaker convection extending from eastern Nebraska to the southwest into the Texas panhandle with hail and wind being the significant danger as no tornadoes were reported.

b. Model Configuration

Version 3.4.1 of the Advanced Research WRF (WRF-ARW; Skamarock et al. 2008) model is used to simulate the three events described in the previous sub-section. The forecast model configurations largely follow those of Romine et al. (2014), Schwartz et al. (2014), and Torn and Romine (2015), where a two-way nested domain configuration (Figure 2) is used, with horizontal grid spacing Δx = 15 km (3 km) on the outer (inner) domain and both domains having 40 vertical levels with the model top set at 50-hPa. Physical parameterization packages used for both domains include the Thompson et al. (2008) bulk microphysical parameterization, Rapid Radiative Transfer Model (RRTM; Mlawer et al. 1997) longwave and the Goddard (Chou and

Suarez 1994) shortwave radiation parameterizations, and Noah land surface model (Chen and Dudhia, 2001; Ek et al. 2003). The Tiedtke (1989) cumulus parameterization scheme is used in the outermost domain, while convection is explicitly resolved on the inner domain. Only the PBL scheme changes between ensembles, as described in Section 2c below.

To conduct the simulations, an ensemble framework comprised of thirty members initialized with slightly different initial states of the atmosphere is used. To obtain the initial conditions (ICs) we utilize the method as proposed by the Schwartz et al. (2015) and the references within. This method differs from previous works (e.g., Done et al. 2004; Kain et al. 2006, 2008; Weisman et al. 2008) which obtained ICs by downscaling analyses from larger-domain, coarser resolution NWP models for deterministic convection-allowing simulations and also differs from similar methods typically used when initializing convection-allowing ensembles with larger and coarser ensembles. The method of Schwartz et al. (2015) describes the process of generating ICs for limited area models via continuously cycling data assimilation (DA) where the NWP model state is routinely updated with real observations. Additionally, the use of an Ensemble Kalman Filter (EnKF; Evensen 1994; Burgers et al. 1998; Houtekamer and Mitchell 1998) in this method allows for the integration of ensemble DA and forecasting as the EnKF system incorporates time-evolving forecast error characteristics and produces ensemble analysis states which can be used as ensemble initialization. Previously, studies by Schwartz et al. (2014) and Romine et al. (2014) examined month long periods while using EnKF analysis/forecast systems with grid resolutions of Δx = 15-36-km to initialize convection-allowing ensembles forecasts and showed that mesoscale EnKFs can initialize skillful forecasts while using convection-allowing ensemble. Additionally, Schumacher and Clark (2014) found

that rainfall forecasts were less biased and tended to be more skillful when initialized with EnKF analyses rather than using ICs derived from coarser, external models.

The ICs as produced for this study are generated by using a 50-member ensemble adjustment Kalman filter (EAKF; Anderson 2001, 2003) from the Data Assimilation Research Testbed (DART; Anderson et al. 2009) software This EAKF system was run at Δx = 15-km grid resolution and assimilated many upper-air and surface observations including, but not limited to, radiosonde observations of temperature, wind, surface pressure, and specific humidity, aircraft reports of temperature and wind, infrared and water vapor channel satellite-tracked wind, global positioning system radio occultation (GPSRO) refractivity, wind profiler observations, surface observations from ship, buoy, surface synoptic observation (SYNOP), and routine METAR data from aviation weather reports. The assimilation process used prior adaptive inflation before the assimilation of the observations to preserve ensemble spread and also utilized vertical and horizontal localization to reduce spurious correlations resulting from sampling error. Readers are referred to Romine et al. (2013, 2014) for additional information regarding the EAKF process. These ensemble initial conditions are obtained by randomly perturbing a first-guess atmospheric analysis from global background error covariances provided by the WRF DA (Barker et al. 2012) system. Both the perturbed lateral boundary conditions (LBCs) for all assimilation cycles and ICs for the original ensemble initialization were then obtained for the EnKF system by interpolating the Global Forecast System (GFS) to a Δx = 15-km grid WRF Model domain and adding to it the random covariance perturbation method of Torn et al. (2006). The National Center for Atmospheric Research (NCAR), for what Schwartz et al. (2015) estimates as the first time, utilized a continuously cycled EnKF DA system to generate the ICs for 48-h ensemble forecasts at convection-allowing (Δx = 3-km) grid resolution to support the MPEX field

campaign. The perturbed LBCs and ICs were first generated 18 UTC on 30 April 2013 and served as the IC's for a 6-h WRF forecast. Then the prior ensemble estimate at 00 UTC on 1 May 2013 became the first EAKF analysis, and after the assimilation of additional observations became the IC's for the following 6-h forecast, providing a second EAKF analysis at 06 UTC on 1 May 2013. This cyclic pattern continued through 12 UTC 16 June 2013. The 48-h ensemble forecasts were initialized twice daily at 00 and 12 UTC between 15 May and 15 June of 2013 and were ran with the WRF-ARW model. Additionally, on the days in which an MPEX IOP occurred, a subsequent 0000 UTC EnKF simulation was launched from the ongoing EnKF analyses with the 6-h cycling interval shortened to 1-h and had MPEX dropsondes assimilated over the period 0900 to 1500 UTC. The ICs used for our three selected events were then the final EnKF analyses at 15 UTC on the day of the event produced by the 0000 UTC EnKF simulations with 1-h cycling intervals which contained all real observations plus the MPEX targeted observations. These analyses are used as the LBCs and ICs for our 15-km outer domain.

c. PBL Schemes

Five PBL schemes are utilized to quantify their influence upon the predictability of CI. Similar to Coniglio et al. (2013), Clark et al. (2015a), and Cohen et al. (2015), the five schemes selected are the MYJ (Janic 1994), YSU (Hong and Pan 1996), QNSE (Sukoriansky et al. 2005), ACM2 (Pleim 2007), and the MYNN (Nakanishi and Niino 2009) schemes. Table 1 summarizes these selections and their associated surface layer parameterizations. These five schemes can be categorized into two main groups based on the closure used to obtain turbulent fluxes from resolved-scale atmospheric quantities: local and non-local. The local closure schemes (MYJ, MYNN, and QNSE) estimate the turbulent fluxes and unknown atmospheric variables at every grid point using only adjacent vertical levels (Stensrud 2007). The three local schemes used here

variations of the MYJ level-2.5 scheme (Mellor and Yamada 1982) and compute TKE by diagnosing the variance and covariance of potential temperature variance and water vapor mixing ratio (Coniglio et al. 2013) and use the TKE predictions to estimate the height of the PBL (Stensrud 2007). The non-local closure schemes (ACM2, YSU) estimate the turbulent fluxes and unknown atmospheric variables at every grid point using multiple vertical levels in a column and attempt to simulate the effects of eddies in the convective PBL (Stensrud 2007). Non-local schemes do not estimate TKE, but rather the height of the PBL is estimated using empirical formulas based on wind speed, vertical gradients of virtual potential temperature, and the critical Bulk Richardson number (Stensrud 2007).

d. CI Algorithm/Identification

For the three cases described in section 2a, we begin by seeking to identify CI within both observed data as well as our ensemble forecast. To look at the observed data, we start with obtaining the observed Next Generation Radar (NEXRAD) level II data from the National Centers for Environmental Information online archive (http://www.ncdc.noaa.gov). Data are obtained every five minutes beginning at 1500 UTC on the day of interest and ending on 0600 UTC the following day. This was done for 42 National Weather Service (NWS) Weather Surveillance Radars, 1988 Doppler (WSR-88D) across the Central Plains of the United States (Figure 3). To assess CI in general we use radar reflectivity at the -10ºC level (e.g. Kain et al. 2008). We adopt a method to define CI which utilizes radar reflectivity $\geq$ 35 dBz at the -10ºC level to remove the effects of bright banding that may artificially amplify the signal of DMC. This method has been informed by the findings of Gremillion and Orville (1999) which has found high probability of detection (POD) and critical success index (CSI) for matching radar observations to observed thunderstorm occurrences. We further require CI exist for at least 30

minutes (Kain et al. 2013; Burghardt et al. 2014) to target CI with an object based approach. To identify CI events in the observed data, the Warning Decision Support System – Integrated Information (WDSS-II; Lakshmanan et al. 2007) spatial analysis tools (Lakshmanan 2012) are used to convert the NEXRAD data files to a WDSS-II readable form, generate a uniform 0.03º x 0.03º gridded domain, interpolate the radar data to the -10ºC isotherm extracted from the 00-h analyses of the Rapid Update Cycle (RUC; Benjamin et al. 2004) model, and lastly merge (w2merger from WDSS-II) all data from the 42 radar locations onto the generated uniform grid (Lakshmanan et al. 2006; Lakshmanan and Humphrey 2013). With this data set we then, similar to Kain et al. (2013) and Burghardt et al. (2014), sought to identify and track CI objects by the aforementioned definitions. These thresholds ensured that CI had occurred and was sustained, eliminating the presence of any pulse or otherwise unsustained deep, moist convection that could falsely represent CI. Using the WDSS-II w2segmotionll tool, individual storm cells are then identified (Lakshmanan et al. 2009) and tracked forward in time (Lakshmanan and Smith 2010). Figure 4 shows an example snapshot of the CI object identification. The output is then used to extract motion estimates and attribute it to each CI object (Lakshmanan and Smith 2010; Lakshmanan et al. 2013). From which we were able to examine the start/end time, size, motion (speed and direction), and location of every CI object which met our criterion, then using this data we specified CI as occurring at the location in which the criterion of Gremillion and Orville (1999) was met for 30 minutes. CI objects are also identified from the ensemble numerical forecasts. Simulated reflectivity interpolated to the -10ºC isotherm is computed at each model time step and stored at five minute intervals. These data are interpolated to the same uniform 0.03º x 0.03º gridded domain, ingested by WDSS-II, and CI objects are then tracked in the same manner as the observed described above. This process allowed us to compare our modeled and

observed CI events using a uniform temporal (5-min) and spatial (0.03º x 0.03º grid) element as well as a consistent identification/tracking algorithm. In this study however we do not attempt to sample the model data in a similar fashion to that of the observations, i.e. we do not account for radar gaps, terrain blocking, or non-uniform data distribution in the vertical over the domain.

e. CI Verification

To compute the verification presented in the results, the technique of Burghardt et al. (2014) is used by sorting the results produced by using equation 1; a spatiotemporal flow-dependent error metric which accounts for spatial and temporal distances (error) between CI objects:

$$C^2 = (E_d)^2 + (VxE_t)^2 \quad (1)$$

The error metric *C* is represented in units of kilometers with E_d representing the absolute error of distance in km and E_t representing the absolute error of time in seconds between the observed and modeled CI objects. *V* corresponds to the estimated average velocity (converted to $km\ s^{-1}$) of the observed CI objects life as determined by the w2segmotionll function within WDSS-II. The velocity component used here differs slightly from that of Burghardt et al. (2014), who used an hourly, area- and layer-averaged middle tropospheric horizontal wind to approximate the deep-layer steering flow. Each ensemble member was verified against observations deterministically given the inherent complexities associated with verifying probabilistic forecasts of locally rare events such as CI (e.g., Lock and Houston 2014). The verification was then expanded to evaluate the aggregated ensemble as a function of its individual members. Further, to limit the spatial and temporal bounds of verification, E_d is constrained to be ≤ 100-km and E_t is constrained to be ≤ 3600-s. These constraints are consistent with Duda and Gallus (2013), who found a mean absolute displacement error of 105-km and no systematic temporal error, and

Burghardt et al. (2014) who limited by the same constraints to represent a balance in time and space errors, and to confine the motion of CI objects to 100 $km\ h^{-1}$. Additionally, we also conduct the verification at various other thresholds to find bounds by which the forecast of CI is limited in terms of the chosen spatial and temporal errors. This verification is done over the entire 0.03º x 0.03º gridded domain used when identifying our CI objects. The extent of this domain stretches from -111º to -93º W and 23º to 50º N as is represented in Figure 3 as the black box around the used NWS radar sites (hereafter referred to the verification domain).

We acknowledge that there exists other error metrics that can be used in place of the one above. For example Skinner et al. (2016) suggests a metric referred to as the interest calculation which eliminates the dependence on the magnitude of the flow. This metric determines a maximum temporal and spatial error that confines a modeled object to a "useful" range of skill such that a hit would still provide important information to a forecaster or end user. Preliminary investigation into the interest factor as comparison against the flow-dependent error metric used in this study yielded small differences between the observed and modeled CI matching, thus we feel there would be no significant change in forecast skill related to the selected error metric however the sensitivity to the chosen metric could fall into future work.

3. Results

a. Model Verification

For the three cases analyzed, we began by looking at the characteristics of the CI events in an aggregate sense by examining the full domain of analysis (as seen in Figure 3) for the full 15 hours of simulation time (15 UTC to 06 UTC the following day). In total there were 238, 75, and 180 observed CI events for RF4, RF10, and RF12 respectively. Our modeled CI counts

exceeded the number of observed events in each case, often exceeding ratios of 2:1 or greater. For RF4, RF10, and RF12 the average number of modeled CI events produced by each ensemble member, ignoring the differing PBL scheme, was respectively 399.88, 117.84, and 325.67 events. To highlight this, the number of CI events per case aggregated over time in a cumulative frequency distribution throughout the entire domain are plotted in Figure 5. This result is similar to Burghardt et al. (2014), who in verifying the occurrence of CI in convection-allowing high resolution deterministic NWP simulations found 641 modeled CI events as compared to 249 observed events with observed to modeled ratios of nearly 4:1 at times.

This overproduction in a temporal sense can be seen in Figure 5 which highlights the cumulative frequency distribution through time. Most notably, early in the simulations around 16-17 UTC the slope of the modeled distributions is steeper than the observed line and this continues until about 01 UTC when the distributions level off. The overproduction can also be seen spatially in Figures 6, 7, and 8 for cases RF4, RF10, and RF12 respectively which show a zoomed-in domain that focuses on the primary regions in which CI occurred. For RF4 and RF10, there exists a significant overproduction of CI in eastern Colorado, northwestern Kansas, and throughout Nebraska with most these overproduced CI events falling within the 16-21 UTC time frame. RF10 was a much more localized event and it was forecast better in terms of the areal coverage of CI from the standpoint of the shown representative member, but still overproduced CI in the Texas panhandle, northwest Missouri, and northern Iowa. The sensitivity to our adopted CI definition was tested with both the modeled and observed data relative to the minimum reflectivity threshold as well as the length of time required for CI to determine if the CI overproduction was a result of our definition. For these sensitivity tests we extended our definition to be more inclusive by testing combinations of 35, 30, and 20 for the minimum dBz

threshold and 30 and 15 minutes for the minimum temporal requirement. Similarly, to make the definition more restrictive we increased the minimum dBz and time thresholds to 40 and 45 dBz and 45 and 60 minutes and tested combinations within. It was found that the overall number of CI objects would increase (decrease) with lower (higher) dBz thresholds and shorter (longer) time restraints, but in general for each ensemble and for each cases the ratio of CI events to that of the observed remained relatively consistent, thus we feel confident our results would be consistent across various CI definitions.

One could speculate that this overproduction results from the PBL schemes tendencies to over-predict mixed-layer CAPE (MLCAPE) and under-predict mixed-layer convection-inhibition (MLCIN) during the daytime in pre-convective environments (Coniglio et al. 2013). Coniglio et al. (2013) showed that on average the PBL schemes over-predicted the MLCAPE by 500-600 $J\ kg^{-1}$ with over-predictions of 700-800 $J\ kg^{-1}$ 25% of the time with the non-local MYJ and QNSE PBL schemes having the largest mean errors relative to the ACM2, MYNN, and YSU. With increased MLCAPE and decreased MLCIN, the modeled environments are more primed to initiate convection than the observed environment and this likely plays a role in the CI overproduction seen here. While this speaks to the possible reasons for CI overproduction relative to thermodynamic field, there is an additional source of possible error related to the kinematic fields, i.e. the degree of forcing as manifest as the CI triggering mechanism. The degree of forcing will vary between the observed environment and the modeled environment but we lack a high resolution mesonet of observed data in the immediate area of CI to quantitatively show to what effect. However, drawing from Ching et al. (2014) we are aware that a major problem of modeling flows at small [Δx O(1-km)] grid spacings is that these scales still too large to fully resolve the turbulent transfer of energy as they fall within the "terra incognita" of

Wyngaard (2004). Consequently these scales may also be too small and not suitable for the available PBL schemes to accurately parameterize the turbulent energy, and in fact Ching et al. (2014) highlights that for some PBL schemes on these higher resolution grids they often over-simulate the amount of TKE in the PBL.

After applying the spatiotemporal threshold (equation 1) to the modeled CI events in each ensemble member and sorting based on the lowest value of the error metric *C*, we placed the results into a 2x2 contingency table (Wilks 1995; Fowle and Roebber 2003; Kain et al. 2013) as described in Table 2. The best match for each observed event, or the match with the lowest *C*-value for each member, were fed in as true-positives or "hits" (a). The remainder of the table was filled with false-positives (B) which occur where the model overproduces observed events, false-negatives or "misses" (C) occur where there existed an observed event and no forecasted event or the forecasted event was outside the spatiotemporal threshold, and true-negatives (D) which occur when neither an observed or forecasted event are present. Given our method of verification and the skill metrics we examine, we ignore section (D) of the contingency table since we only aim to verify the correct forecast of CI and thus ignore the correct "non" forecast, or "true negative," since an event in and of itself is a positive occurrence. For cases RF4 and RF12, the 30 member ensemble using the MYJ PBL scheme (hereafter MYJ; similarly for ACM2, MYNN, QNSE, and YSU) had the most hits and the fewest misses, but also had the most false alarms with QNSE second in those categories for RF12 as shown in Table 3. Similarly for RF10, MYJ continued to have the most false alarms but was second in hits to the QNSE run by 0.1 hits. This result was to be expected given the largest overproduction of CI events seen by the local mixing MYJ and QNSE ensembles (refer to Figure 5). The ACM2 simulations had the fewest false alarms along with the smallest standard deviation amongst the ensemble for all three cases.

Consequently ACM2 also had the fewest number of hits and the most misses for all events speaking to the "duality of error" as noted by Doswell (2004) which describes how increasing the probability of detection (POD) for a fixed forecast accuracy results with an increase in false alarm ratio (FAR) as false negatives become false positives.

To examine the results of Table 3 we compute common verification metrics deterministically for each ensemble member and then average them for the ensemble mean. The verification metrics are all computed using the aforementioned contingency table and are presented in equations 2-5, with the letters A, B, and C corresponding to the section of the contingency table (See Table 2). The metrics computed are the POD [Eq. (2)], which highlights the ratio of correctly forecasted objects to the total number of observed CI objects; FAR [Eq. (3)], which looks at the ratio of overproduced forecasted CI objects to the total number of CI objects forecasted; Bias, which examines the ratio of the modeled CI forecasts (both hits and misses) to the number of observed CI objects [Eq. (4)]; and lastly the critical success index [CSI; Eq. (5)], or threat score [TS; Wilks (1995)], which is the ratio of correctly forecasted objects (hits) to the total number of observed and forecast CI objects.

$$\frac{A}{A+C} = POD \tag{2}$$

$$\frac{B}{B+c} = FAR \tag{3}$$

$$\frac{A+B}{A+C} = bias \tag{4}$$

$$\frac{A}{A+B+c} = CSI \tag{5}$$

POD, FAR, and CSI range between 0 and 1, with 1 being the best possible score for POD and CSI and 0 being the best for FAR. Bias begins at 0 and is not bounded by an upper limit but values greater (less) than 1 imply, in this sense, an overproduction (underproduction) of CI.

Figure 9 depicts the results of the above calculations plotted on the performance diagram of Roebber (2009) which allows for the visualization of multiple forecast quality measurements.. Noting that the setup of the diagram results in better-performing forecasts (i.e. High POD and SR and also a bias near 1) to trend toward the upper right corner where all values approach 1. RF4 had the highest forecast skill in three categories with a mean POD of 0.686, mean SR of 0.431, and mean CSI of 0.359 but was more biased than RF10 with 1.604. RF12 had the lowest mean values of POD, SR, and CSI with 0.547, 0.326, and 0.255 respectively and the highest mean bias of 1.697. RF10 fell between the other 2 cases with mean values of POD, SR, and CSI equal to 0.630, 0.428, and 0.337 respectively and was the least biased by 0.057 events relative to RF4. The positive bias (bias greater than 1), as seen in all three cases, is a product of the CI overproduction previously mentioned and as noted in Table 3. The bias is largest for the MYJ ensembles and smallest for ACM2 in all three cases. The variability of skill for ensemble means between cases is relatively small overall and holds true between ensemble members for RF4 and RF12. RF10 however, has significantly more variability between the ensemble means and the subsequent ensemble members as seen in Figure 9. This could be an artifact of the convection occurring into the evening hours of RF10 as compared to RF4 and RF12. As Coniglio et al. (2013) highlighted, MLCIN in the evening hours is under predicted when strong capping inversions are present which is the case for most areas in RF10 (shown later in section 3b). Similarly they point out that evening MLCAPE tends to be over predicted by MYJ and QNSE schemes when the observed MLCAPE is relatively small ($\leq 1500\, J\, kg^{-1}$) and is under predicted by ACM2, MYNN, and YSU when the observed MLCAPE is large ($\geq 1500\, J\, kg^{-1}$). Additionally, RF10 saw a majority of the CI result from a dryline in Oklahoma, which from Clark et al. (2015a) we know there exist biases in the forecasted location of the dryline between

each PBL scheme. This is important because the variability seen between ensemble members and their means could be related to forecasted dryline positioning. Knowing that the strength of capping inversions increases eastward from the dryline since the vertical circulations induced by the moisture gradient of the dryline are localized to the dryline position, differences in the location of the dryline as well as the PBL depth can result in very large moisture differences. Where some members/ensembles are underneath a stronger stable cap and others are within the well-mixed PBL; i.e. small positional errors in each member's forecast can result in large differences in the environment where CI is trying to occur which could directly translate in varying measures of forecast skill. Later in section 3b (Figure 16), we also show that RF10 had the largest mean error (ME) relative to 10-m wind speed (WS) and 2-m relative humidity (RH) to the forecast of the diurnal cycle. These points suggest that the variability within the ensembles themselves and the ensemble means between cases is related to different biases that are present within the PBL schemes during the day and at night as well as with forecasting dryline positioning; i.e. CI during the evening or CI resulting from dryline convergence may be inherently less predictable.

Figure 10 shows violin plots of the aggregated ensemble distribution of CSI and the associated probability density function for each case and the ensembles within. For all cases the PDF of the data is approximately normal, with RF10 having a larger spread between ensemble member solutions. RF4 and RF10, as described in section 2a, were events with more well-organized large scale forcing and had convection initiating along more well-defined triggering mechanisms and from an instability standpoint had more CAPE present. Thus, from a NWP sense it is assumed that there would be by some measure better skill in forecasting these more well defined features in RF4 and RF10. Additionally, work done by Van Klooster and Roebber

(2009) showed that increasing the CAPE regardless of the amount of vertical wind shear increases the probability of convection occurring. Their Figure 1 further suggests that variations in the stability fields within an ensemble on more unstable days would have less impact than those same variations on more stable days. For example a 500 $J\ kg^{-1}$ error in the CAPE field, which is reasonable given Coniglio et al. (2013), would have much more of an impact on the probability of convection occurring if the actual CAPE field was measured at 1500 $J\ kg^{-1}$ than if it were measured to be 4000 $J\ kg^{-1}$, this assuming the wind shear was the same in both examples. Given the sizeable values of instability in RF4 and RF10 relative to RF12, we add to the above hypothesis in that it is expected that the more unstable environments should be more predictable. That hypothesis is verified in Figure 9 with the higher POD and SR values and further verified in Figure 10 with cases RF4 and RF10 having higher overall mean CSI scores of 0.359 and 0.337 relative to 0.255 for RF12. While our hypotheses are verified for these situations, we understand that given the sample size of three our results are conclusive and require additional analyzed cases.

To again attempt to understand the aforementioned CI overproduction, we visualize the temporal distribution of the misses and false alarms resulting from the 2x2 contingency table in Figure 11. Given the cumulative distributions seen in Figure 5 that show the overproduction of CI occurring most notably between 16-21 UTC, it is fitting to see the false alarms mostly occurring in that same time frame and tapering off through time. RF4 and RF12 similarly produce most of the false alarms early then as these decrease an increase in the misses becomes apparent as the ratio of modeled CI objects to observed approaches 1 after 22 UTC. The structure of the misses, CI observed and no forecasted object, is different in all three cases but there are some shared characteristics. For example the misses are much less frequent in the early hours,

which is again to be expected given the overproduction and false alarms, and increases toward a peak around 00 UTC. This is most notable in RF4, but the overall pattern also be seen in RF12 and slightly in RF10 which also has many more misses towards the end of the simulation resulting from initiation at that time that the other 2 cases did not see. Most notably in Figure 11, the false alarms, similar to shape of the cumulative distributions in Figure 5, show that there appears to be some connection to the diurnal cycle which will be examined more closely later in this study.

Further examination of the misses and false alarms in the early hours led us to analyze the CAPE field in which these events were occurring. Utilizing the 1800-2100 UTC 00-h forecast of surface based CAPE (SBCAPE) from the 20-km RAP analysis valid on the respective day of each case as the observed CAPE field, we found that the areas of daytime CI overproduction early in our forecasts (also previously shown in Figures 6-8) occur in regions of less than 2000 $J\ kg^{-1}$ (not shown). This leads us to the idea of looking at the modeled CAPE and sounding structures in these areas which is done later in section 3b. 2000 $J\ kg^{-1}$ was subjectively selected as a comparison for the SBCAPE field based on Figure 1 from Van Klooster and Roebber (2009) which shows that increasing SBCAPE increases the probability of convection being realized. They show here that even with minimal shear there is roughly a 75% or greater probability of convection occurrence when surface convergence is present (all three cases had surface based convection in some form).

To further understand the forecast metrics shown in Figures 9 and 10, the "hits," or true-positives (also seen in Table 3) are analyzed further. Figures 12 and 13 examine the temporal and spatial difference respectively of the observed event and the modeled event matched to it. Interestingly, RF4 and RF10 have near normal temporal distributions centered around t=0, where

RF12 has a slightly skewed distribution with more negative temporal differences (indicating modeled objects occurred too late) but still has resemblance of a normal distribution. This result in itself is intriguing and is similar to Kain et al. (2013) who found a similar shaped distribution in matching CI observed and modeled events with a temporal threshold of ± 5-h in 15-24-h forecasts (their Figure 4a and 4b). Similarly these results suggest that when CI occurred in both observations and models there was little systematic bias in CI timing for the 5 PBL ensembles at least for the more organized RF4 and RF10, and the difference between PBL schemes in each case here is negligible. Albeit the variance in timing was large the distribution of temporal differences likely provides, as suggested by Kain et al. (2013), a glimpse into the true temporal uncertainty related to CI which may prove useful in the development of probabilistic guidance for forecasting the timing of CI.

In Figure 13, analyzing the spatial error of the "hits," we see that the three cases have mean spatial errors of 43.789, 42.629, and 47.495 -km and 55.34, 55.45, 53.45 % probabilities that the spatial error will be less than the mean respectively for RF4, RF10, and RF12. The three cases, including the varying PBL ensembles within, all have similar distributions case by case which are skewed towards the higher spatial difference with RF12 having the least amount of skew. The mode of the distributions occur between 24 and 31 -km for all five ensembles in all three cases with overall mean modes of 26.865, 29.873, and 26.632 -km creating probabilities of 29.55, 34.75, and 25.52 % that the spatial difference will be less than the mode for RF4, RF10, and RF12 respectively. The skewness of these spatial PDFs and the normality of the temporal PDFs suggest that the matching of a CI event is approaching a limit as it nears 100-km and 60-min since outside of those ranges the data are in the tails of the distributions. This is further analyzed by exploring the sensitivity to the spatiotemporal threshold. Here we explore five

additional thresholds, 15-min/25-km, 30-min/50-km, 90-min/150-km, 120-min/200-km, and 180-min/300-km. Across all thresholds we see the shape of the distributions remain consistent (not shown) and as the thresholds become more inclusive beyond 60-min/100-km there is still an observed "drop-off" in the amount of data contained outside of our accepted threshold. To visualize this, in Figure 14 we compute the cumulative distribution function (CDF) of CI hits for the aggregated ensemble members in both time and space out to 180-min/300 km. These graphics show that at the smallest thresholds the spatial characteristics impact the skill more since there is significantly less data confined to the smallest bins then there is in the temporal bins. From 0 to the 60-min/100-km there is a steep increase in the area under the CDF and after which the slope levels off and increases steadily through 180-min/300-km.

For RF4, RF10, and RF12 respectively, there is 75.366%, 78.453%, and 66.692% of the hits occurring prior to ± 60-min and 83.490%, 87.319%, and 74.910% within 100-km. This is important, the data here are suggesting that increasing the spatial threshold beyond 100-km and/or the temporal threshold beyond 60-min will have impact on the skill of the forecast, but it will not be a significant improvement. The temporal aspect, by extending it to 90-min, would change the CDF so that it becomes similar to the structure of the spatial CDF in that both the CDFs would flatten out right at our chosen thresholds. This is confirmed by again computing the area under the CDF with 86.7%, 89.1%, and 81.0% of the hits occurring within 90-mins for RF4, RF10, and RF12 respectively. The amount of hits occurring within each threshold would then be comparable, however, to remain consistent with Burghardt et al. (2014), we continue to conduct our verification using the 60-min/100-km verification threshold but also consider further verification using the 90-min/100-km threshold. Figure 15 highlights the skill of all the previously mentioned thresholds (15-min/25-km, 30-min/50-km, 60-min/100-km, 90-min/100-

km, 90-min/150-km, 120-min/200-km, and 180-min/300-km) on the performance diagram of Roebber (2009). This plot shows the drastic jump in skill from the 2 most lenient thresholds to the 60-min/100-km, and then another smaller increases from 60-min/100-km to 180-min/300-km. This reiterates what was highlighted in the CDFs of Figure 14, by extending out to 60-min/100-km, there is information available which improves the overall skill of the forecast. However, beyond both 60-min and 100-km, the amount of useful information is approaching its limitation, and there is minimal skill added to a forecast. To elaborate, while increasing the thresholds from 60-min/100-km to 180-min/300-km does increase the CSI for each ensemble within each case, subjectively the amount gained skill does not translate to a better forecast since a 180-min and/or 300-km forecast error for a CI event is likely outside the realm of usefulness. This analysis confirms that the CI threshold of 60-min/100-km used both here and in Burghardt et al. (2014) is viable in maximizing the skill of forecast while still utilizing restrictive thresholds. It also suggests there are bounds that can be placed on the forecast verification of CI, with those bounds similar to, or a variant of, the analyzed 60-min/100-km or 90-min/100-km within this study. It is also hypothesized that the skill of future studies will likely be sensitive to the large scale flow, geography, and speed of convective systems in which the verification is conducted. However, this study seeks to quantitatively highlight thresholds that similar future work verifying the forecast of CI in an object based approach could apply and provide justification for the selection of those thresholds.

Relative to the PBL scheme on the forecast of CI, it is seen that there exists no major differences in the overall skill of the forecast between ensemble means. The more restrictive the threshold, the less the discrepancy between CSI of the ensemble means. As the thresholds become more lenient there is a more discernable difference but it is related to the bias, or CI

overproduction; i.e. The MYJ (ACM2) ensemble has the highest (lowest) bias, so with this PBL scheme you are more likely to have a higher (lower) POD but will also have a higher (lower) FAR, again speaking to the "duality of error" as noted by Doswell (2004). As previously mentioned the ACM2 ensemble in all three events was the least biased and as Figures 9 and 15 show, increasing the thresholds does not change the bias since it is the ratio of forecasted to observed objects. Thus it can be concluded that the ACM2 PBL scheme provides the highest limit of forecast skill related to CSI, FAR, and SR as the thresholds applied increase since one can linearly follow the bias line until it reaches a POD or SR of 1. This suggests that although the chosen PBL scheme does have slight effect on the overall forecast skill of CI as seen in the performance diagrams, there appears to be minimal effect on the temporal and spatial biases of the modeled CI objects that are matched to the observed.

b. Sensitivity of PBL Scheme on CI Environment

Considering the verification in the previous subsection, we now seek to highlight the differences in the near-surface as well as vertical structures of the environment between PBL schemes to showcase the sensitivities of a selected scheme on a given forecast. We begin by examining 2-m temperature (Tc), dew point temperature (Td), and RH along with the 10-m wind speed from each of the five PBL ensembles for the three selected cases. We use the National Center for Environmental Prediction (NCEP) Meteorological Assimilation Data Ingest System (MADIS) database to access Meteorological Aviation Report (METAR) observations gathered at the surface over the CONUS. Taking the 1-h MADIS METAR data we extract the observations from over our verification domain and compute the ME from each of the ensemble means. The mean error is computed hourly throughout the 15-h simulations with on average just over 2,000

observations being compared per hour per case. Highlighted in Figure 16 we see throughout time the ME for Tc, Td, RH, and WS computed using the following equation:

$$ME = \frac{1}{N}\sum_{i=1}^{N}(x_i - o_i) \quad (6)$$

Where in equation (6) o_i is in reference to each observation per hour per case, x_i corresponds to the ensemble mean, and o_i represents the MADIS metar observation. This domain average ME highlights any consistent biases that exist from any one PBL scheme as well as any temporal biases that may exist. Examining the ME of Tc, Td, and RH, we immediately notice a cool and moist bias throughout time in RF4 and RF10 and to a lesser extent RF12. The non-local mixing PBL schemes ACM2 and YSU are the least cool and moist biased and are even warm and dry biased in the last 6 to 8-h of RF12. In contrast, the local mixing PBL schemes MYJ and QNSE have the most pronounced cool and moist biases with the "hybrid" MYNN PBL scheme falling in the middle. These findings support previous studies by Hu et al. (2010), Coniglio et al. (2013), and Clark et al. (2015a) who all found tendencies for the non-local schemes to produce PBLs that are too deep and dry and vice versa for the local schemes. While these ME plots do not discuss the vertical structure of the PBL, they do highlight similar biases at the surface in that although the domain averaged ME shows cool and moist biases, the local (non-local) schemes are the most (least) cool and moist biased. Overall, the ACM2 scheme is the least biased relative to the MADIS observations where the MYJ is the most. There are also apparent biases with the IC analyses used to initialize the 30 member ensembles as all the simulations are initialized too cool and moist. This could likely be a result of the MYJ PBL scheme being used in the generation of the ICs, since it has previously been documented to have a cool and moist bias. This is relevant to the CI overproduction, with a cool moist bias at the surface the modeled environment has lower lifting condensation levels (LCL) than the observed which could entice

parcels to more easily become positively buoyant and initiate DMC. Another interesting characteristic observed here is the error associated with the diurnal cycle of the PBL. This is most notable in RF10, to a lesser extent RF4, which sees a sharp decrease in ME of Td and WS around 01 UTC suggesting an issue with forecasting the daytime to nighttime transition of the PBL.

Next we aim to highlight the vertical structure of the PBL and the environment prior to convection to illuminate biases and differences between PBL schemes and the possible effects they could have on CI forecasts. To do this vertical profiles of ME are generated by computing the difference between modeled ensemble mean soundings and observed soundings. The observed soundings are a combination NWS additional 18 UTC and the standard 00 UTC soundings at select sites along with soundings from MPEX mobile units. The MPEX mobile soundings were launched by participants in the field program and sought to capture vertical profiles in the pre- and post- convective environments. The selected soundings were manually examined to ensure that only soundings occurring in a "pristine," or convectively undisturbed environment were used in this analysis. On average the ME was computed from roughly 20 observed soundings for each case, while the sample size is not as large as we would like we do feel that what is presented is an accurate representation of the biases within the PBL and vertical structure given similarities to previous works (e.g. Hu et al. 2010, Coniglio et al. 2013; Clark et al. 2015a; and Cohen et al 2015). The soundings used spanned from as early as 16 UTC to as late as 03 UTC on the following day for each event. To provide some context to the location of soundings used, Figure 17 displays the location of all soundings used within each case atop of the accumulated 3 to 9-h (18 to 00 UTC) precipitation from the 20-km 15 UTC RAP forecast for each event. Figure 18 then computes the ME for Tc, Td, RH, and WS at every 25-hPa using

equation (6) again. The vertical ME profiles show some variability case to case which is to be expected given the rather large differences in flow patterns and synoptic set-ups between the three cases. Within this variability however there are some reoccurring artifacts that arise from the chosen PBL scheme. For example in Figure 18, in the ME of Tc, Td, and RH we observed that the local-mixing MYJ and QNSE are always cool and moist biased relative to the other PBL schemes up to 850-hPa and vice versa above 850-hPa. This is again consistent with what we expect to see given previous work and with our understanding of local vs non-local mixing PBL schemes. The more vigorous mixing associated with the non-local ACM2 and YSU results in a drier PBL that is often too deep given the mixing of the dry air from above the PBL and the opposite is true for local MYJ and QNSE with the MYNN often being between the two mixing groups. Which again these findings remain consistent with the results from Coniglio et al. (2013) and Clark et al. (2015a).

In addition to examining the ME resulting from the sounding comparisons, we looked closely at each individual sounding which went into the calculations. Another artifact that was realized was that all PBL schemes within each ensemble struggled to resolve moderate to strong capping inversions. This point was briefly discussed in Coniglio et al. (2013) and shown in their Figure 15 as they had seen a similar artifact resulting from the WRF soundings used in their study. This had also been a point of conversation within the Hazardous Weather Testbed (HWT) Spring Forecasting Experiment (SFE) which is discussed later in this section. Figures 19 and 20 show the observed NWS sounding from Springfield, MO (SGF) and the ensemble mean from each of the five ensembles for RF10 at 18 UTC on 31 May and 00 UTC on 01 June. The observed sounding at 18 UTC on 31 May has a well-defined capping inversion just below 700-hPa, however the ensembles drastically smooth over this feature. This affects the environment in

numerous ways, while the ensemble mean SBCAPE is similar to that of the observed in the shown example, numerous examples analyzed had large discrepancies in the CAPE field with the modeled profiles have much higher values. Additionally, the convective inhibition (CIN) in the ensembles is almost exclusively less as compared to the observed soundings in all of the examples used for the analysis (not shown). By the 00 UTC 01 June SGF sounding, the strong capping inversion has been partially mixed out increasing the CAPE and decreasing the CIN. This is also observed in the modeled soundings, the erosion of the cap is much less realized relative to the observed but the CAPE does still slightly increase. The examples shown in Figures 19 and 20, albeit just 2 soundings from RF10, are representative of what we see in other soundings for this case as well as in the other 2 cases when a capping inversion is present. This suggests an inherent problem related to the implicit damping associated the spatial and temporal finite differencing schemes as well the vertical diffusion associated with the chosen PBL from within the model. Without the ability to replicate these capping inversions at all, the potential for parcels to take advantage of the inflated available energy and decreased inhibition is realized much sooner and could result in timing issues related to CI. Or more significantly, when CI otherwise would not be realized in the real atmosphere the modeled environment which is more conducive to initiation given the lack of capping structure could prematurely or artificially produce convection (i.e. weaker vertical bursts could allow parcels to push through the cap to positive buoyancy in the simulation when in fact it would not be sufficient forcing in the observed environment).

Within the WRF-ARW model, the vertical diffusion is handled by the chosen PBL scheme if one is used. Given that our five ensembles use and differ only in the selected PBL scheme, the influence of PBL parameterization on model error also reflects the influence of

vertical diffusion on model error. Our simulations use the default 3rd order accurate finite difference approximation for vertical derivatives as well as the Runge-Kutta 3rd order temporal discretization scheme. Both of these schemes contain implicit numerical diffusion of all wavelengths and thus will act to weaken the magnitude of gradients that exist within the model. While likely the vertical diffusion is not the only cause of this deficiency, we still test two other means of computing vertical diffusion in WRF in which a PBL scheme is not used within one deterministic member from our ensemble. These methods still yielded similar results in that there was no improvement in the representation of the sounding or capping structure (not shown). Another hypothesis tested was related to the number of vertical levels used in our simulations. Our model configuration uses 40 specified sigma levels, as a test we increased this number of vertical levels to both 75 and 100 and run one deterministic simulation using the same model configuration and the MYJ PBL scheme. For these tests we run the same 15-h simulations initialized at 15 UTC, but we use 0.25 degree GFS analyses as the ICs and LBCs which contains data on 27 vertical levels rather than using the perturbed EnKF analyses originally used to initialize our ensembles. This was done as the initial EnKF analyses were provided to us and thus we could not readily alter the vertical levels within. Again, as was the results from the vertical diffusion test, the increase in number of vertical levels provided very little improvement in the model's ability to resolve these capping features (not shown). One could speculate that the short, 3- to 6-h model spin up prior to the examination of the vertical profiles may hinder the model's ability to resolve these caps anyway since we begin with 27 vertical levels and interpolate to 75 or 100. However, this test was conducted as a preliminary investigation into this deficiency within our ensembles and the results were at the very least worth noting.

Not all models however suffer from this inability to resolve capping inversions. For example the United Kingdom Meteorology Office has conducted 48-h forecasts over the CONUS with their convection-allowing versions of the Unified Model (UM) for the 2013, 2014, and 2015 HWT SFEs and found much more promising results. During the 2014 SFE, it was noticed that the UM significantly outperformed the National Severe Storms Laboratory (NSSL) version of the WRF in representing vertical soundings, especially those which contained strong capping inversions (see Figure 2 of Clark et al. 2015b). One way the UM model differs from WRF is that it is semi-lagrangian in its 3-dimensional treatment of advection terms, meaning unlike WRF it does not use implicit diffusion or finite differencing and thus does not have the issue of weakening vertical gradients mentioned previously.

4. Summary

Convection-allowing WRF-ARW 30 member ensembles varying only in the selected PBL scheme were simulated to analyze the impact of the PBL parameterization on the forecast of CI. Using a combination non-local and local mixing PBL schemes, five ensembles were simulated for each of the fourth, tenth, and twelfth research flight (RF4, RF10, and RF12) days from the MPEX field campaign. These three events were selected given their varying large scale flow regimes as an attempt to identify the predictability of differing synoptic scale setups. The simulations utilize the additional sub-synoptic scale observations gathered during MPEX in the EnKF produced Initial analyses and are simulated for 15-h. With these simulations we investigate the ability of convection-allowing ensembles to accurately represent CI by evaluating the forecast skill of an object-based approach which classifies CI as spatially and temporally sustained radar reflectivity-derived objects.

Utilizing a consistent object identifying and tracking algorithm from within the WDSS-II suite for both the observed and modeled CI objects, a spatiotemporal flow-dependent error (*C*) was computed between the objects. For all three events an overproduction of CI objects was detected as the tracking of the observed reflectivity data yielded 238, 75, and 180 objects while the modeled reflectivity yielded on average, ignoring the differing PBL ensembles, 399.88, 117.84, and 325.67 events for cases RF4, RF10, and RF12 respectively. The non-local MYJ and QNSE schemes overproduced by the largest margins in all three cases with the overproduction mostly occurring within the first 6-h of the forecast. The ability of the ensembles to accurately forecast a CI object within 100-km and 60-min was provided via the performance diagrams of Figures 9 and 15. RF4 had the highest values of POD, SR, and CSI across the ensembles with RF10 and RF12 following in that order. Between each case the non-local MYJ and QNSE schemes had the highest POD but simultaneously had the highest FAR and bias resulting from the noted overproduction. The local ACM2 scheme yielded the lowest POD in each of the three cases but also had the lowest FAR and was the least biased. In general the chosen PBL scheme does not significantly improve the skill of forecast since the CSI within any one case had very subtle changes between PBL ensembles, but this study highlights artifacts related to each scheme's behavior of producing CI. Additionally, examining the hits for each case and ensemble member, it was found there exists little systematic bias in the timing of CI when it was both observed and forecasted by a model as the distribution of errors was approximately normal with mean errors of about 0. This result was interesting as it echoed the findings from within Kain et al. (2013). Further, the distribution of spatial errors from the hits showed a mean distance error of 43.789, 42.629, and 47.495 -km for RF4, RF10, and RF12 but these distributions were skewed right and had on average about 55% of the data less than the mean. More notably was when the

statistics were recomputed with the spatial and temporal thresholds increased to 300-km and 180-min, CDFs showed that even at these extended thresholds if an object was matched as a hit the probability of it being within 100-km was 83.490%, 87.319%, and 74.910% and within 60-min was 75.366%, 78.453%, and 66.692% for RF4, RF10, and RF12. Thus suggesting there is a limit on the added value to forecast skill by increasing the temporal and spatial constraints. This study highlights that using 60-min and 100-km as the thresholds to conduct forecast verification in an object-based approach provides a good representation of the forecast skill while still constraining the results to relatively restrictive thresholds.

Analysis of the environment in which CI occurred is a direct representation of the biases associated with each PBL scheme. While it was observed that the different PBL schemes had little effect on the CSI and associated forecast skill, the biases on environmental conditions can provide insight into possible causes for the forecast skill sensitivity. For example it was observed that there existed a cool and moist bias within the domain of the ICs, this could have resulted in lower LCLs allowing for convection to be more readily initiated. Additionally, from an PBL to PBL perspective, the MYNN and local mixing ACM2 and YSU scheme were consistently warmer and drier that the non-local MYJ and QNSE. This coupled with these non-local schemes having the largest CI overproduction, there appears to be some connection with the larger cool and moist bias and the overproduction. Another bias which was noticed in the PBLs representation of the pre-convection vertical environment was the local mixing PBL schemes were too vigorous in the vertical mixing resulting in deeper drier PBLs relative to the other schemes. Similar findings have been found in recent works Hu et al. (2010), Coniglio et al. (2013), and Clark et al. (2015a), this result suggests that with the stronger mixing of the non-local schemes, any temperature inversions will be mixed out quicker decreasing the CAPE and

subsequently decreasing the probability of convection occurring. Additionally this is opposite for the non-local schemes and could also play part in the documented CI overproduction. This theory holds true in the analysis of individual observed soundings compared with modeled soundings. Once a capping inversion began being mixed out, the local schemes often eroded the CAPE field quicker. This analysis also brought to light the model's inability to accurately reproduce strong capping inversions, a feature also pointed out in Coniglio et al. (2013). The modeled forecasts here have shown a tendency to smooth over sharp gradients of temperature resulting in an over prediction of CAPE and an under prediction of CIN. This issue persists even when the number of vertical levels is increased from 40 to 75 and up 100 suggesting that the issue possibly lies within the vertical diffusion of WRF.

The results presented here are a beginning analysis in the sensitivity of the PBL parameterization on the forecast of CI and attempt to discuss the effects that the biases these schemes related to the PBL environment have of the forecast skill CI in an object-based approach Our results, relative to other studies and future works, aim to identify the impact a chosen PBL parameterization scheme has on an ensemble forecast of CI. Within, we identify such impacts and also justify the definition of CI and the use of discussed spatial and temporal thresholds. While robust, our results are explicitly valid for only the chosen model configuration. We acknowledge that our sample size of three events is small, and expanding to encompass a wider range of events is dedicated to future work. Further future work involves a more in-depth analysis of the kinematic forcing which leads to CI, an expansion of the verification to test multiple techniques, and examination of the temporal evolution of forecast skill within our simulations.

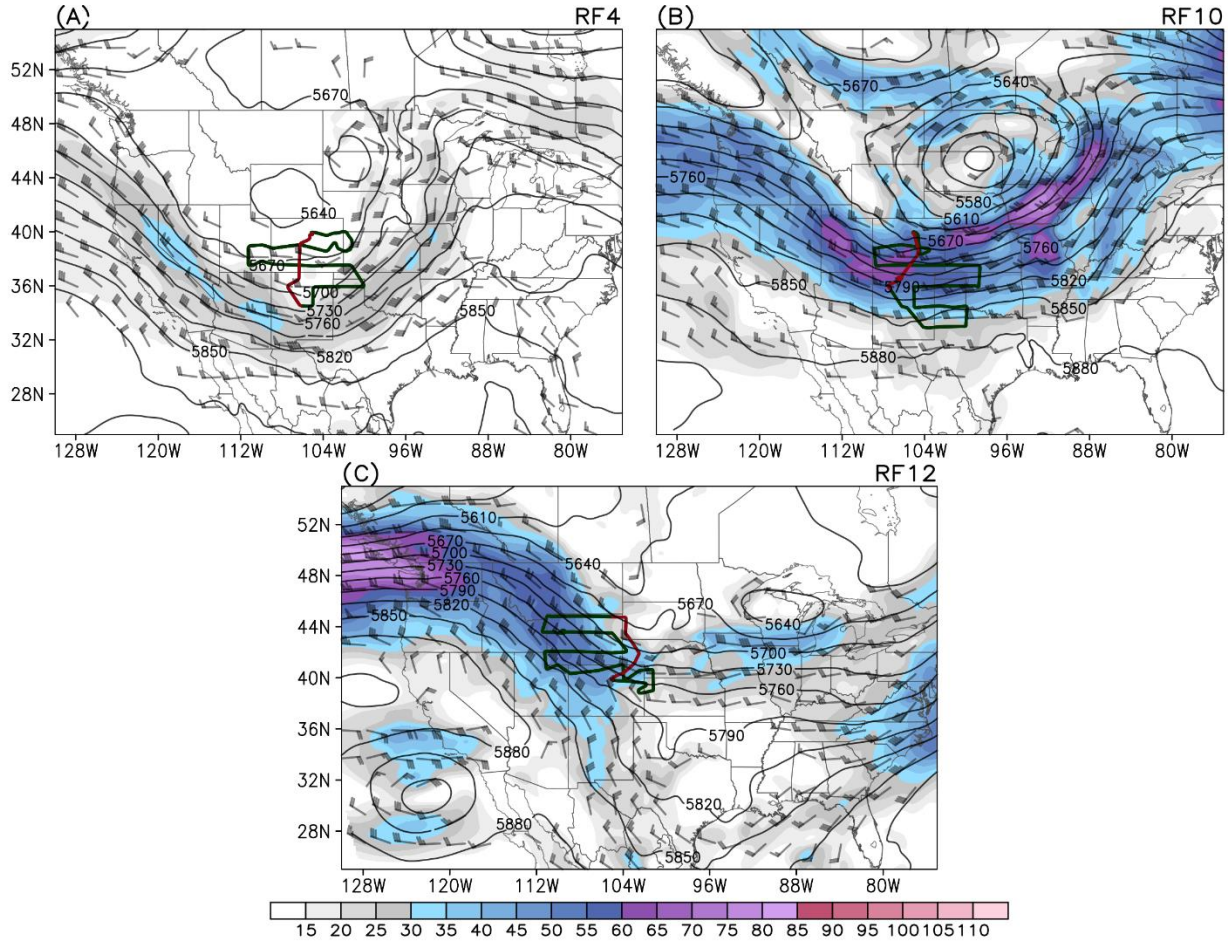

Figure 1: 00 h forecast analysis time valid at 15z on the day of interest from the 20 km Rapid Refresh (RAP; Benjamin et al. 2015). Data obtained from the NCAR Research Data Archive (http://rda.ucar.edu/) and showcases the synoptic scale pattern of the three selected events as well as the targeted region for MPEX additional sub-synoptic observations via the observed flight path of the NCAR GV aircraft. Plot shows 500 hPa heights (contour) in [m/s] and wind speed (barbs and isotachs) in [m/s]. The GV aircraft track is contoured in dark green with the final hour of the flight in red. Panels (A), (B), and (C) correspond to RF4 [19 May 2013], RF10 [31 May 2013], and RF12 [8 June 2013], respectively.

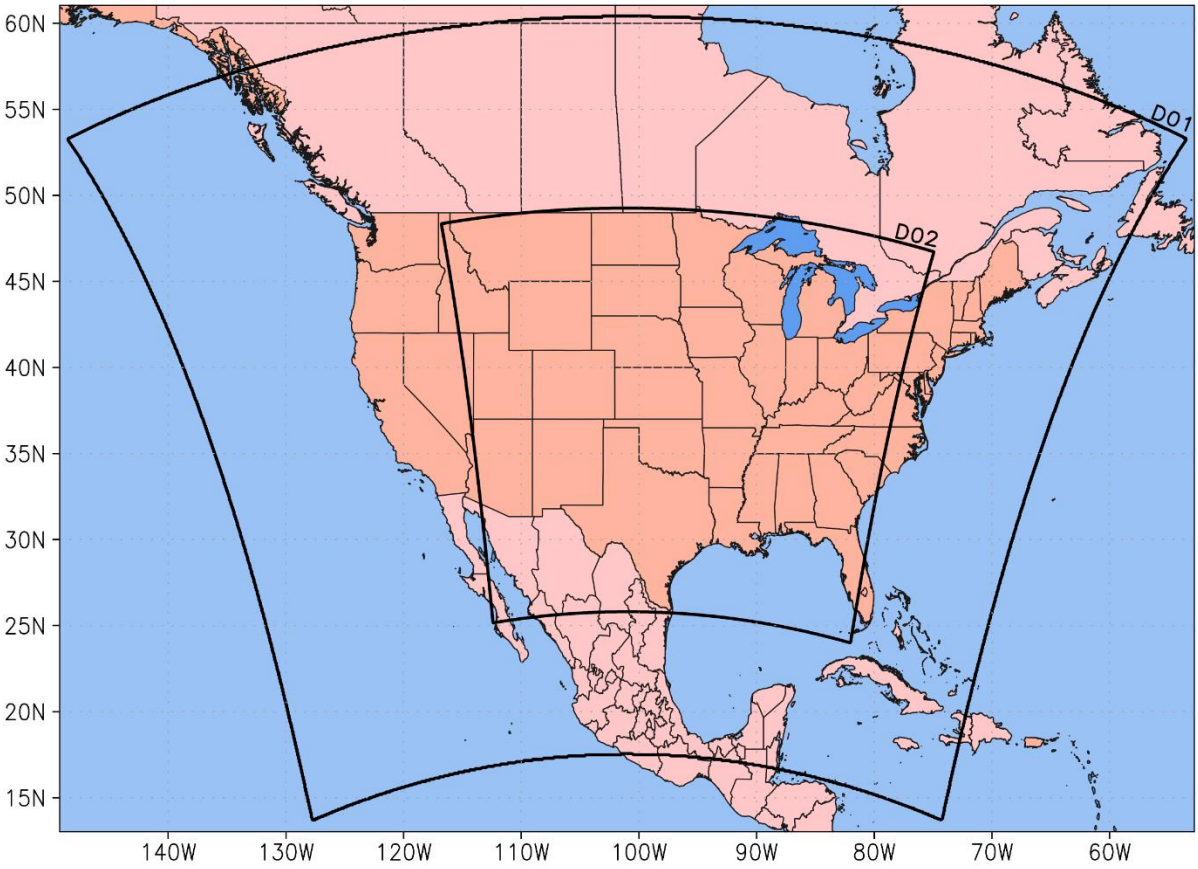

Figure 2: Areal extent of both domain 1 (D01) and domain 2 (D02) used for the cycled analysis (D01 only) and ensemble forecasts (D01 and D02).

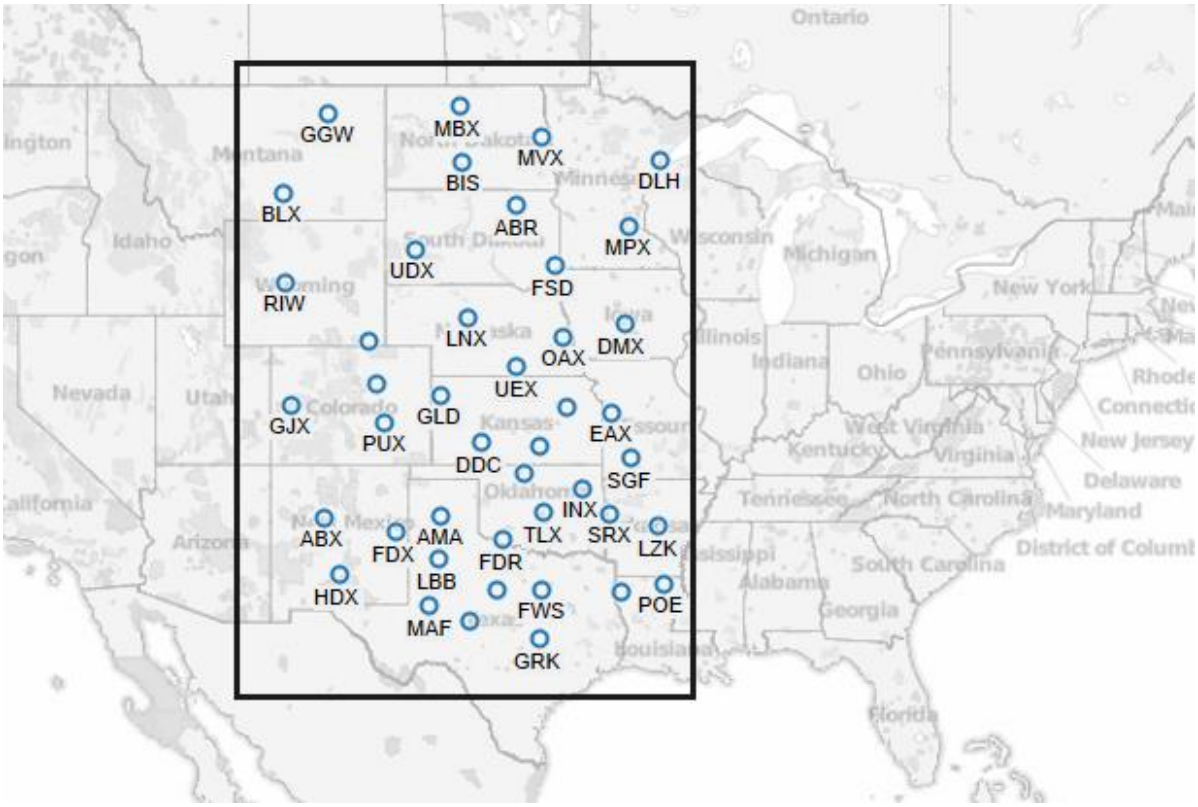

Figure 3: WSR-88D sites (blue circles, some labeled with station identifier) used to determine observed CI events using WDSS-II. The black box represents the areal bounds of the 0.03° x 0.03° grid to which both the observed and modeled radar reflectivity are interpolated for verification purposes.

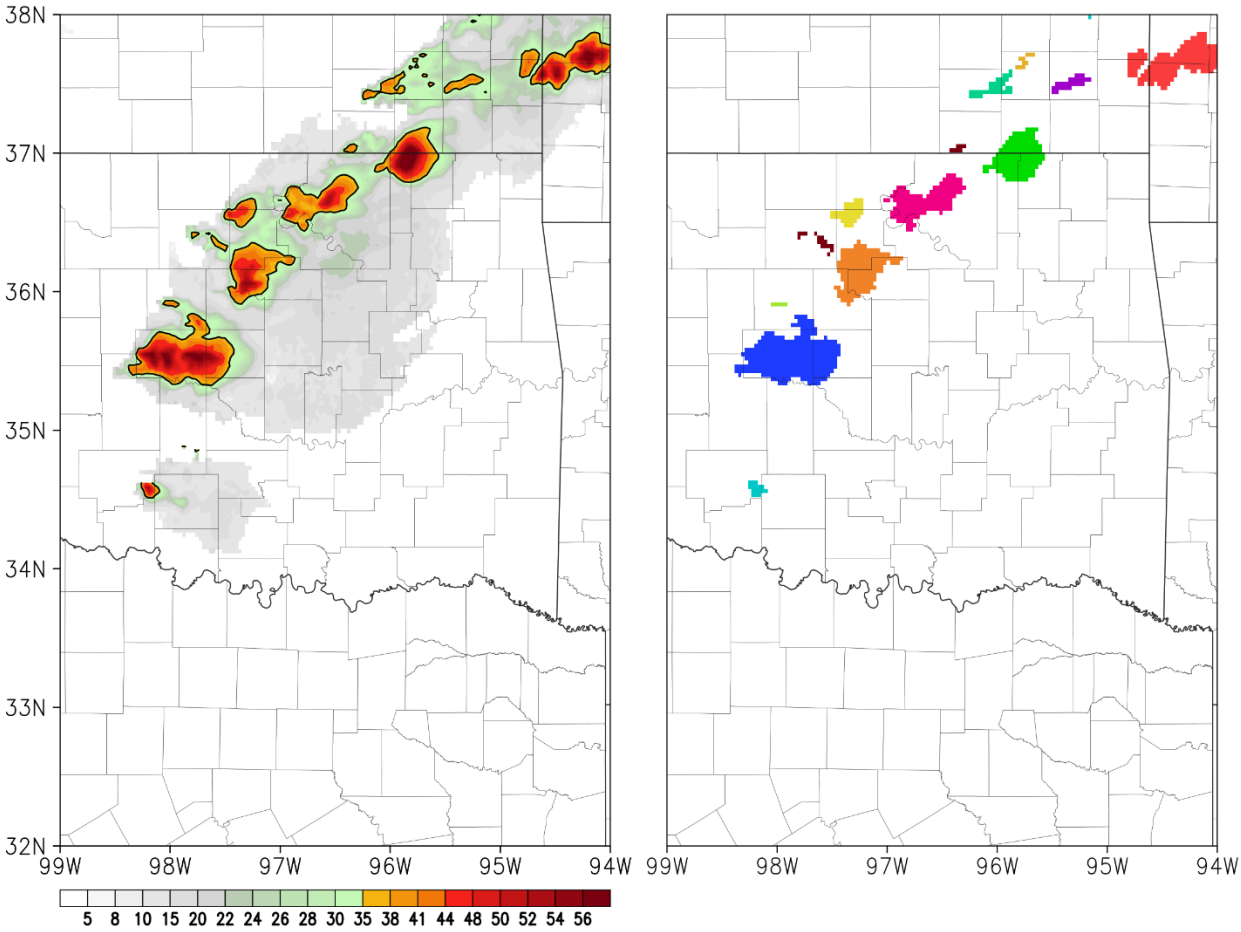

Figure 4: Example of WDSS-II object identification and tracking. Reflectivity (shaded; dBz) at -10℃ isotherm from 42 NEXRAD locations merged onto uniform grid with 35 dBz contoured in black (left). CI Objects as identified by WDSS-II (right).

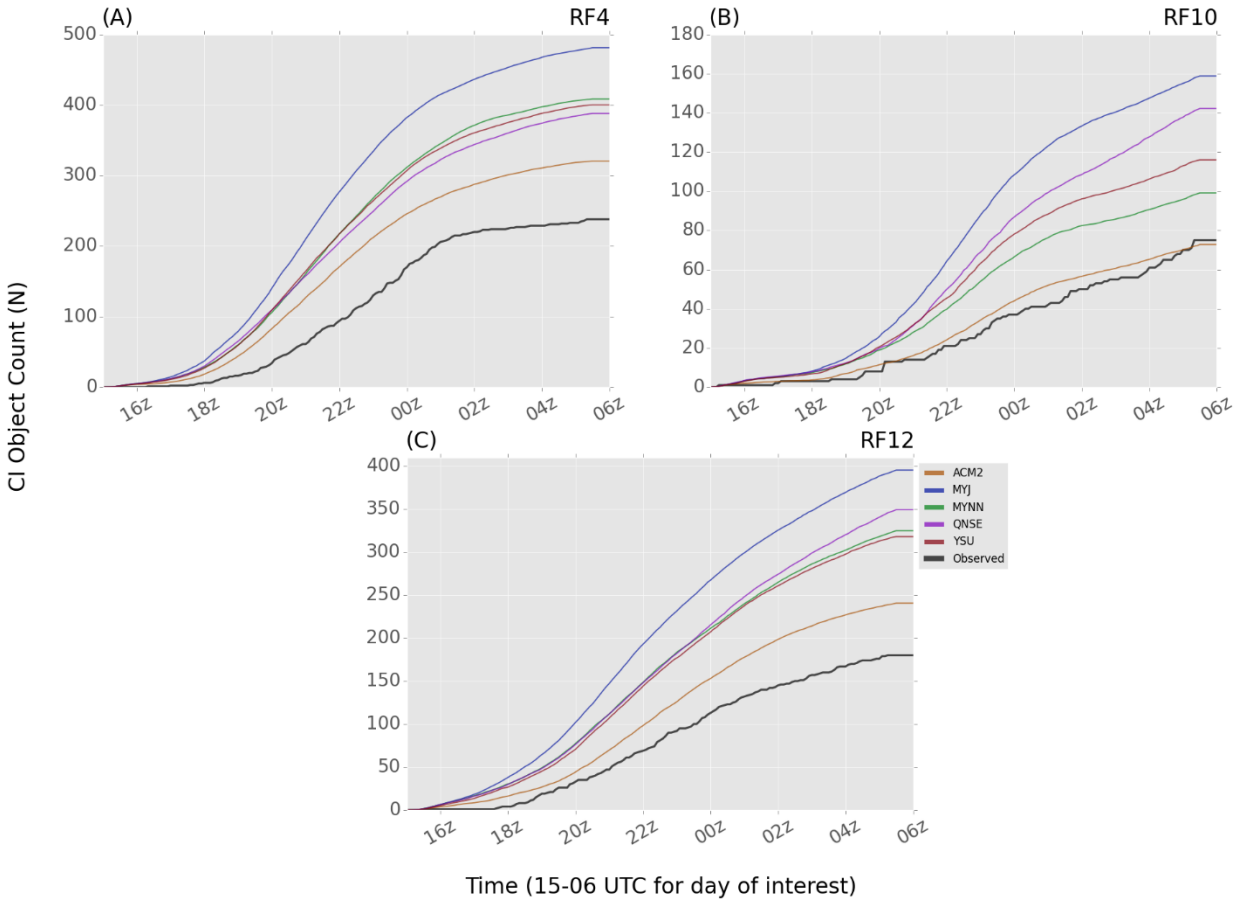

Figure 5: Count of CI objects within the entire domain, as aggregated through the 15 hours of each simulation. Colored lines represent the ensemble mean from each different PBL ensemble and black line represents observed total. Panels (A), (B), and (C) correspond to RF4 [19 May 2013], RF10 [31 May 2013], and RF12 [8 June 2013], respectively.

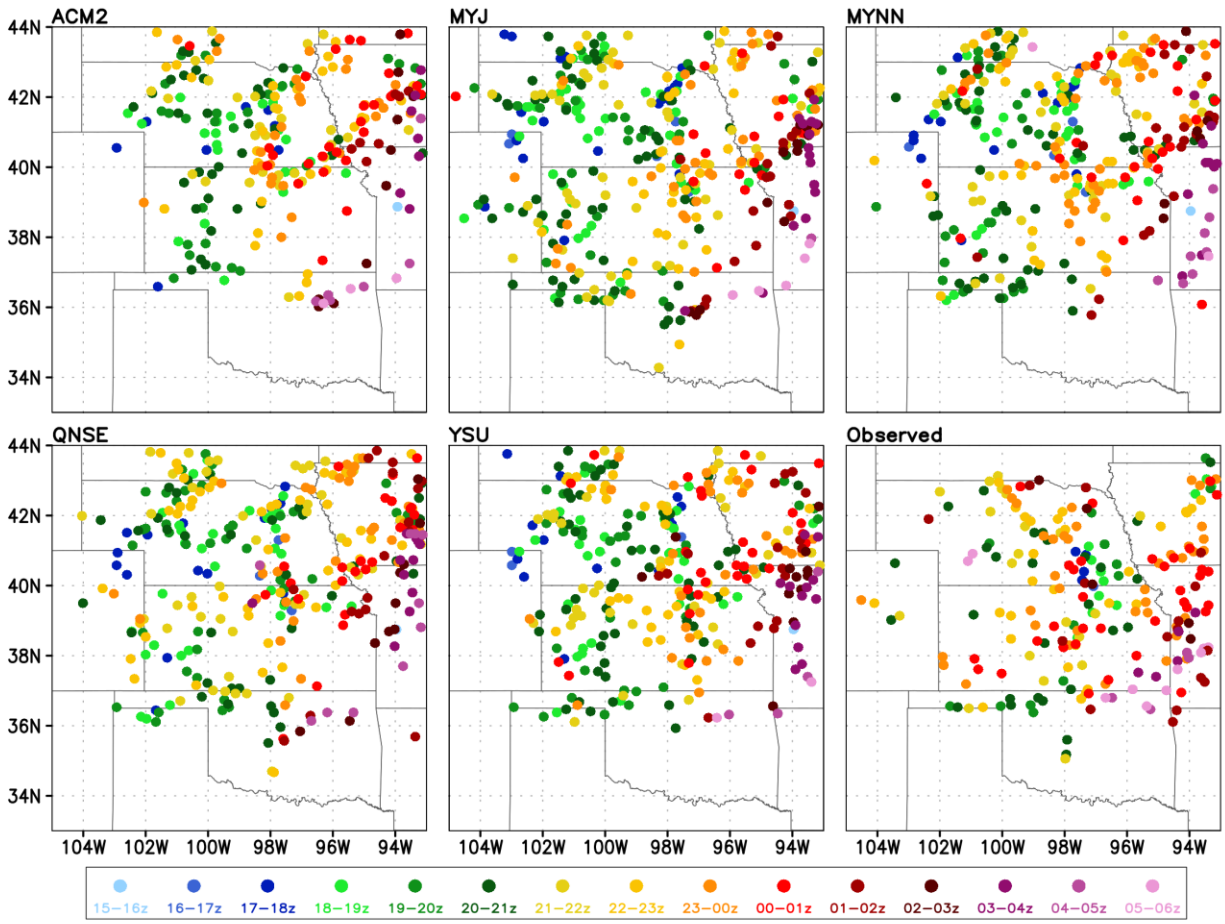

Figure 6: Paintball plot for RF4 [19 May 2013]. Dots color-coded and located based on timing and location of when each event was first identified as CI. Panels show 1 randomly selected member (26) from each ensemble with the observed in the lower right panel.

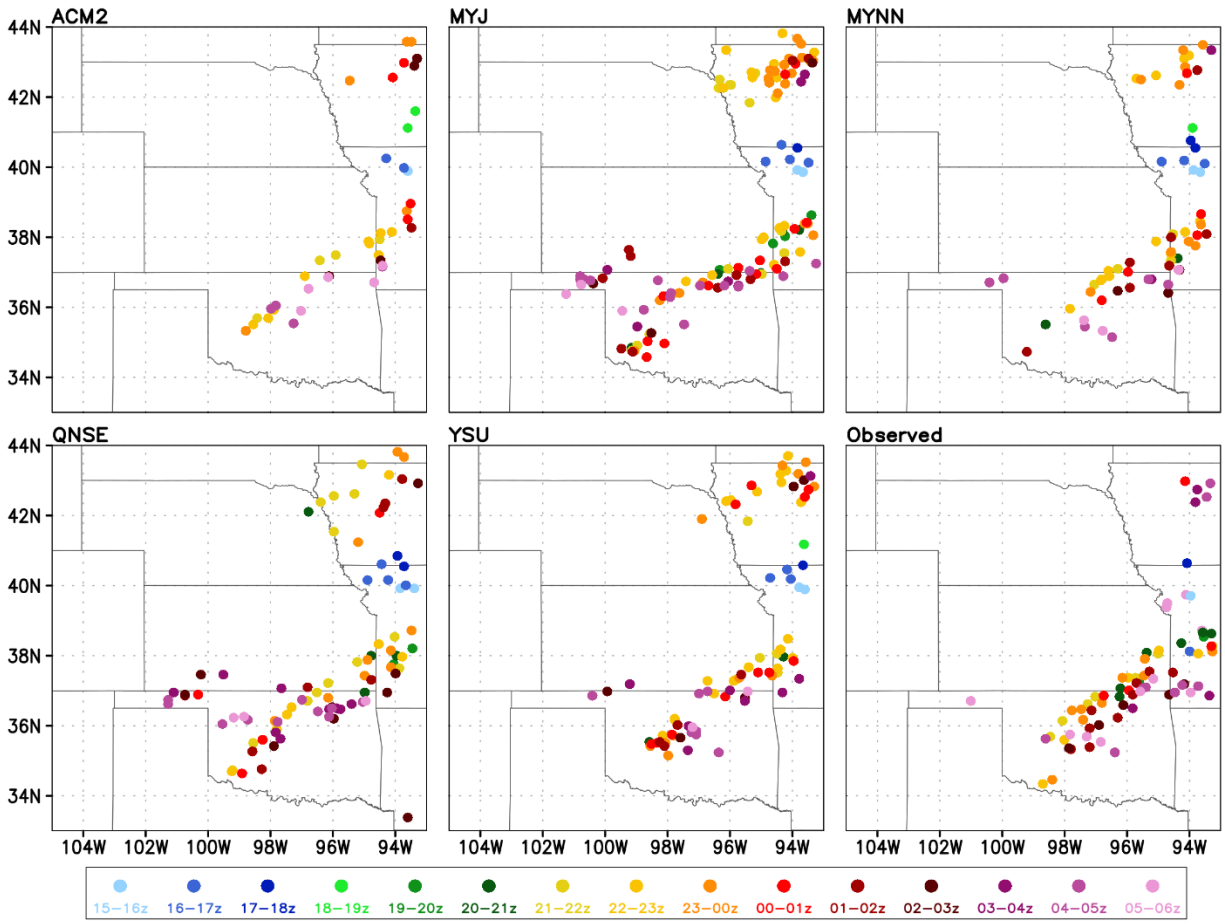

Figure 7: Paintball plot for RF10 [31 May 2013]. Dots color-coded and located based on timing and location of when each event was first identified as CI. Panels show 1 randomly selected member (26) from each ensemble with the observed in the lower right panel.

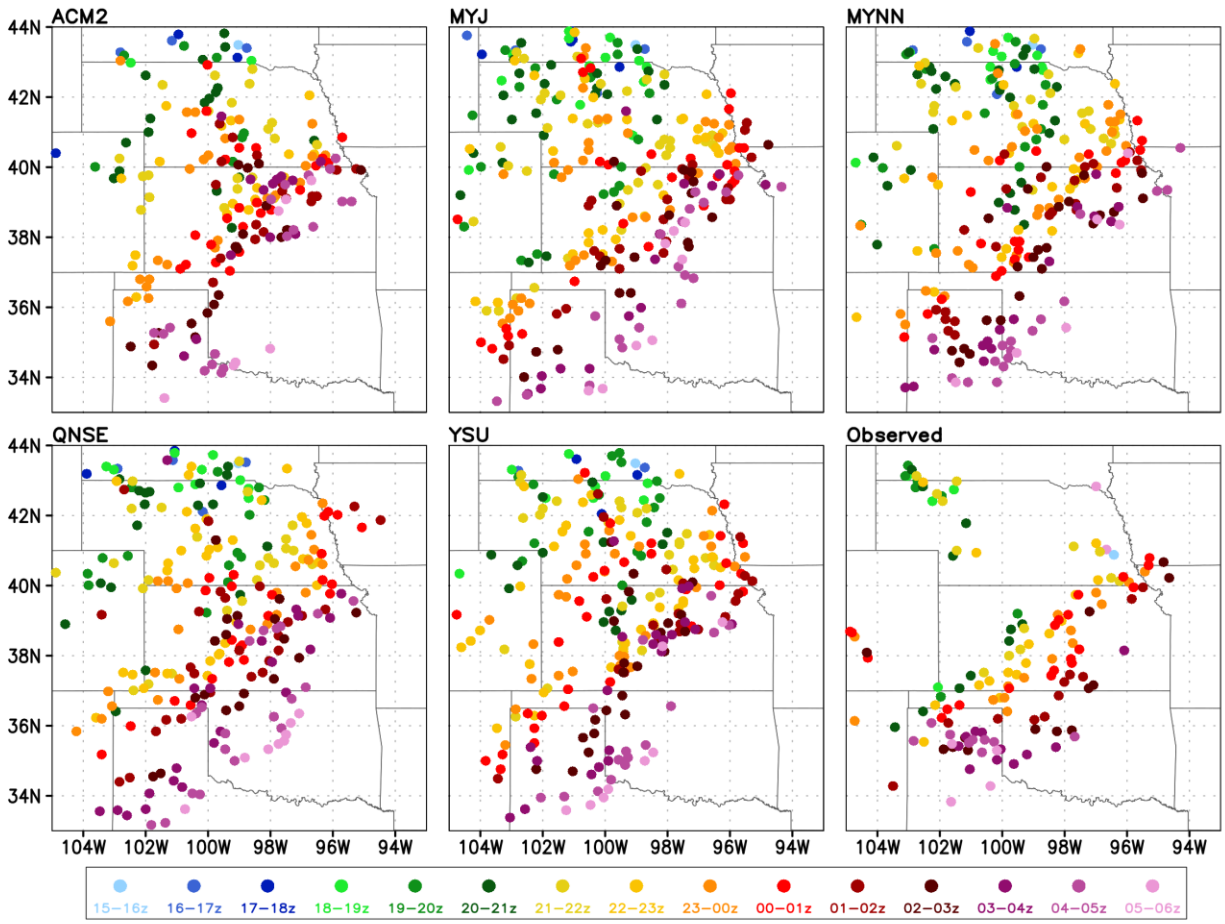

Figure 8: Paintball plot for RF12 [8 June 2013]. Dots color-coded and located based on timing and location of when each event was first identified as CI. Panels show 1 randomly selected member (26) from each ensemble with the observed in the lower right panel.

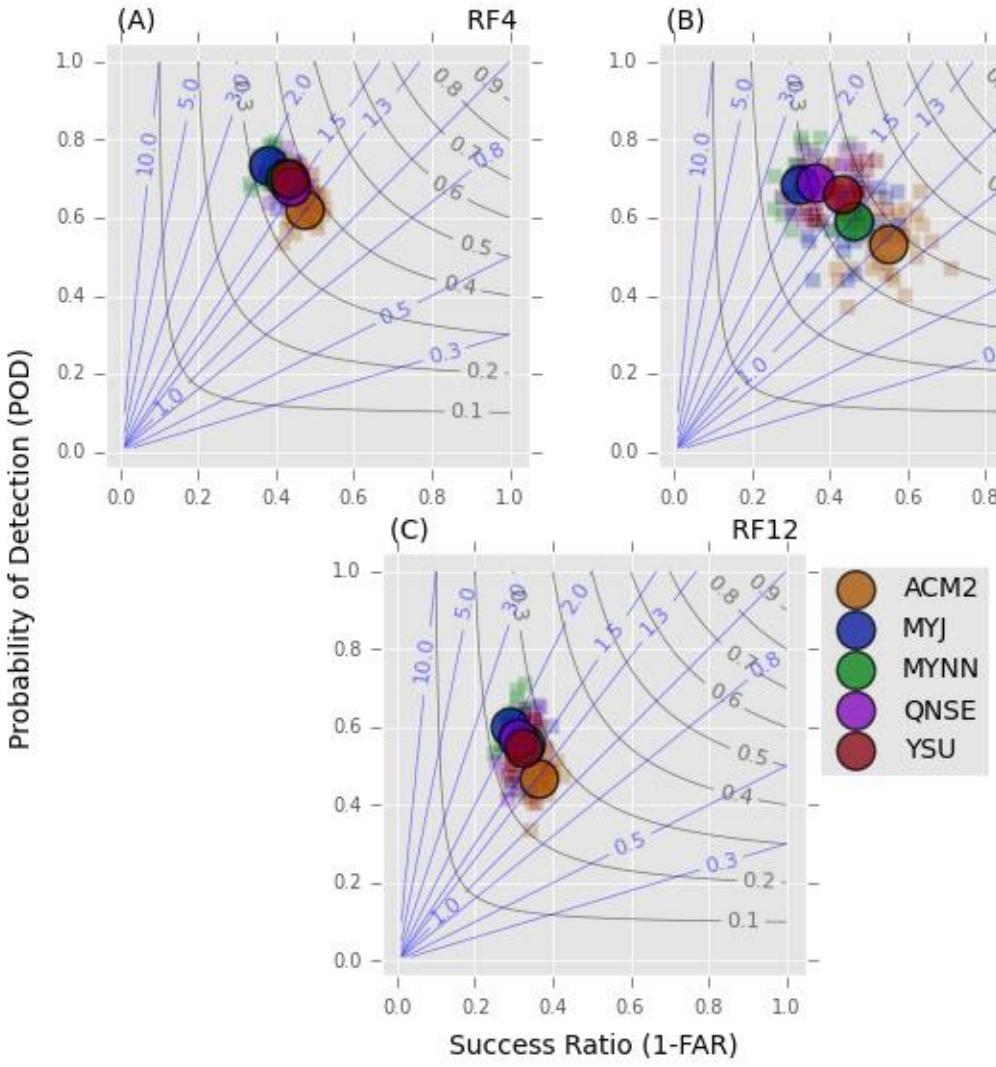

Figure 9: Roebber (2009) performance diagrams for RF4 [(A); 19 May 2013], RF10 [(B); 31 May 2013], and RF12 [(C); 8 June 2013]. Large color coded circles correspond to ensemble mean statistics while smaller colored boxes represent each ensemble member.

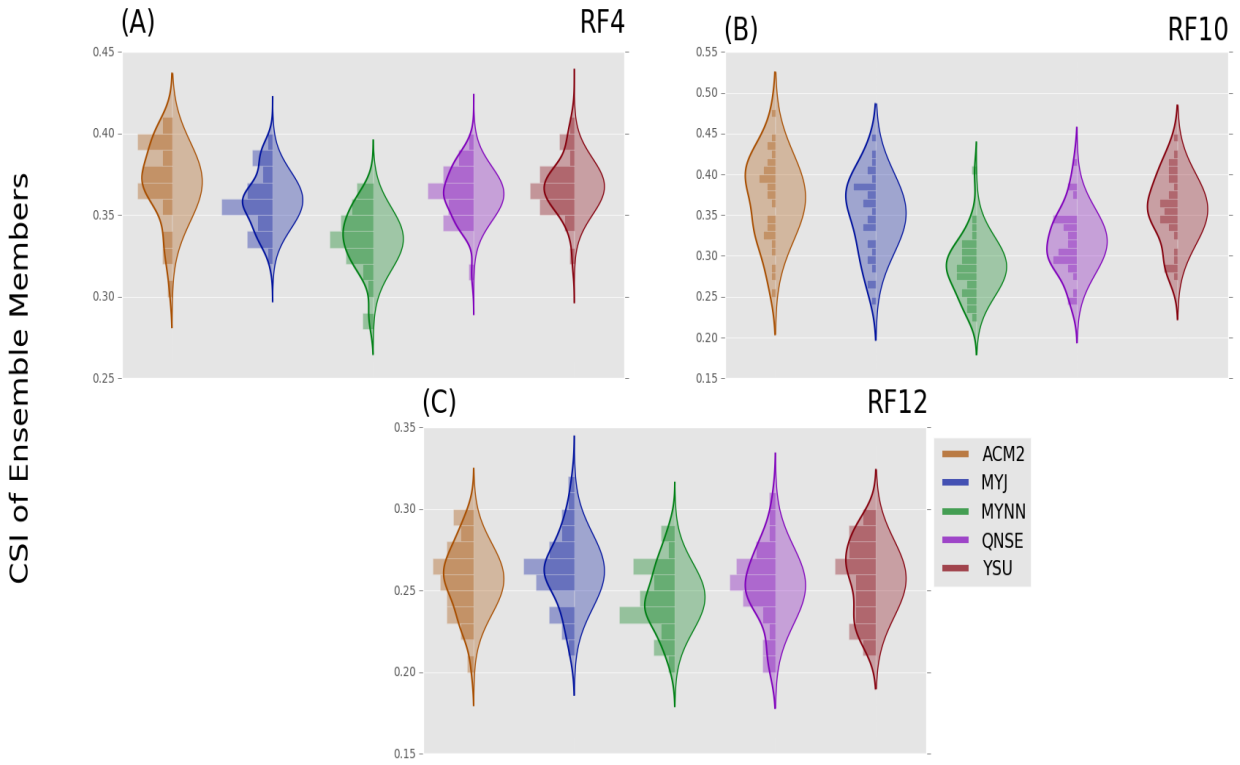

Figure 10: Violin plots of CSI distribution within each ensemble. Left side of violin shows values from 30 member ensemble binned by 0.01 increments with PDF fit from KDE. Right side of plot shows normalized PDF as computed from the mean and standard deviation of the CSI distribution. Panels (A), (B), and (C) correspond to RF4 [19 May 2013], RF10 [31 May 2013], and RF12 [8 June 2013], respectively.

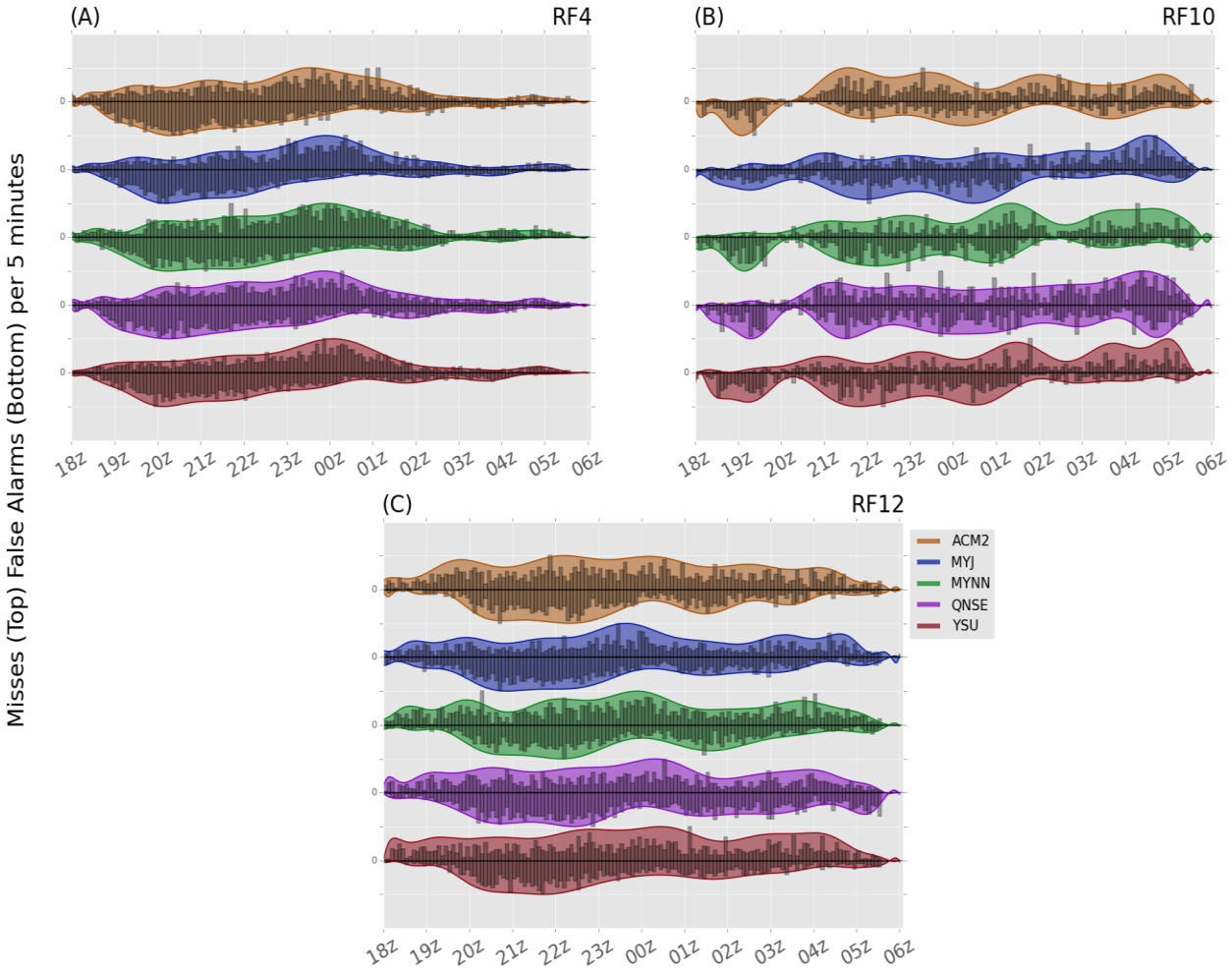

Figure 11: Histograms of the misses (top) and false alarms (bottom) aggregated over all the ensemble members in 5 minute bins throughout the 15 hour simulation. Histograms are scaled by the max value within the distribution. A polynomial function is then fit to the histogram. Panels (A), (B), and (C) correspond to RF4 [19 May 2013], RF10 [31 May 2013], and RF12 [8 June 2013], respectively.

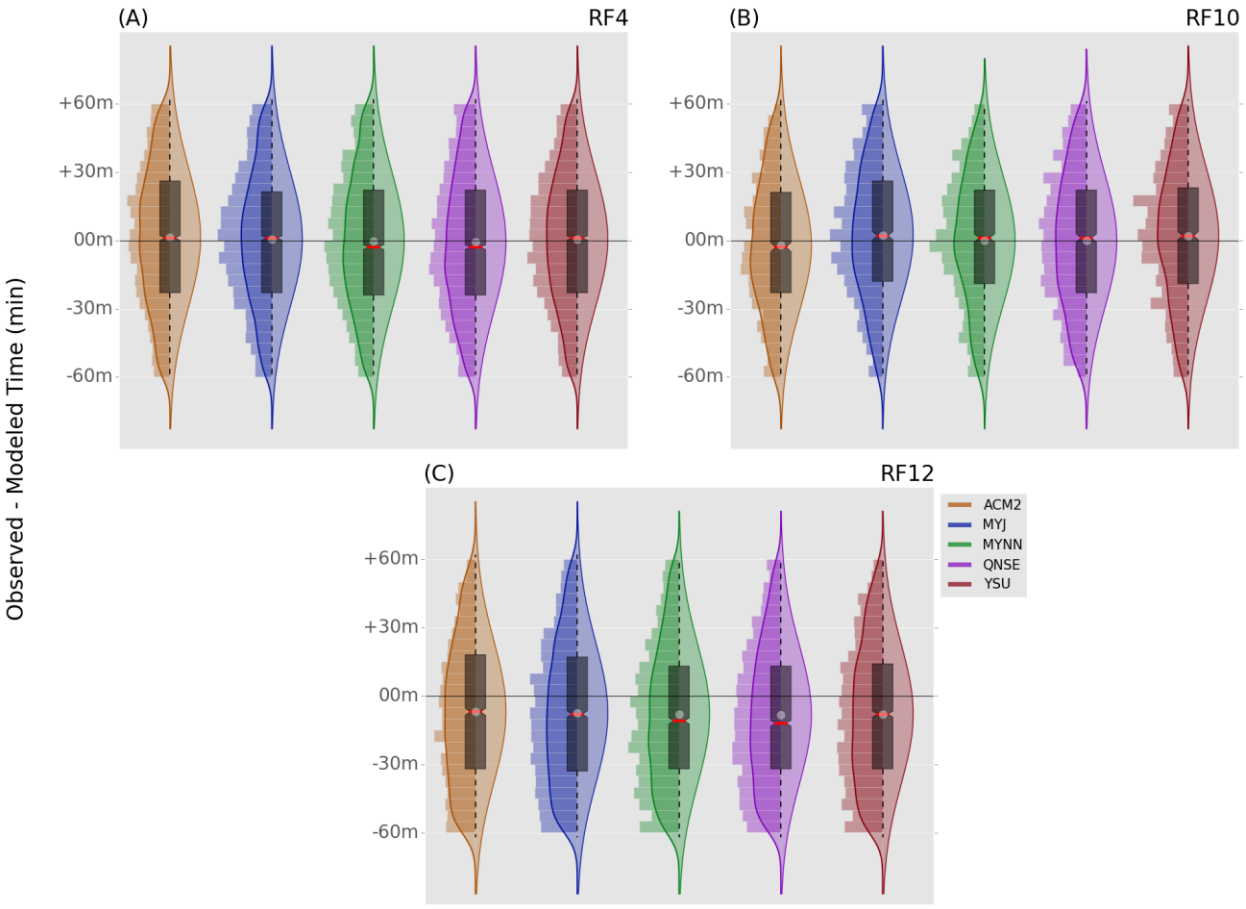

Figure 12: Violin plots of temporal difference between observed and modeled CI events within each ensemble. Left side of violin shows aggregate values from each PBL ensembles binned by 5-m increments with PDF fit from KDE. Right side of plot shows normalized PDF as computed from the mean and standard deviation of the temporal histogram. Panels (A), (B), and (C) correspond to RF4 [19 May 2013], RF10 [31 May 2013], and RF12 [8 June 2013], respectively.

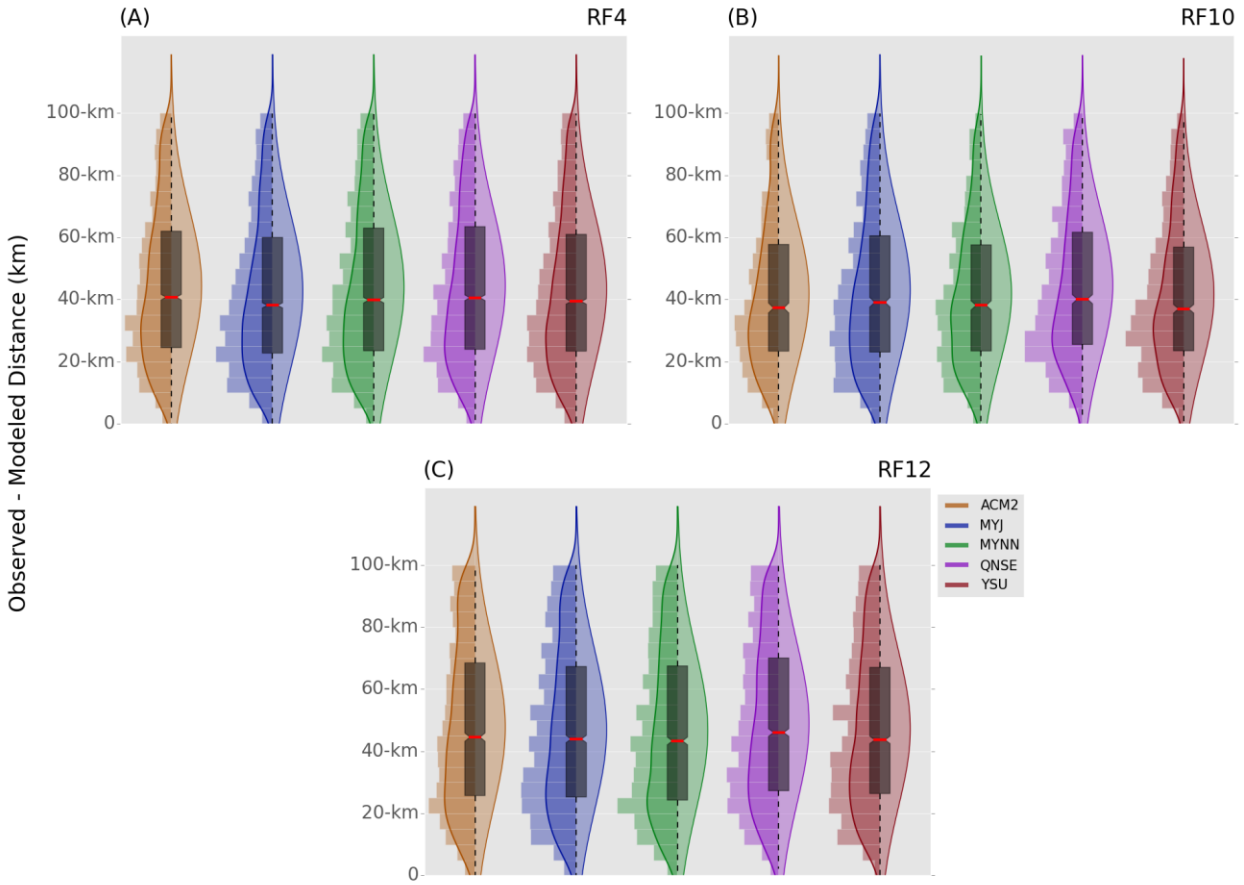

Figure 13: Violin plots of spatial difference between observed and modeled CI events within each ensemble. Left side of violin shows aggregate values from each PBL ensembles binned by 5-km increments with PDF fit from KDE. Right side of plot shows normalized PDF as computed from the mean and standard deviation of the spatial histogram. Panels (A), (B), and (C) correspond to RF4 [19 May 2013], RF10 [31 May 2013], and RF12 [8 June 2013], respectively.

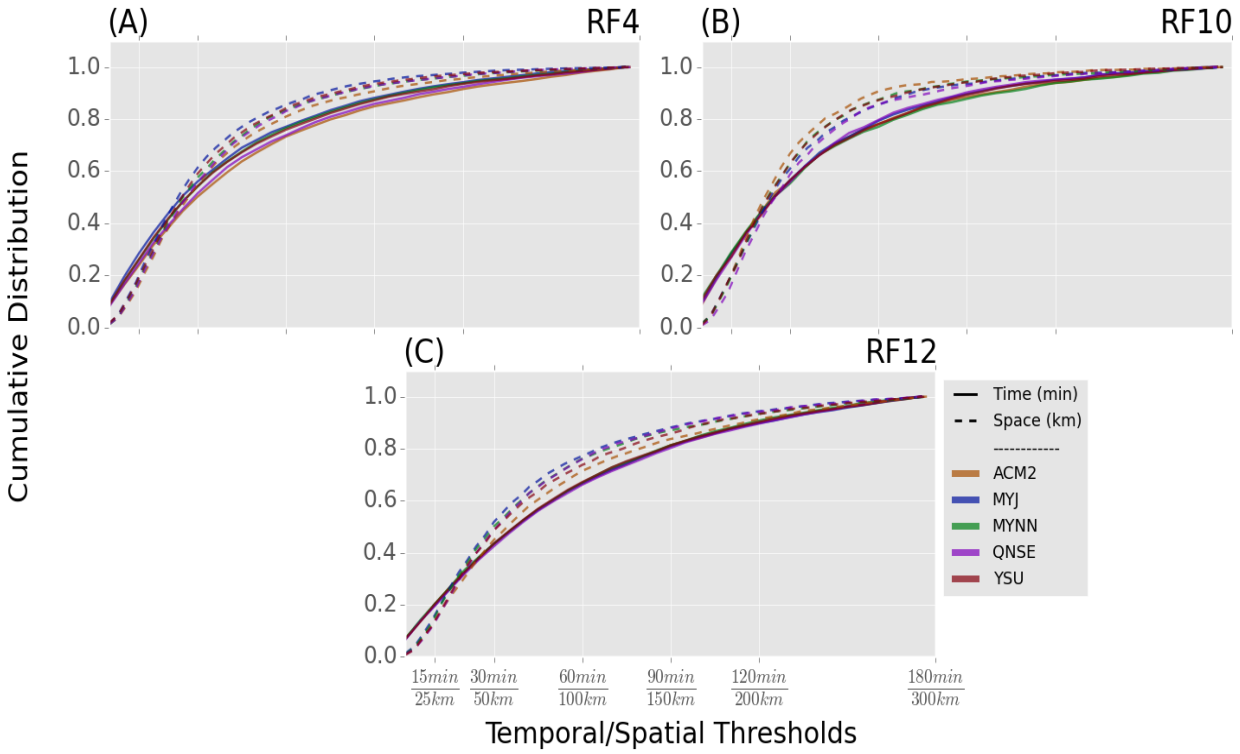

Figure 14: CDF of aggregated ensemble members hits binned in 5-min and 5-km bins based on error from observed. Solid (dashed) lines show the temporal (spatial) CDF, y-axis shows the total number of hits per ensemble, and x-axis shows time/space thresholds. Panels (A), (B), and (C) correspond to RF4 [19 May 2013], RF10 [31 May 2013], and RF12 [8 June 2013], respectively.

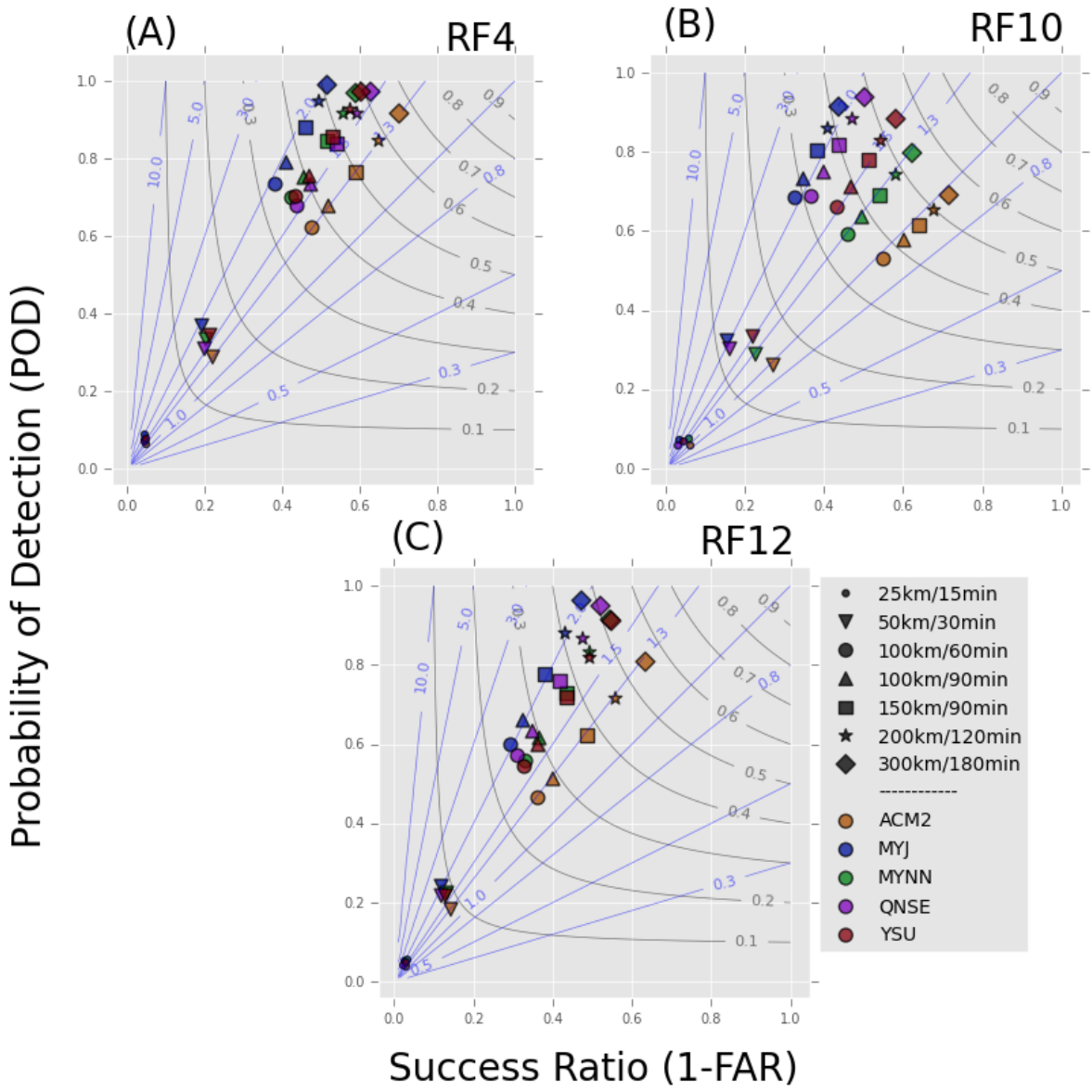

Figure 15: Roebber (2009) performance diagrams for RF4 [(A); 19 May 2013], RF10 [(B); 31 May 2013], and RF12 [(C); 8 June 2013]. Large color coded shapes correspond to ensemble mean statistics while shape corresponds to various thresholds.

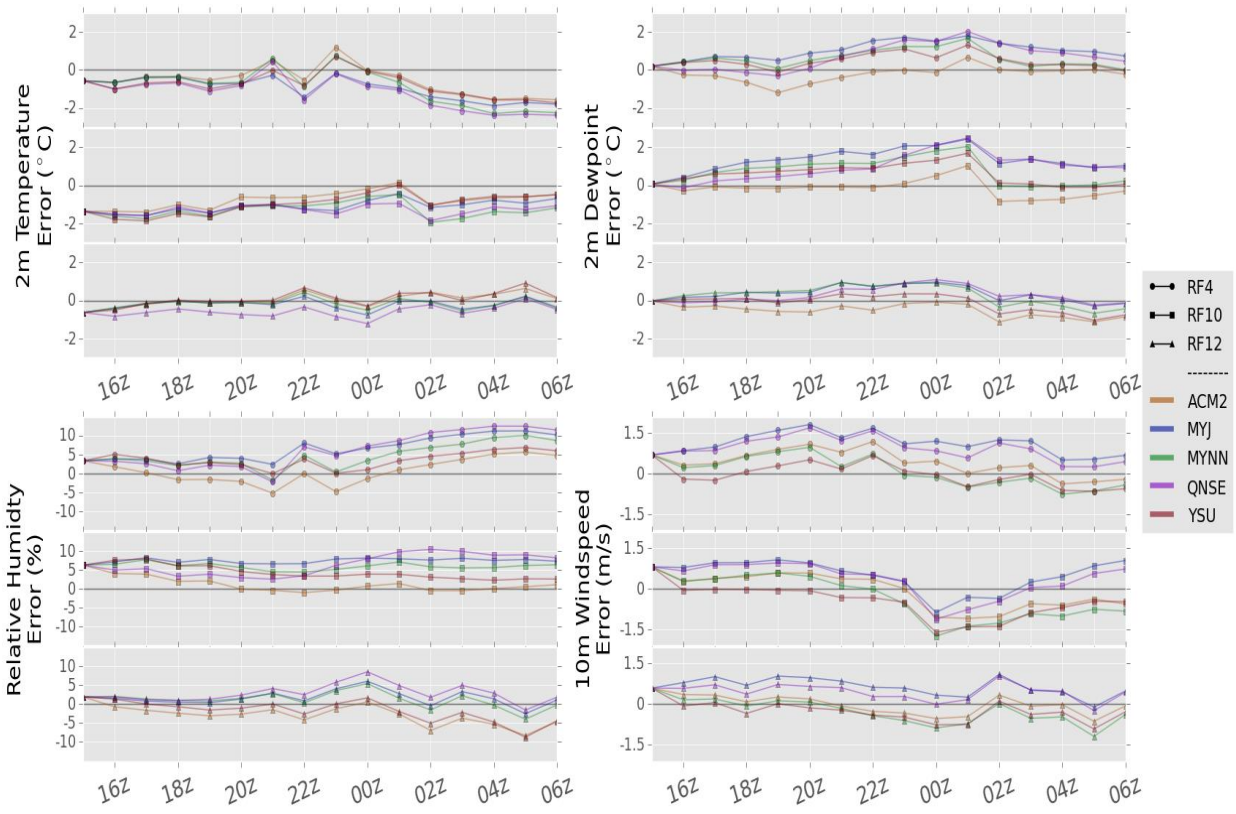

Figure 16: Mean Error from ensemble mean and NCEP MADIS metar data as averaged over the domain throughout time. 2-m temperature error is the top left, 2-m dewpoint error is the top right, 2-m RH is the bottom left, and 10-m wind speed is the lower right panel. Within each subpanel, RF4, RF10, and RF12 are plotted separately in that order, from top to bottom.

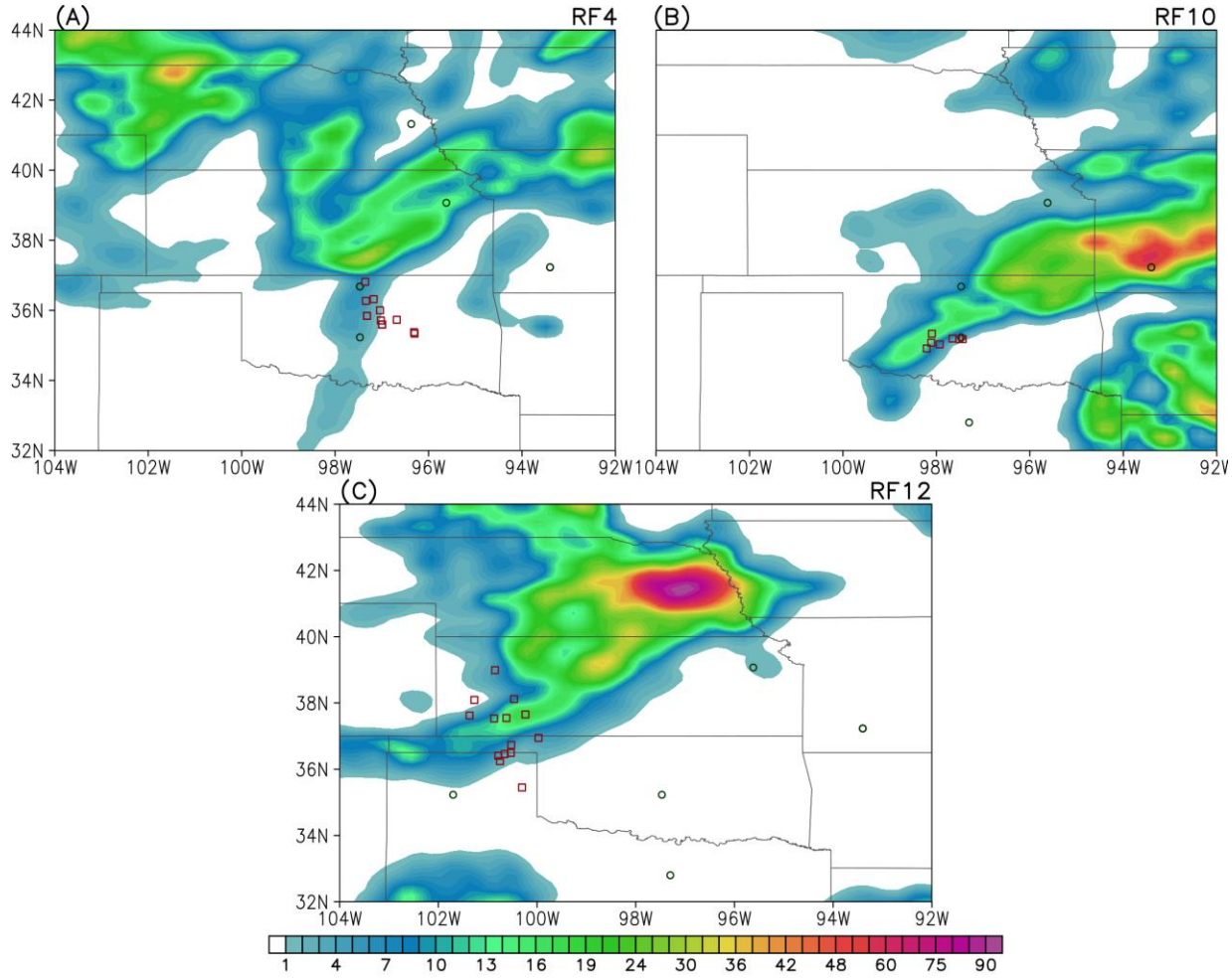

Figure 17: Locations of soundings used to calculate vertical profiles of ME in pre-convective environments between modeled and observed soundings. These are plotted over the 6-h (18-00 UTC on day of event) accumulated precipitation [mm] from the 20-km 15 UTC RAP forecast. Red squares represent the MPEX mobile soundings and green circles are the NWS 00 UTC and subsequent 18 UTC soundings. Panels (A), (B), and (C) correspond to RF4 [19 May 2013], RF10 [31 May 2013], and RF12 [8 June 2013], respectively.

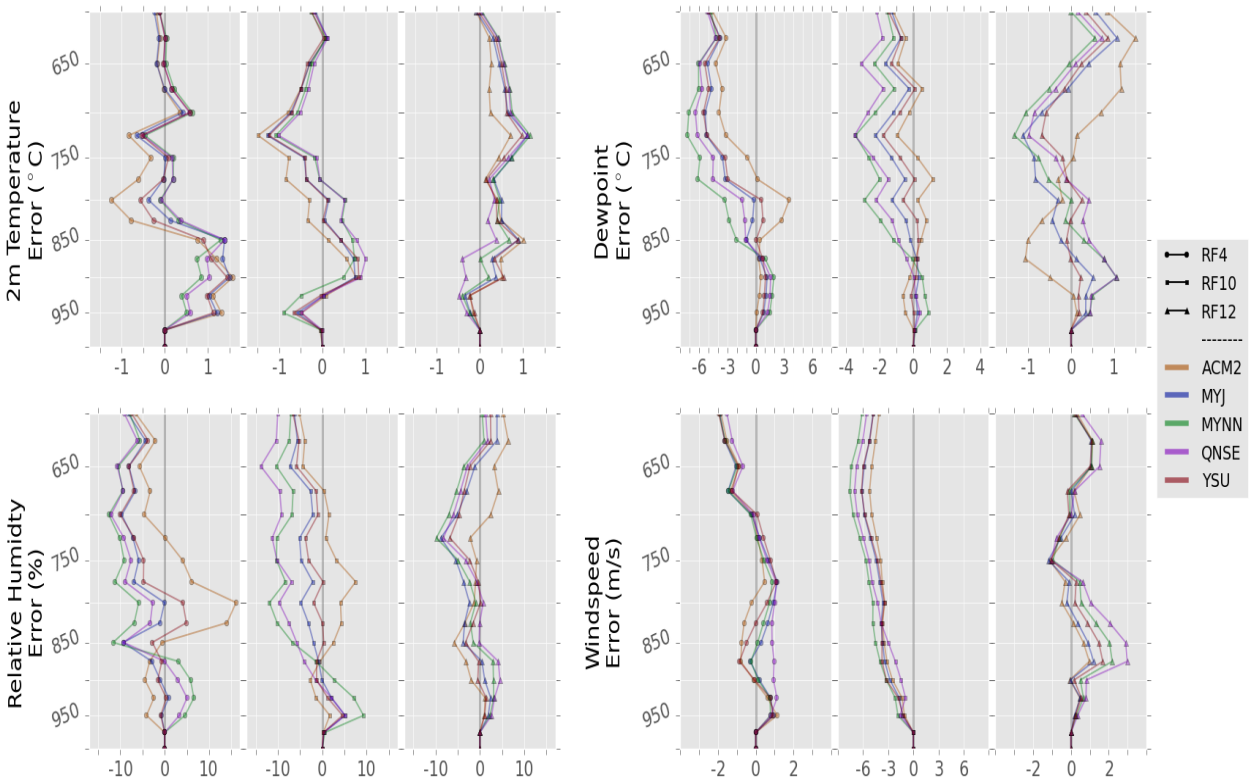

Figure 18: Vertical profiles of ME in pre-convective environments between modeled and observed soundings. 2-m temperature error is the top left, 2-m dewpoint error is the top right, 2-m RH is the bottom left, and 10-m wind speed is the lower right panel. Within each subpanel, RF4, RF10, and RF12 are plotted separately in that order, from left to right. [**Note within the 2-m dewpoint and 10-m wind speed errors that the scales change between the three presented cases].

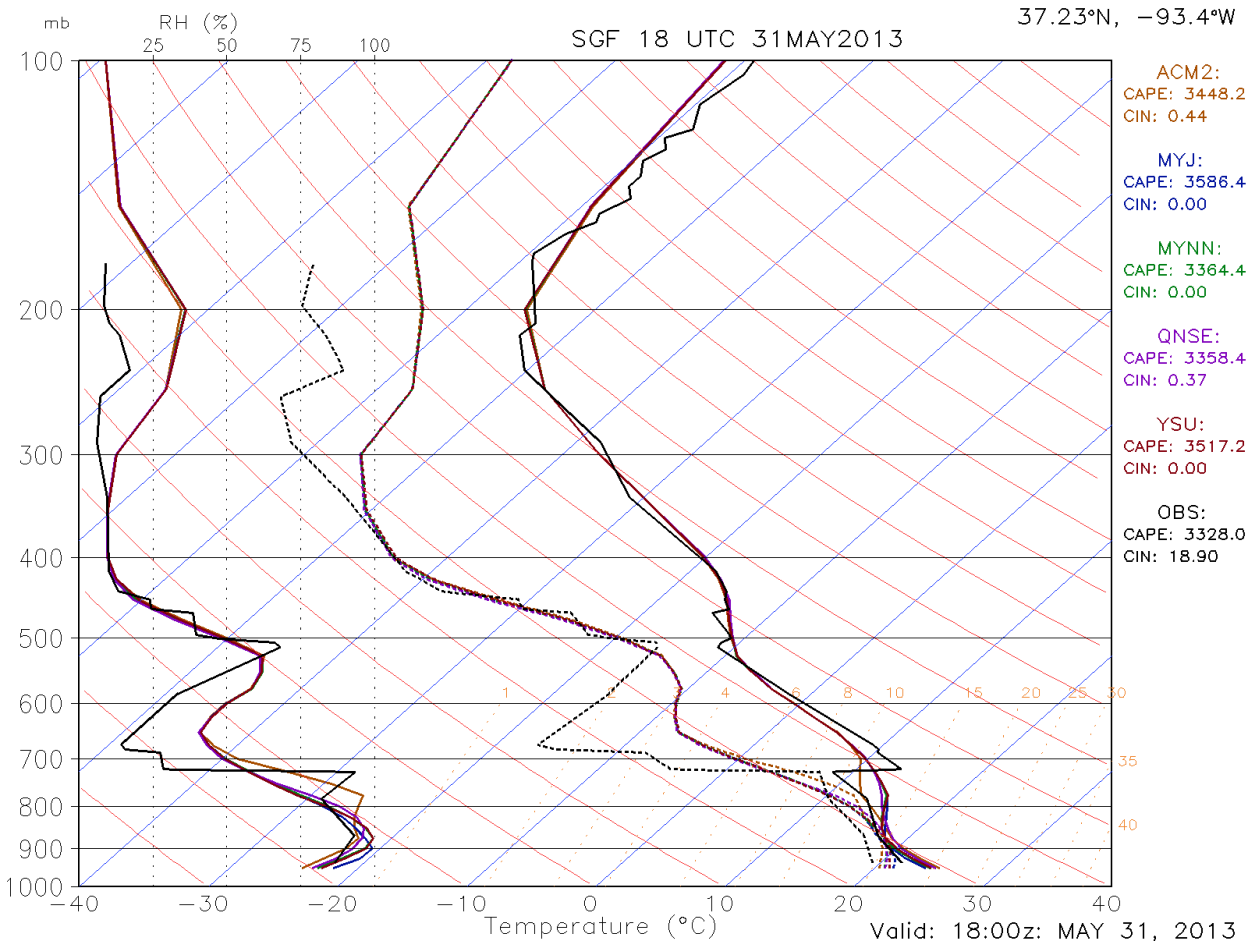

Figure 19: NWS sounding from RF10 for Springfield, MO (SGF) valid at 18 UTC on 31 May compared with the ensemble mean soundings from the same location. Left side of the plot shows RH profiles (solid) and dewpoint temperature profiles (dashed) and temperature profile (solid) are on the right.

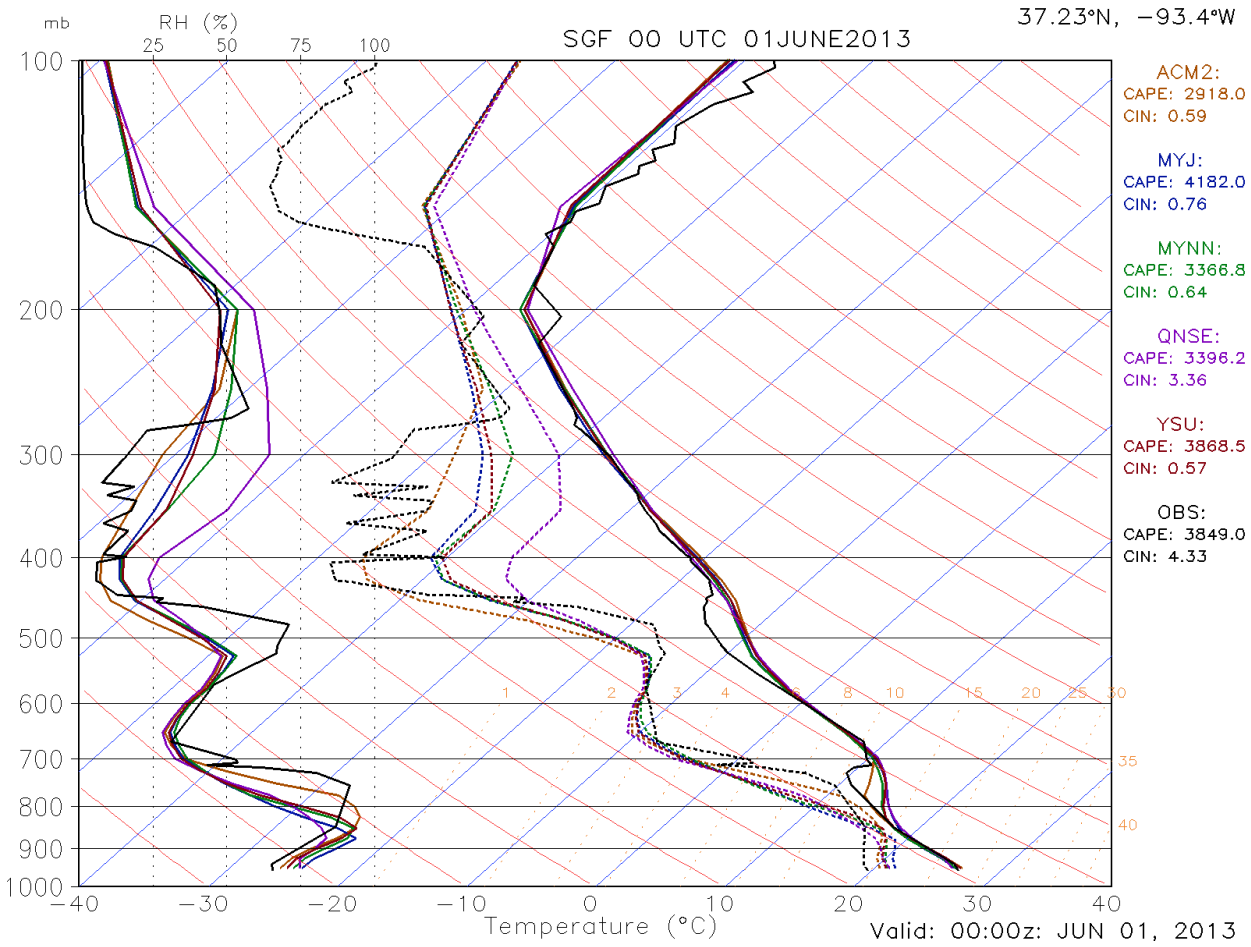

Figure 20: NWS sounding from RF10 for Springfield, MO (SGF) valid at 00 UTC on 01 June compared with the ensemble mean soundings from the same location. Left side of the plot shows RH profiles (solid) and dewpoint temperature profiles (dashed) and temperature profile (solid) are on the right.

bl_pbl_physics (PBL Scheme)	sf_sfclay_physics (Surface Layer Scheme)
YSU	MM5 similarity: Based on Monin-Obukhov with Carslon-Boland viscous sub-layer and standard similarity functions from look-up tables (sf_sfclay_physics = 1).
MYJ	Eta similarity: Used in Eta model. Based on Monin-Obukhov with Zilitinkevich thermal roughness length and standard similarity functions from look-up tables (sf_sfclay_physics = 2).
QNSE	QNSE surface layer. Quasi-Normal Scale Elimination PBL scheme's surface layer option (sf_sfclay_physics = 4).
MYNN2	MYNN surface layer. Nakanishi and Niino PBL's surface layer scheme (sf_sfclay_physics = 5).
ACM2	MM5 similarity: Based on Monin-Obukhov with Carslon-Boland viscous sub-layer and standard similarity functions from look-up tables (sf_sfclay_physics = 1).

Table 1: PBL and associated surface layer parameterizations used within WRF (v3.4.1) as the differentiation between ensembles.

		Observed	
		Yes	**No**
Forecasted	**Yes**	**(A)** - CI object observed - CI object forecasted within 100 km and 1 hour of observed	**(B)** - CI objects forecasted but not observed - CI objects forecasted and matched to observed, but not best match from ensemble
	No	**(C)** - CI object observed - CI object *NOT* forecasted within 100 km and 1 hour of observed	**(D)** - CI object *NOT* observed - CI object *NOT* forecasted ***This box is ignored***

Table 2: Description of how 2x2 contingency table was filled for purpose of computing forecast skill metrics.

		Contingency Table Results Ens. Mean | Ens. Standard Deviation					
Case	**PBL**	**A (Hits)**		**B (False Alarms)**		**C (Misses)**	
RF4	**ACM2**	144.933	8.082	158.933	16.229	**88.067**	8.082
	MYJ	**170.533**	**8.476**	**276.333**	23.624	62.467	**8.476**
	MYNN	162.4	7.383	219.867	21.705	70.6	7.383
	QNSE	157.867	6.869	202.6	22.353	75.133	6.869
	YSU	163.267	6.202	212.167	**23.95**	69.733	6.202
RF10	**ACM2**	38.2	**5.357**	31.167	7.776	**33.8**	**5.357**
	MYJ	49.367	4.997	**101.8**	**19.099**	22.633	4.997
	MYNN	42.5	5.123	49.567	11.732	29.5	5.123
	QNSE	**49.467**	3.784	85.3667	16.664	22.533	3.784
	YSU	47.4	3.929	62.067	13.818	24.6	3.929
RF12	**ACM2**	81.967	7.167	143.167	16.141	**94.033**	7.167
	MYJ	**105.567**	9.722	**253.767**	**23.944**	70.433	9.722
	MYNN	97.9	**9.857**	197.767	18.049	78.1	**9.857**
	QNSE	100.3	9.353	220.737	19.616	75.7	9.353
	YSU	95.633	8.927	196.433	23.507	80.367	8.927

Table 3: Ensemble mean and standard deviation as computed for sections A, B, and C of the 2x2 contingency table. Darker (lighter) shading and bold (underlined) number corresponds to maximum (minimum) value from the five PBL ensembles for each section of each case.

REFERENCES

Anderson, J. L., 2001: An ensemble adjustment Kalman filter for data assimilation. Mon. Wea. Rev., 129, 2884–2903.

Anderson, J. L., 2003: A local least squares framework for ensemble filtering. Mon. Wea. Rev., 131, 634–642.

Anderson, J. L., T. Hoar, K. Raeder, H. Liu, N. Collins, R. Torn, and A. Arellano, 2009: The Data Assimilation Research Testbed: A community facility. Bull. Amer. Meteor. Soc., 90, 1283–1296.

Barker, D. M., 2005: Southern high-latitude ensemble data assimilation in the Antarctic Mesoscale Prediction System. Mon. Wea. Rev., 133, 3431–3449.

Barker, D. M., and Coauthors, 2012: The Weather Research and Forecasting Model's Community Variational/Ensemble Data Assimilation System: WRFDA. Bull. Amer. Meteor. Soc., 93, 831–843.

Baxter, M.A., 2011: Impacts of Very Small Initial Condition Errors on Mesoscale Aspects of Two Cyclones, Electronic Journal of Operational Meteorology, published online as 2011-EJ1.

Benjamin, S. G., and Coauthors, (2015): A North American Hourly Assimilation and Model Forecast Cycle: The Rapid Refresh. "in press."

Benjamin, S. G., and Coauthors, 2004: An hourly assimilation– forecast cycle: The RUC. Mon. Wea. Rev., 132, 495–518.

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

Burgers, G., P. J. van Leeuwen, and G. Evensen, 1998: Analysis scheme in the ensemble Kalman filter. Mon. Wea. Rev., 126, 1719–1724.

Chen, F., and J. Dudhia, 2001: Coupling an advanced land surface– hydrology model with the Penn State–NCAR MM5 modeling system. Part I: Model description and implementation. Mon. Wea. Rev., 129, 569–585.

Ching, J., R. Rotunno, M. A. LeMone, A. Martilli, B. Kosovic, P. A. Jimenez, and J. Dudhia, 2014: Convectively induced secondary circulations in fine-grid mesoscale numerical weather prediction models. Mon. Wea. Rev., 142, 3284– 3302.

Chou, M.-D., and M. J. Suarez, 1994: An efficient thermal infrared radiation parameterization for use in general circulation models. NASA Tech. Memo. 104606, 3, 85 pp.

Clark A. J., M. C. Coniglio, B. E. Coffer, G. Thompson, M.Xue, and F. Kong, 2015a: Sensitivity of 24-h Forecast Dryline Position and Structure to Boundary Layer Parameterizations in Convection-Allowing WRF Model Simulations. Wea. Forecasting, 30, 613–638.

Clark, A. J., and Coauthors, 2015b: Experimental Forecast Program Spring Experiment 2015: Program overview and operations plan. NOAA/NSSL/SPC, 24 pp.

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, 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, 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.

Cohen A.E., S. M. Cavallo, M. C. Coniglio, and H. E. Brooks, 2015: A Review of Planetary Boundary Layer Parameterization Schemes and Their Sensitivity in Simulating Southeastern U.S. Cold Season Severe Weather Environments. Wea. Forecasting, 30, 591–612.

Coniglio M. C., J. Correia Jr., P. T. Marsh, and F. Kong, 2013: Verification of Convection-Allowing WRF Model Forecasts of the Planetary Boundary Layer Using Sounding Observations. Wea. Forecasting, 28, 842–862.

Crook, A. N., 1996: Sensitivity of Moist Convection Forced by Boundary Layer Processes to Low-Level Thermodynamic Fields. Mon. Wea. Rev., 124, 1767–1785.

Done, J., C. Davis, and M. Weisman, 2004: The next generation of NWP: Explicit forecasts of convection using the Weather Research and Forecasting (WRF) model. Atmos. Sci. Lett., 5, 6. 110–117.

Doswell, C. A., III, 1987: The distinction between large-scale and mesoscale contribution to severe convection: A case study example. Wea. Forecasting, 2, 3–16.

Doswell, C. A., III, 2004: Weather forecasting by humans—Heuristics and decision making. Wea. Forecasting, 19, 1115–1126, doi:10.1175/ WAF-821.1.

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

Ek, M. B., 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 National Centers for Environmental Prediction operational mesoscale Eta Model. J. Geophys. Res., 108, 8851,

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J. Geophys. Res., 99, 10143–10162.

Fowle, M. A. and P. J. Roebber, 2003: Short-range (0-48 h) numerical prediction of convective occurrence, mode and location. Wea. Forecasting, 18, 782-794.

Gold, D. A., and J. W. Nielsen-Gammon, 2008: Potential vorticity diagnosis of the severe convective regime. Part III: The Hesston tornado outbreak. Mon. Wea. Rev., 136, 1593–1611

Gremillion, M. S., and R. E. Orville, 1999: Thunderstorm characteristics of cloud-to-ground lightning at the Kennedy Space Center, Florida: A study of lightning initiation signatures as indicated by the WSR-88D. Wea. Forecasting, 14, 640–649.

Houston, A.L, and Niyogi, D., 2007: The Sensitivity of Convective Initiation to the Lapse Rate of the Active Cloud-Bearing Layer. Mon. Wea. Rev., 135, 3013–3032.

Houtekamer, P. L., and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. Mon. Wea. Rev., 126, 796–811.

Hu, X.-M., J. W. Nielsen-Gammon, and F. Zhang, 2010: Evaluation of three planetary boundary layer schemes in the WRF model. J. Appl. Meteor. Climatol., 49, 1831–1843.

Jonathan E. Pleim, 2007: A Combined Local and Nonlocal Closure Model for the Atmospheric Boundary Layer. Part I: Model Description and Testing. J. Appl. Meteor. Climatol., 46, 1383–1395.

Jorgensen, D. P., and T. M. Weckwerth, 2003: Forcing and organization of convective systems. Radar and Atmospheric Science: A Collection of Essays in Honor of David Atlas, Meteor. Monogr., No. 52, Amer. Meteor. Soc., 75–103.

Kain, J. S., and Coauthors, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. Wea. Forecasting, 23, 931–952.

Kain, J. S., and coauthors, 2013: A feasibility study for probabilistic convection initiation forecasts based on explicit numerical guidance. Bull. Amer. Meteor. Soc., in press.

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.

Kong, F., and Coauthors, 2009: A real-time storm-scale ensemble forecast system: 2009 Spring Experiment. Preprints, 23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction, Omaha, NE, Amer. Meteor. Soc., 16A.3.

Lakshmanan, V., 2012: Automating the Analysis of Spatial Grids: A Practical Guide to Data Mining Geospatial Images for Human and Environmental Applications. Springer, 2012.

Lakshmanan, V., and T. W. Humphrey, 2013: A mapreduce technique to mosaic continental-scale weather radar data in real-time, IEEE J. of Select Topics in Appl. Earth Obs. and Remote Sensing, in Review.

Lakshmanan, V., and T. Smith, 2009: Data mining storm attributes from spatial grids, J. Ocea. and Atmos. Tech., vol. 26, no. 11, pp. 2353-2365.

Lakshmanan, V., and T. Smith, 2010: An objective method of evaluating and devising storm tracking algorithms, Wea. and Forecasting, pp. 721-729, vol. 29 no. 3.

Lakshmanan, V., K. Hondl, and R. Rabin, 2009: An efficient, general-purpose technique for identifying storm cells in geospatial images, J. Ocean. Atmos. Tech., vol. 26, , no. 3, pp. 523-37.

Lakshmanan, V., R. Rabin, and V. DeBrunner, 2013: Multiscale storm identification and forecast, J. Atm. Res., vol. 67, pp. 367-380.

Lakshmanan, V., T. Smith, K. Hondl, G. J. Stumpf, and A. Witt, 2006: A real-time, three dimensional, rapidly updating, heterogeneous radar merger technique for reflectivity, velocity and derived products, Weather and Forecasting, vol. 21, no. 5, pp. 802-823.

Lakshmanan, V., T. Smith, G. J. Stumpf, and K. Hondl, 2007: The warning decision support system - integrated information, Weather and Forecasting, vol. 22, no. 3, pp. 596-612.

Lilly, D. K., 1990: Numerical prediction of thunderstorms—Has its time come? Quart. J. Roy. Meteor. Soc., 116, 779–798.

Lock, N. A. and A. L. Houston, (2015), Spatiotemporal distribution of thunderstorm initiation in the US Great Plains from 2005 to 2007. Int. J. Climatol., 35: 4047–4056.

Markowski, P., and C. Hannon, 2006: Multiple-Doppler radar observations of the evolution of vorticity extrema in a convective boundary layer. Mon. Wea. Rev., 134, 355–374.

Markowski, P., and Y. Richardson, 2010: "Chapter 4: The Boundary Layer. "Mesoscale Meteorology in Midlatitudes. Chichester: Wiley-Blackwell, 397 pp.

Mercer, A. E., C. M. Shafer, C. A. Doswell III, L. M. Leslie, and M. B. Richman, 2012: Synoptic composites of tornadic and nontornadic outbreaks. Mon. Wea. Rev., 140, 2590-2608.

Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmosphere: RRTM, a validated correlated-k model for the longwave. J. Geophys. Res., 102, 16663–16682.

Murphey, H. V., R. M. Wakimoto, C. Flamant, and D. E. Kingsmill, 2006: Dryline on 19 June 2002 during IHOP. Part I: Airborne Doppler and LEANDRE II analyses of the thin line structure and convection initiation. Mon. Wea. Rev., 134, 406–430.

Nakanishi M., H. Niino, 2009. Development of an improved turbulence closure model for the atmospheric boundary layer. J. Meteorol. Soc. Jpn 87: 895-912.

Roebber, P. J., 2009: Visualizing Multiple Measures of Forecast Quality. Wea. Forecasting, 24, 601–608.

Roebber, P. J., D. M. Schultz, and R. Romero, 2002: Synoptic regulation of the 3 May 1999 tornado outbreak. Wea. Forecasting, 17, 399-429.

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

Romine, G. S., C. S. Schwartz, C. Snyder, J. Anderson, and M. Weisman, 2013: Model bias in a continuously cycled assimilation system and its influence on convection-permitting forecasts. Mon. Wea. Rev., 141, 1263–1284.

Schumacher R. S., 2015: Resolution Dependence of Initiation and Upscale Growth of Deep Convection in Convection-Allowing Forecasts of the 31 May–1 June 2013 Supercell and MCS. Mon. Wea. Rev., 143, 4331–4354.

Schumann, M. R., and P. J. Roebber, 2010: The influence of synoptic forcing on convective morphology: An observational analysis. Mon. Wea. Rev., 138, 463-474.

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

Schwartz, C. S., Z. Liu, and X.-Y. Huang, 2015a: Sensitivity of limited-area hybrid variational-ensemble analyses and forecasts to ensemble perturbation resolution. Mon. Wea. Rev., 143, 3454– 3477.

Shafer, C. M., A. E. Mercer, C. A. Doswell III, M. B. Richman, and L. M. Leslie, 2009: Evaluation of WRF forecasts of tornadic and nontornadic outbreaks when initialized with synoptic scale input. Mon. Wea. Rev., 137, 1250–1271.

Skamarock, W. C, and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN– 4751STR, 113 pp. [Available online at http://www.mmm.ucar. edu/wrf/users/docs/arw_v3.pdf.]

Skinner, P. S., L. J. Wicker, D. M. Wheatley, Knopfmeier, K. H., 2016: Application of two spatial verification methods to ensemble forecasts of low-level rotation. Wea. Forecasting, in press.

Song-You Hong and Hua-Lu Pan, 1996: Nonlocal Boundary Layer Vertical Diffusion in a Medium-Range Forecast Model. Mon. Wea. Rev., 124, 2322–2339.

Stensrud, D. J., 2007: Parameterizations Schemes: Keys to Understanding Numerical Weather Prediction Models. Cambridge University Press, 459 pp.

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.

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

Tiedtke, M., 1989: A comprehensive mass flux scheme for cumulus parameterization in large-scale models. Mon. Wea. Rev., 117, 1779–1800.

Torn, R. D., G. J. Hakim, and C. Snyder, 2006: Boundary conditions for limited-area Ensemble Kalman filters. Mon. Wea. Rev., 134, 2490–2502.

Trier, S.B., and Coauthors, 2015: Mesoscale Thermodynamic Influences on Convection Initiation near a Surface Dryline in a Convection-Permitting Ensemble. Mon. Wea. Rev., 143, 3726–3753.

Van Klooster, S. L., and Roebber, P. J., 2009: Surface-Based Convective Potential in the Contiguous United States in a Business-as-Usual Future Climate. J. Climate, 22, 3317–3330

Weckwerth, T. M., and D. B. Parsons, 2006: A review of convection initiation and motivation for IHOP_2002. Mon. Wea. Rev., 134, 5–22.

Weckwerth, T. M., H. V. Murphey, C. Flamant, J. Goldstein, and C. R. Pettet, 2008: An observational study of convection initiation on 12 June 2002 during IHOP_2002. Mon. Wea. Rev., 136, 2283–2304.

Weisman M. L., and Coauthors, 2015: The Mesoscale Predictability Experiment (MPEX). Bull. Amer. Meteor. Soc., 96, 2127–2149.

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.

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.

Weiss, S., and Coauthors, 2011: Experimental Forecast Program Spring Experiment 2011: Program overview and operations plan. NOAA/NSSL/SPC, 62 pp.

Wilks, D. S., 1995: Statistical Methods in Atmospheric Sciences: An Introduction. Academic Press, 500 pp.

Wyngaard, J. C., 2004: Toward Numerical Modeling in the "Terra Incognita". J. Atmos. Sci., 61, 1816–1826.

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

Zaviša I. Janjić, 1994: The Step-Mountain Eta Coordinate Model: Further Developments of the Convection, Viscous Sublayer, and Turbulence Closure Schemes. Mon. Wea. Rev., 122, 927–945.

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

Zhang, F., C. Snyder, and R. Rotunno, 2002: Mesoscale predictability of the ''surprise'' snowstorm of 24–25 January 2000. Mon. Wea. Rev., 130, 1617–1632.

Ziegler, C. L., T. J. Lee, and R. A. Pielke, 1997: Convective initiation at the dryline: a modeling study. Mon. Wea. Rev., 125, 1001–1026.