Exploring Model Error through Post-processing and an Ensemble Kalman Filter on Fire Weather Days

A Dissertation Presented

by

Michael J. Erickson

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Marine and Atmospheric Science

Stony Brook University

May 2015

Stony Brook University

The Graduate School

Michael J. Erickson

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Dr. Brian A. Colle – Dissertation Advisor
Professor – School of Marine and Atmospheric Sciences

Dr. Minghua Zhang - Chairperson of Defense
Professor – School of Marine and Atmospheric Sciences

Dr. Edmund K.M. Chang
Professor – School of Marine and Atmospheric Sciences

Dr. Ryan Torn
Assistant Professor – Atmospheric and Environmental Sciences
University at Albany – State University of New York

Dr. Joseph J. Charney
Research Meteorologist – Climate, Fire and Carbon Cycle Sciences
United States Forest Service

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Exploring Model Error through Post-processing and an Ensemble Kalman Filter on Fire Weather Days

by

Michael Jon Erickson

Doctor of Philosophy

in

Marine and Atmospheric Science

Stony Brook University

2015

The proliferation of coupling atmospheric ensemble data to models in other related fields requires a priori knowledge of atmospheric ensemble biases specific to the desired application. In that spirit, this dissertation focuses on elucidating atmospheric ensemble model bias and error through a variety of different methods specific to fire weather days (FWDs) over the Northeast United States (NEUS). Other than a handful of studies that use models to predict fire indices for single fire seasons (Mölders 2008, Simpson et al. 2014), an extensive exploration of model performance specific to FWDs has not been attempted.

Two unique definitions for FWDs are proposed; one that uses pre-existing fire indices (FWD1) and another from a new statistical fire weather index (FWD2) relating fire occurrence and near-surface meteorological observations. Ensemble model verification reveals FWDs to have warmer (> 1 K), moister (~ 0.4 g kg^{-1}) and less windy (~ 1 m s^{-1}) biases than the climatological average for both FWD1 and FWD2. These biases are not restricted to the near surface but exist through the entirety of the planetary boundary layer (PBL). Furthermore, post-processing methods are more effective when previous FWDs are incorporated into the statistical training, suggesting that model bias could be related to the synoptic flow pattern.

An Ensemble Kalman Filter (EnKF) is used to explore the effectiveness of data assimilation during a period of extensive FWDs in April 2012. Model biases develop rapidly on FWDs, consistent with the FWD1 and FWD2 verification. However, the EnKF is effective at removing most biases for temperature, wind speed and specific humidity. Potential sources of error in the parameterized physics of the PBL are explored by rerunning the EnKF with simultaneous state and parameter estimation (SSPE) for two relevant parameters within the ACM2 PBL scheme. SSPE helps to reduce the cool temperature bias near the surface on FWDs, with the variability in parameter estimates exhibiting some relationship to model bias for temperature. This suggests the potential for structural model error within the ACM2 PBL scheme and could lead toward the future development of improved PBL parameterizations.

Table of Contents

List of Tables

Chapter 3 **Page**

Chapter 4

Chapter 5

List of Figures

Chapter 2 **Page**

Chapter 3

Chapter 4

Chapter 5

Chapter 6

List of Abbreviations

3D-VAR: 3-dimensional variational data assimilation
4D-VAR: 4-dimensional variational data assimilation
ACARS: Aircraft Comminucations Addressing and Reporting System
ACARSP: ACARS profiles
ACM2: Asymmetric convective model, version 2 PBL scheme
ARW: Advanced Research WRF model
ASOS: Automated Surface Observing System
AWOS: Automated Weather Observing System
BLK: Blackadar PBL scheme
BMA: Bayesian Model Averaging
BSS: Brier skill scores
CA: Cluster analysis
CB: Contingency bias
CBC: Conditional bias correction
CDF: Cumulative distribution function
CONUS: Contiguous United States
CFSR: Climate Forecast System Reanalysis
CRN: Climate research network surface observations
CSI: Critical success index
CT: Conditional training
DA: Data assimilation
DREAM: DiffeRential Evolution Adaptive Metropolis
EOF: Empirical orthogonal function
KF: Kain Fritsch cumulus scheme
EnKF: Ensemble Kalman Filter
EnSRF: Ensemble square root ensemble filter
ETS: Equitable threat score
FAR: False alarm ratio
FPI: Fire Potential Index
FWD: Fire weather day
FWI: Fire weather index
GFS: Global Forecast System
HDW-E: High density winds – experimental (satellite observations)
HIT: Hit rate
IC: Initial condition
IMS: Ice Mapping System
LBC: Lateral boundary condition
LRM: Logistic regression model
LSM: Land surface model
MADIS: Meteorological Assimilation Data Ingest System
MAE: Mean absolute error
MARINE: Buoys and ships offshore data
MCMC: Markov Chain Monte Carlo
ME: Mean error

METAR: ASOS surface observations
MESONT: Mesonet surface observations
MIXR: Mixing ratio
MM5: Mesoscale Model Verison 5
MULT-P: Multi-agency profiler observation data
MY: Mellor-Yamada PBL scheme
MYJ: Mellor-Yamada-Janjic PBL scheme
MYNN: Mellor-Yamada Nakanishi and Niino PBL scheme
NAM: North American Mesoscale model
NARR: North American Regional Reanalysis
NCAR: National Center for Environmental Research
NCDC: National Climatic Data Center
NCEP: National Centers for Environmental Prediction
NEPP: New England Pilot Project surface observations
NEUS: Northeast United States
NFDRS: National Fire Danger Rating System
NICC: Northeast Interagency Coordination Center
NMM: Nonhydrostatic Mesoscale Model
NYC: New York City
OTIS: Optimum Thermal Interpolation System
PBL: Planetary boundary layer
PC: Principal component
PCP: Cumulative daily precipitation
PDF: Probability density function
PIT: Probability integral transform
POES: Polar Orbital Environmental Satellites
PSU: Pennsylvania State University
QC: Quality control
RAP: Rapid Refresh model
RELH: Relative humidity
RELH0: 0 day lagged relative humidity
RELH1: 1 day lagged relative humidity
RELH2: 2 day lagged relative humidity
RELH20: 20 day lagged average relative humidity
RG: Relative greenness
RI: Reliability index
RMSI: Root mean squared innovation
RRTM: Rapid Radiative Transfer Model
RSM: Regional Spectral Model
RUC: Rapid Update Cycle model
SAO: Standard Aviation Observation
SBC: Sequential bias correction
SBU: Stony Brook University
SCL: Successive covariance localization
SLP: Sea level pressure
SPHU: Specific humidity

SREF: Short Range Ensemble Forecast
SSPE: Simultaneous state and parameter estimation
ST: Sequential training
stdev: Standard deviation
SST: Sea surface temperature
TAMDAR: Tropospheric Airborne Meteorological Data Reporting
TEMP: Temperature
THTA: Theta
TMPD: Dew point
US: United States
UWND: U component of the wind
VWND: V component of the wind
WFAS: Woodland Fire Assessment System
WNDS: Wind speed
WRF: Weather Research and Forecast model
WSM6: WRF Single Moment 6-class microphysics scheme
YSU: Yonsei University PBL scheme

Acknowledgments

I would like to thank my advisor, Dr. Brian Colle for his patience, guidance, and help over these many years. He showed me how to more effectively clear my mind to see that there are many useful ways in approaching a scientific question. I would like to extend a grateful thank you to my committee Dr. Minghua Zhang, Dr. Edmund Chang, Dr. Ryan Torn, and Dr. Joseph Charney for all their help, feedback, and patience. I would also like to thank the staff at SoMAS for their help and enjoyable sanity breaks from work. Many thanks go to Mark Lang and Dr. Ping Liu for their help with our local machines and system setup. I would not have been able to finish as quickly without their help.

There have also been a number of people who have kindly provided help and code from outside my committee circle and SoMAS. Thanks to Jasper Vrugt for supplying the base BMA code and Jun Du for his help with the SREF. I am extremely grateful to Fuqing Zhang at PSU for supplying me the PSU EnKF code and to several PSU students (Christopher Melhauser, Jerry Zhang, Benjamin Green, and Yue Ying) and Yonghui Weng for their gracious help.

I would like to thank my friends for all of their help and support over these long years. The sanity breaks from research provided perspective, and with that, the realization of a light at the end of the tunnel. I would like to thank my parents, Ivy and Rudy Erickson, my brothers Bill and Fred, and all of my family for their never ending support and love. I would also like to thank my lovely girlfriend, Melissa Yencho, for her support, love, strength and keeping me sane. Sometimes it is easy to get lost in the research “bubble,” but it is the personal connections and experiences we make in life that are the most important.

Chapter 1:

Introduction

1.1 Background

Wildfires represent a relatively rare high-impact forecast problem over the Northeast United States (NEUS). For instance, NEUS wildfires have resulted in 13,633 acres burned on average annually, or about 0.27% of the total acres burned for the contiguous United States (CONUS; Pollina et al. 2013). However, wildfires over the NEUS have a potentially greater societal impact than in other parts of CONUS due to the region's higher population density. For example, the August 1995 "Sunrise Fire" burned about 7000 acres in the Pine Barrens region of eastern Long Island (Hamilton and Ostapow, 2009), resulting in destroyed homes, businesses, and road closures. On 17 April 2008, a wildfire near New Paltz, NY burned over 3000 acres of the Minnewaska State Park Preserve resulting in several road closures. More recently on 9 April 2012, a wildfire burned between 1000 and 2000 acres around Manorville, NY on Long Island, disrupting traffic and threatening homes. Despite the importance of NEUS fire events, the majority of wildfire research studies over CONUS are focused in the western U.S. due to the greater abundance of large observed wildfires in that region. However, the characteristics of western U.S wildfires may not be the same as in the NEUS due to differences in terrain, climate, land-use characteristics and anthropogenic influences. More attention should be given to the characteristics and predictability of NEUS wildfires particularly with respect to meteorological influences.

Atmospheric ensemble modeling can be used to explore the meteorological contribution to wildfire predictability. Operational ensembles quantify forecast uncertainty through the use of different initial conditions, physical parameterizations, and model cores. Ideally, the spread of model solutions creates a representative sample of all possible future outcomes from which probabilistic forecasts can be derived. This attempt to capture the distribution of all potential forecasting outcomes is extremely valuable for operational forecasters and in other related fields that rely on accurate atmospheric simulations, such as air pollution and dispersion modeling (Seaman and Michelson 2000, Delle Monache et al. 2006, Yahya et al. 2015), storm surge modeling (Mel and Lionello 2014), electricity consumption forecasts (Salamanca et al. 2013), hydrologic modeling (Brown et al, 2012; Demargne et al. 2014), and fire weather efforts (Hoadley et al. 2004, Hoadley et al. 2006, Mölders 2008, Mölders 2010, Simpson et al. 2013, Simpson et al. 2014). Unfortunately, ensembles exhibit systematic biases in both their mean (Colle et al. 2003; Jones et al. 2007) and spread (Hamill and Colucci 1997, Eckel and Mass 2005) that can adversely impact operational forecasts and users who couple atmospheric ensemble data with other dynamical or statistical models. These biases can be complex and vary by model, spatially, diurnally, seasonally and even with the synoptic flow pattern. This dissertation addresses the topics of quantifying, exploring and correcting ensemble model bias conditional on fire weather days (FWDs) over the NEUS through a variety of pre-processing and post-processing methodologies.

a. Fire weather synoptic patterns over the Northeast United States

The abundance of previous fire weather studies over the NEUS focus on identifying synoptic patterns responsible for fire occurrence. Schroeder et al. (1964) examines periods of high fire danger in 14 regions of the United States (US) and associates those periods with synoptic weather patterns. They show fire danger events in the NEUS occur with four types of setups: Canadian high, Pacific high, Bermuda high and Atlantic storm. Simard et al. (1987) proposes a methodology for predicting extreme fire potential that includes the NEUS, which combines daily weather observations with components of the National Fire Danger Rating System (Bradshaw et al. 1983). Takle et al. (1994) employ the Yarnal (1993) synoptic weather classification system to identify different types of surface high and low pressure patterns associated with wildfire events in West Virginia. Barbero et al. (2014) explore the influence of inter-annual, sub-seasonal and synoptic weather on very large fires over the eastern United States and conclude that very large fires occur during coincident long-term and short-term droughts.

Pollina et al. (2013) presents a spatial and temporal climatology of major wildfire events in the NEUS and examines the associated meteorological conditions while also using the Yarnal (1993) classification technique. Pollina et al. (2013) finds a wildfire occurrence peak during April and May before green-up (leaf out). In addition, Pollina et al. (2013) finds that fire occurrence along the coastal plain is more commonly associated with a high pressure system building in from the northwest and a low pressure center off the mid-Atlantic coast. This results in a dry continental polar air mass advecting into the NEUS with aided downsloping from the Appalachians coincident with a broad region of mid-tropospheric quasigeostrophic decent. Furthermore, it is possible to have fire occurrence with a Bermuda high-like setup, provided the low level air mass originates over the continental United States than the Atlantic Ocean (Pollina et al. 2013). Fire occurrence days are likely amplified before the spring green-up period, when evapotranspiration is minimized allowing for a drier, warmer and deeper planetary boundary layer (PBL).

b. Mesoscale modeling of fire weather events

To elucidate understanding of the meteorological contribution to extreme wildfire events, case studies of large wildfires have been performed using mesoscale models. Simulations of the 1300 acre New Jersey “Double Trouble” fire on 2 June 2002 reveal a significant surface drying resulting from a combination of a deepening PBL and downward transport of dry, high momentum air from aloft (Kaplan et al. 2008, Charney and Keyser 2010). In addition, high-resolution atmospheric simulations are commonplace outside of the NEUS and have been performed for significant individual wildfire events in Colorado (Johnson et al. 2014), Australia (Mills 2005a,b; Mills 2008a,b, Fox-Hughes 2012, Engel et al. 2013), New Zealand (Simpson et al. 2013), and Iran (Mofidi et al. 2015). In general, the suspected meteorological influences on fire development vary somewhat between cases but are usually enhanced by topography and sometimes associated with dry air intrusions from aloft.

Atmospheric model simulations of a longer duration (~ 1 month or greater) have been employed during a fire season for the Pacific Northwest (Hoadley et al. 2004, 2006), Alaska (Mölders 2008, 2010), and New Zealand (Simpson et al. 2014a, 2014b). These longer duration simulations allow for model bias to be explored in greater detail. Using the Pennsylvania State University (PSU) / National Center for Environmental Research (NCAR) Mesoscale Model (MM5), Hoadley et al. (2004) notes a negative surface temperature bias (< -3 ^{o}C in all cases) and positive surface relative humidity bias (> 9.8% in all cases) during the daytime, resulting in the underestimation of National Fire Danger Rating System (NFDRS) indices (Hoadley et al. 2006). Mölders et al. (2008) finds similar biases compared to Hoadley et al. (2006) while using the Weather Research and Forecast (WRF; Skamarock et al. 2008) model for average daily maximum temperature (- 4.1 ^{o}C) and average daily minimum relative humidity (12%). More recently, Simpson et al. (2014a, 2014b) verify WRF performance for the 2009-10 New Zealand fire season and note a systematic WRF underprediction on high-end fire weather days of temperature (averaging -1.7 ^{o}C) and relative humidity (averaging 1 %) accompanied by an overprediction of wind speed (averaging 1.4 m s^{-1}) and precipitation (averaging 0.35 mm). These combined biases result in a general underprediction of fire threat, but an overprediction of extreme fire threat days, and emphasize the importance of careful post-processing.

c. *Ensemble post-processing approaches*

As mentioned earlier, post-processing methods are effective at removing the average ensemble bias; with examples including the running mean bias removal (Stensrud and Yussouf 2003; Stensrud and Yussouf 2005), the multivariate regression model (Glahn et al. 2009), the Kalman filter (Libonati et al. 2008; Müller 2011), gene expression programming (Bakhshaii and Stull 2009), binning and correcting forecasts by value (Hamill and Colucci 1998; Gallus and Segal 2004; Gallus et al. 2007; Stensrud and Yussouf 2007), and the correction of model bias in a gridded format (Eckel and Mass 2005; Gallus et al. 2007; Stensrud and Yussouf 2007; Yulia 2007; Mass et al. 2008; Glahn et al. 2009). However, PBL physics errors contribute significantly to lower atmospheric model error (Pleim 2007; Hu et al. 2010a; Nielsen-Gammon et al. 2010), suggesting that bias correction may be important when forecasting phenomena sensitive to near-surface meteorological conditions. In addition, most ensembles exhibit insufficient spread (i.e. are underdispersed; Eckel and Mass 2005; Jones et al. 2007; Erickson et al. 2012), which is worse in the lower atmosphere (Hamill and Colucci 1997) due to heavy reliance on parameterized physics. There is a large history of ensemble post-processing techniques designed to improve model spread and probabilistic scores; including the prior rank histogram method (Hamill and Colucci 1998), Bayesian Processor of Forecast (Krzysztofowicz and Evans 2008), ensemble reforecasts (Hamill et al. 2006), Bayesian Model Averaging (BMA, Raftery et al. 2005), ensemble “dressing” techniques (Wang and Bishop 2005), extended logistic regression (Wilks 2009), Ensemble Model Output Statistics (EMOS, Gneiting et al. 2005) and calibrated error sampling/randomly calibrated resampling (Eckel et al. 2012).

Ensemble post-processing is generally effective at correcting ensemble model biases in the mean and spread with a large enough sample size (i.e. Hamill et al. 2006). However, the variability of model biases conditional on weather regime or synoptic flow pattern are not clear given the inherent complexities in defining a “weather regime” over a finite area. As a result, assuming temporally invariant model biases may be inappropriate for specific operational applications such as creating fire weather forecasts during anomalous events.

A few recent studies explore the topic of regime-based model post-processing. Greybush et al. (2008) finds varying optimal ensemble member weights dependent on the synoptic regime over the Pacific Northwest. An analog technique developed in Hamill et al. (2006) searches for historical days over the United States with the smallest local root mean squared difference to the current ensemble mean forecast using Global Forecast System (GFS) reforecasts. Delle Monache et al. (2011) use an analog method that matches previously modeled WRF forecasts with the current forecast to improve 10-m wind speed predictions by 20% to 25% over the western United States compared to non-analog bias correction methods. However, any quantification of FWD ensemble model bias (mean and spread) and how that relates to the climatological average model bias has not been explored.

d. Data assimilation with the Ensemble Kalman Filter

The main goal with data assimilation is to statistically combine assimilated observations and a background state (in this case a short-term model forecast) to create the optimal analysis. Typically data assimilation techniques are used to create historical reanalysis products or initial condition fields for model simulations. There are a variety of methods for data assimilation including nudging (Stauffer et al. 1991), variational data assimilation (3D-VAR and 4D-VAR; Talagrand and Courtier 1987) and the Ensemble Kalman Filter (EnKF; Evenson 1994). Both variational and ensemble filtering methods have shown encouraging results at creating an improved analysis state (Kalnay et al. 2007). Both 4D-VAR and the EnKF consider flow-dependent statistics, whereas 3D-VAR uses stationary background covariances. This allows EnKF to be competitive with 4D-VAR (Kalnay et al. 2007), while producing consistently better results than 3D-VAR (Anderson et al. 2009). Since EnKF derives statistics from the ensemble covariances, it is significantly easier to set up and implement than 4D-VAR (Anderson et al. 2009), which requires the derivation of a tangent linear model and backward integrations.

Although the EnKF is competitive with other data assimilation techniques, it is still sensitive to sampling errors when calculating ensemble statistics such as model state variable covariances. These sources of sampling error result in variance estimates that are too low (Anderson et al. 2009). The source of these errors is difficult to isolate, but generally include a lack of covariance in model state variables (Whitaker et al. 2008), a lack of ensemble spread (Bonavita et al. 2008; Li et al. 2009) and model biases (Bonavita et al. 2008; Mass et al. 2008). Insufficient variance causes artificial overconfidence of the ensemble in the EnKF, which can result in the filter drifting away from reality (i.e. filter divergence). Additional causes of sampling error come from spurious long distance correlations in model state variables, which are corrected by a variety of localization techniques (Anderson and Anderson 1999; Gaspari and Cohn 1999; Mitchell and Houtekamer 2000; Ott et al. 2004). The constraining radius for localization can vary depending on the localization method, assimilated observations, and resolution of the ensemble. Therefore, care must be taken when running an EnKF since the filter may require some tuning depending on its intended use.

Finally, the increasing popularity of ensemble filters has led to the growth of simultaneous state and parameter estimation (SSPE). SSPE allows parameters within the derived model physics schemes to be part of the model state in conjunction with conventional variables (Annan and Hargreaves 2004). SSPE is a way to account for model error in ensemble filters while adjusting incorrectly specified parameters (Nielson-Gammon et al. 2010). SSPE has been attempted within idealized low-dimensional models (Annan and Hargreaves 2004; Aksoy et al. 2006a; Stroud and Bengtsson 2007), for microphysical parameters in more complex models (Tong and Xue 2008a; Tong and Xue 2008b; Jung et al. 2010; Xue et al. 2010), soil moisture modeling (Dumedah and Walker 2014), coupled general circulation models (Liu et al. 2014a, b) and to estimate PBL parameters (Aksoy et al. 2006b; Hu et al. 2010b). In general, SSPE produces encouraging results that can improve EnKF performance. For instance, Hu et al. (2010b) shows that SSPE with two PBL parameters has the potential to correct both synoptic cold air advection and a lack of vertical mixing near the surface. The estimation of the optimal parameters within a PBL scheme may allow for significant improvements in vertical mixing representation within and above the PBL (Nielsen-Gammon et al. 2010). However, this approach does not correct any potential misspecification of a term's functional form (i.e. structural model error) within the parameterization scheme (Golaz et al. 2007). In this respect, developing more accurate parameterization schemes is a more direct and better pathway (Xue et al. 2010). Nonetheless, parameter estimation could be used to reveal structural model error parameterizations by examining the variability of the parameter values (Golaz et al. 2007; Hu et al. 2010b). Exploring SSPE on FWDs or as a function of synoptic flow pattern has never been attempted. The impact of SSPE on the parameterized PBL scheme for FWDs could be significant given the presence of a deep and dry PBL (Charney and Keyser 2010, Pollina et al. 2013).

1.2 Research goals and approach

While long duration atmospheric simulations during fire seasons note the presence of atmospheric model biases that can degrade fire index forecasts (Hoadley et al. 2004, Hoadley et al. 2006, Mölders et al. 2008, Mölders et al. 2010, Simpson et al. 2014 a,b), no study attempts to correct these biases by using post-processing. This is particularly worrying because systematic near-surface cool and wet biases (except in Simpson et al. 2014 a, b, which note a slight dry bias) found in both the MM5 and WRF simulations would likely result in the underestimation of common fire threat indices such as the NFDRS on FWDs. Furthermore, the average model bias over a fire season is not necessarily the average model bias on an FWD, since there will likely be some daily synoptic variability. As a result, any potential differences in model performance between FWDs and non-FWDs have not been studied. Knowing these differences is critical, since standard model post-processing may be insufficient if conditional biases apply on FWDs. Similarly, the distribution of meteorological conditions common to FWDs may deviate significantly from the climatological average. As a result, FWD model biases may differ since they focus on an anomalous region of the observed probability density function. Conditional model biases could have adverse effects as dynamical coupling between atmospheric ensembles and wildland fire models become more common (i.e. WRF-Fire; Coen et al. 2013).

Elucidating ensemble model biases on fire weather days are further complicated by the lack of a strict definition for what constitutes a FWD. As mentioned earlier, meteorology is recognized as a major influence on fire initiation and behavior. However, for spatial scales representative of the NEUS, the relevant physical processes involving fire initiation are not well understood (Potter 2012) and representations of physical processes are often used in conjunction with statistical methods that link meteorological conditions and fire activity. Numerous indices have been constructed that consider the local meteorology and other potentially important factors to wildfire formation and development such as fuel conditions, fuel abundance, landuse category, and topography. A few of these indices include the McArthur Forest Fire Danger Index (Luke and McArthur, 1978), the United States NFDRS (Deeming et al. 1978), the Fosberg Index (Fosberg, 1978), the Canadian Forest Fire Weather Index System (Van Wagner, 1987), and the Haines Index (Haines, 1988). A quick and potentially beneficial method would be to use these already existing indices to subset FWDs. Although this approach is subjective, it could be constrained with trial and error and logical thresholding of index or category values.

A potential caveat of using preexisting indices to subset FWDs is with the lack of a known quantitative relationship (statistical or dynamical) between the meteorological representations in the index and fire occurrence over the NEUS. In other words, it might be impossible to quantitatively understand what an index or category value represents with respect to the meteorological impacts on fire occurrence or behavior. Despite these caveats, the use of preexisting indices to subset FWDs will still carry over some benefit since the meteorological conditions responsible for fire threat are likely to be captured even if they can't be statistically quantified.

In addition to using preexisting indices, it would likely be advantageous to develop a complimentary fire weather index (FWI) with a less complex statistical form based solely on near-surface meteorological variables. While previous studies (Yarnal 1993, Takle et al. 1994, Pollina et al. 2013, Barbero et al. 2014) provide a context for the meteorological variables over the NEUS that likely affect wildfire occurrence, a comprehensive study that constructs and evaluates a longer term predictive statistical relationship between fire activity and near-surface meteorological variables has not been conducted.

Finally, data assimilation with an EnKF has never been implemented or evaluated for an extended period of elevated FWDs. The flow-dependent information obtained from the background error covariances could propagate the benefit of the assimilated observations vertically within the PBL and across model state variables. Therefore, EnKF analyses could improve model simulations on FWDs after the filter has had a suitable time to spin-up. The EnKF is also quite versatile and can be adjusted to run ensemble members with multiple PBL schemes. This adaption could improve the filter performance by increasing ensemble spread in the PBL. Furthermore, SSPE of PBL parameters similar to Hu et al. (2010b) could be implemented to both improve filter performance and explore structural model errors by evaluating variations in the parameter values over time. FWDs provide a unique test bed for SSPE with the EnKF since they feature a less complex atmosphere consisting of a deep well-mixed PBL without the complications of cloud microphysical parameterizations.

This dissertation addresses the following motivational questions:

- Can a predictive statistical FWI be formulated using meteorological observations and fire occurrence data over the NEUS?
- How do ensemble model biases (mean and spread) differ between FWDs and the climatological average and how effective is post-processing at removing model bias?
- Are similar ensemble model biases captured when using two different FWD methodologies?
- How sensitive is BMA performance to the subset of members selected from the total ensemble?
- Does the EnKF perform similarity on FWDs compared to non-FWDs and what is the impact of assimilating observations on model performance?
- Is there any benefit to using a multi-PBL ensemble on FWDs?
- Does SSPE with the EnKF improve filter performance and is there significant evidence of structural model error in the PBL parameterized physics?

These questions will be addressed using a combination of available observational data, existing fire indices, archived ensemble model data, and coupling the WRF model with an EnKF. Chapter 2 describes two different methodologies for capturing FWDs; one that uses preexisting fire indices (FWD1) and another based off a new statistical FWI that uses meteorological variables and fire occurrence data (FWD2). Chapter 3 describes the ensemble model verification and post-processing using FWD1, while gridded verification and post-processing with FWD2 days are explored in chapter 4. Chapter 5 describes the implementation and fine tuning of the EnKF over the NEUS to explore FWDs for a control run, a multi-PBL run and a run with SSPE. Chapter 6 provides a summary of all the results with additional discussion on future work topics.

Chapter 2:

Identifying Fire Weather Days Through the use of Existing Fire Indices and the Development of a new Fire Weather Index

2.1 Introduction

As mentioned in section 1.2, there is no formal definition for what meteorological conditions represent a fire weather day (FWD). Therefore, preexisting indices can be used that incorporate some influence of fire weather or new indices can be developed that quantify a statistical relationship between meteorology and fire threat. Fortunately, a proliferation of meteorological reanalysis datasets [i.e. NCEP (National Centers for Environmental Prediction), NARR (North American Regional Reanalysis), CFSR (Climate Forecast System Reanalysis)] and abundant access to meteorological station observations allow for the examination of statistical relationships between fire activity and the overlying atmosphere. This dissertation makes use of both approaches by defining two different FWDs; one using preexisting indices and another by creating a new statistical fire weather index (FWI). In later chapters, these techniques will be used to subset FWDs, allowing for conditional biases to be quantified, explored and corrected.

2.2 Data and methods

a. Using previous indices to identify FWDs (FWD1)

The first classification of fire weather days (FWD1) employs the Fire Potential Index (FPI) from the Woodland Fire Assessment System (WFAS; Burgan et al 1998) when available and the National Fire Danger Rating System (NFDRS; Deeming et al. 1972) when FPI is unavailable. The FPI considers relative greenness (RG) maps derived from the state of current vegetation, fuel moisture and the daily weather conditions to determine fire threat on a scale from 0 to 100 (Preisler et al. 2009). The NFDRS rates fire danger based on a complex combination of physical, statistical, and engineering equations that consider the local meteorology, terrain height, and fuel conditions. The FPI is similar to the Energy Release Component within the NFDRS in that both indices are moisture related and neither considers the effect of wind. For this dissertation, an FPI value of 50 or an NFDRS rating of "high" is considered to be a FWD. A FWD must have at least 10% of the inner domain (Fig. 2.1) experiencing a high category, while the remainder of the inner domain is in the moderate category (FPI 21-50). In this manner, at least localized fire weather events within the domain are captured while eliminating days with large spatial discontinuities in weather (i.e. rain on the edge of the domain). This technique of thresholding is similar to the NFDRS classification of fire threat days in Pollina et al. (2013). Both the FPI and NFDRS are available daily and their values are collected between 2007 and 2009. FWD1 is not an index because there are only two possible binomial outcomes (i.e. there is a FWD or there is not), representing the entirety of the inner domain in Figure 2.1.

b. Development of a fire weather index to identify FWDs (FWD2)

The second methodology for determining FWDs (hereafter referred to as FWD2) uses a predictive statistical regression technique to relate meteorological variables to fire occurrence data over the NEUS. Although meteorological variables are known to play a significant role in fire threat (Potter 2012), it is not obvious which predictors to include in the optimal regression model. Furthermore, the configuration of the optimal regression model is unknown. As a result, a statistical formulation is proposed and multiple atmospheric variables are tested. The output from the optimal predictive statistical regression model is used to create a FWI. Finally, the FWI is used to determine the FWDs (a binomial variable).

1 Atmospheric variable selection

Hourly Automated Surface Observing System (ASOS) weather observations between 1999 and 2008 are compared with fire occurrence data to test different regression configurations. Potentially important predictors of daily fire occurrence are determined by comparing the difference in near surface weather conditions on observed wildfire days to their climatological average. Regional sensitivity to potential predictors is assessed by considering two separate domains: domain 1 (D1) includes Long Island, Connecticut, and the New York City (NYC) metropolitan region, while domain 2 (D2) includes most of New Jersey, far eastern Pennsylvania and far northeastern Delaware (Fig. 2.2). Domains are employed instead of single stations in order to assess the statistical probability of observed wildfire occurrence during favorable meteorological conditions. Therefore, this dissertation assumes fire initiations are ubiquitous and that quantifiable anomalous meteorological conditions are significant predictors in their detection (in this case referred to as fire occurrence). The domain size encompasses both an urban population and suburban sprawl while being sufficiently small to keep daily indicators of meteorological conditions roughly homogenous throughout the region (i.e. ~30,000 km^2 in the NEUS).

Pollina et al. (2013) compiled the Northeast Interagency Coordination Center (NICC) and the Pennsylvania Bureau of Forestry wildfire occurrence data from 1999 to 2008 to develop a NEUS climatology of wildfire events, although they only presented results for major wildfire events (i.e. wildfires that burned more than 100 acres). This dissertation assesses the potential for a statistical model to predict the occurrence of all wildfire sizes. Therefore, the entire fire occurrence database compiled by Pollina et al. (2013) is used to quantify the effectiveness of different statistical model formulations in predicting the occurrence of small wildfires (less than 20 acres) and larger wildfires (20 to 10,000 acres).

A comparison between hourly meteorological conditions on wildfire occurrence days and the climatological average is done with the following variables: 2-m temperature (TEMP), 2-m relative humidity (RELH), 2-m dew point (TMPD), 2-m specific humidity (SPHU), 2-m mixing ratio (MIXR), 10-m wind speed (WNDS), and cumulative daily precipitation (PCP). Only a single daily value is used: the daily maximum value for TEMP and WNDS, and the daily minimum value for RELH, TMPD, SPHU, and MIXR. Daily mean values are also tested for TEMP, WNDS, RELH, and TMPD with similar or less robust results (not shown). In addition, zero to five day lagged values of the meteorological variables is employed to assess whether days preceding the wildfire occurrence day are associated with wildfire occurrence. A single value for each D1 or D2 domain is computed by taking the spatial median of the daily meteorological variables. The spatial mean is also tested for both D1 and D2 with similar results (not shown). In addition, the day is excluded if any station within the domain reports snow cover.

Before comparing to climatology, the mean of the variables is removed with respect to their 1) annual averages, 2) annual averages with standardization (defined as removing the mean and dividing by the standard deviation), 3) daily climatological averages and 4) daily climatological averages with standardization. Note that only the anomaly is computed with PCP with no standardization. Method 2) exhibits the largest differences between climatology and fire occurrence days for all variables (not shown). Significance of the meteorological variable differences is assessed via bootstrapping (Wilks 2011); where both the climatology and wildfire occurrence datasets are resampled with replacement 10,000 times. For both datasets, the size for each resample is equivalent to the total number of days in the original wildfire occurrence dataset (643 days in D1 and 964 days in D2).

2 Logistic regression model

Creating indices and warnings based on thresholds of atmospheric or other relevant variables is a common practice for fire weather [e.g. the Haines Index (Haines 1988), the ventilation index (Hardy et al. 2001), and National Weather Service Red Flag Warnings]. However, the process of defining the thresholds is often ad hoc and poorly documented, which can contribute to subjective interpretations of what the indices mean. Using a multi-linear regression with wildfire occurrence data regressed on atmospheric predictors is one methodology for documenting the index development process. However a multi-linear regression is not necessarily appropriate for wildfire occurrence in the NEUS because the binomial distribution of the daily NEUS wildfire occurrence data violates the normal distribution assumption. Furthermore, it is not obvious what useful non-probability based output (i.e. fire size) could be generated for wildfire managers or operational forecasters by using a simple multi-linear regression, since the relationship is likely to be weak. A variant of the linear regression known as a binomial logistic regression model is employed to overcome these challenges.

The binomial logistic regression model produces a probabilistic output of a binary predictand conditional on one or more predictors (Wilks 2011, his Eq. 7.29a). For this dissertation, the predictand is the binary occurrence of wildfire on a given day in a region trained with near-surface weather predictors. The functional form for the logistic regression model is:

$$\ln\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_m x_{mi} \qquad (1)$$

where p is the probability of a wildfire occurring in the domain, i is each data sample (i.e. each day), b are the regression coefficients, x are the weather predictors, and m are the number of predictors in the regression. This is similar in formulation to the logistic regression implementation in Mondal and Sukumar (2014) to predict probability of fire occurrence over southern India. In order to isolate useful predictors, potentially significant atmospheric variables are tested in different combinations. After the predictors are demeaned and spatially averaged, uncertainty in the parameter estimates are determined by randomly splitting the data 10,000 times into two equally sized separate 5-year calibration and independent verification periods. This is analogous to the resampling bootstrap method described in the previous subsection. Confidence is assessed by looking at the 2.5th and 97.5th percentile of the 10,000 resampled datasets. Once the optimal binomial logistic regression configuration is determined, the same regression is repeated using ASOS data between 1979 and 2013 to develop a climatology of FWI.

2.3 Results

a. Observed wildfire weather climatology

Figure 2.3 shows the monthly frequency of observed wildfires from the 1999 to 2008 climatological dataset for small wildfires (less than 20 acres) and large wildfires (20 to 10,000 acres) within D1 and D2. Wildfire occurrence peaks in both D1 and D2 in April for all sizes, with a smaller secondary peak in November for all size D1 wildfires and smaller D2 wildfires. The April/November peaks in wildfire occurrence are consistent with the results from Pollina et al. (2013; their Fig. 4), although their study uses a larger domain and only considers wildfire sizes greater than 100 acres. The wildfire occurrence data also suggests a stronger seasonal influence on smaller wildfire occurrence in D1 compared to D2, as indicated by the greater abundance of smaller wildfires occurring for any given month within D2.

b. Identifying relevant atmospheric variables

Anomalies with confidence intervals from the 1979 to 2013 surface TEMP, TMPD, RELH, and WNDS are shown for the climatological average (blue) and for wildfire occurrence days (red) for small and larger wildfires in the two domains (Fig. 2.4). RELH is significantly different than climatology (defined as the 2.5th and 97.5th percentile of the resampled confidence intervals not overlapping) on wildfire occurrence days for D1 small wildfires [-0.53 standard deviation (stdev); >99.9% confidence], D1 large wildfires (-1.00 stdev; >99.9% confidence), D2 small wildfires (-0.28 stdev; >99.9% confidence) and D2 large wildfires (-0.88 stdev; >99.9% confidence). TMPD differences are also statistically significant for D1 small wildfires (-0.18 K stdev; >95% confidence) and large wildfires (-0.55 K stdev; >99% confidence). Although not statistically significant, TEMP is slightly warmer for D1 small wildfires (0.10 K stdev), D2 small wildfires (0.11 K stdev) and D2 large wildfires (0.18 K stdev). There is also a statistically significant WNDS difference on days with larger wildfires for D1 (0.44 m s-^{1} stdev; > 95% confidence) but not D2 (0.35 m s-^{1} stdev). SPHU and MIXR are not significant for any domain or fire size (not shown) and are not analyzed further in this dissertation.

These results suggest that surface RELH and TMPD are the most important meteorological predictors for wildfire occurrence days over the NEUS, while surface TEMP might help distinguish wildfire occurrence days in some cases, and WNDS might distinguish wildfire occurrence days when larger wildfires are present. The clear separation in RELH between fire occurrence days and climatology is consistent with the results of Simard et al. (1987), who noted that RELH was the best discriminator in statistical tests for extreme fire environments in the NEUS.

Figure 2.5 shows the average zero to five day lagged PCP anomalies with respect to the climatological average on wildfire occurrence days for D1 and D2. Day zero lagged PCP is significantly different than climatology at the 95% confidence level on wildfire occurrence days for D1 small wildfires (-2.0 mm; >99.9% confidence), D1 large wildfires (-3.0 mm; >99% confidence), and D2 small wildfires (-1.0 mm; >99% confidence), but not for D2 large wildfires (-1.8 mm). There are no statistically significant differences in lagged PCP between one and five days except for day one lagged average PCP for small wildfires in D1 (-1.4 mm; >95% confidence). From this analysis, it is not clear on how to incorporate PCP in a statistical FWI when only the day-zero PCP value indicates a significant relationship. Future work could utilize the potential impact of local short-term drought using PCP deficit information.

c. Evaluating logistic regression model formulations for FWD2

For this dissertation, predictors are used that are determined from section 2b to be potentially significant and therefore likely to improve the statistical model skill. Since the differences between the climatological average and fire occurrence days are most robust with RELH, lagged combinations of RELH are also tested as potential predictors in the logistic regression model. These additional predictors include one-day lagged RELH (RELH1), 2-day lagged RELH (RELH2) and the 20-day averaged anomaly of lagged RELH (RELH20). The results for WNDS and TMPD are never significant in terms of the p-value from the logistic regression model parameter estimate (not shown) and are not mentioned further in this dissertation. The following logistic regression model predictor configurations are presented:

1) Logistic Regression Model 1 (LRM1) consists of 2 predictors – TEMP and RELH0
2) Logistic Regression Model 2 (LRM2) consists of 3 predictors – TEMP, RELH0 and RELH1
3) Logistic Regression Model 3 (LRM3) consists of 4 predictors – TEMP, RELH0, RELH1 and RELH2
4) Logistic Regression Model 4 (LRM4) consists of 3 predictors – TEMP, RELH0 and RELH20

To evaluate the contribution of each predictor to the logistic regression model, the significance values (p-values) of b_0 and the predictor coefficients are shown for each logistic regression model formulation in the calibration period (Fig. 2.6). The error bars represent the 2.5 and 97.5 percentile of the resampled parameter estimates in the calibration period's LRM1 (red), LRM2 (green), LRM3 (blue) and LRM4 (cyan) model for D1 (X's) and D2 (circles). Parameter estimates below the p-value of 0.05 (solid black line) are statistically significant at greater than 95% confidence. For all logistic regression model formulations, TEMP and RELH0 are significant, suggesting they would likely add skill in the verification period. RELH1 is also significant for LRM2 but not LRM3, which is an indication that including too many parameters results in overfitting for both domains. Otherwise, all logistic regression model formulations have statistically significant p-values for D1. D2 parameter estimates are in all cases less significant than D1 (although most are still statistically significant), suggesting that wildfire occurrence in D2 might be less dependent upon overlying atmospheric conditions than wildfire occurrence in D1.

In order to test whether these results are affected by the inclusion of small wildfires in the wildfire occurrence database, LRM1 through LRM4 are repeated using different minimum wildfire size thresholds (Fig. 2.7). The p-values associated with TEMP, RELH0 and RELH1 generally increase (i.e. become less significant) as more of the smallest fires in the database are excluded, although RELH0 remains significant for all domains and all wildfire sizes. An exception to the trend of decreasing significance occurs with the RELH0 and RELH20 predictors in D2, where parameter estimates become more statistically significant after eliminating fire sizes of less than one acre. This supports the above assertion that the occurrence of very small wildfires in D2 may not strongly depend upon overlying atmospheric conditions. However, most parameter estimates are statistically significant for all fire sizes, suggesting that meteorological conditions are a factor in the occurrence of most small and large fires in the NEUS.

Before determining the best logistic regression model from which to develop an FWI, it is critical to analyze the performance of LRM1 through LRM4 in a randomly sampled independent verification period. This is accomplished by comparing the Brier Skill Scores (BSS; Wilks 2011, his Eq. 8.37) of each logistic regression model to climatology. In this case, climatology is defined as the twice-smoothed 30-day running mean probability of wildfire occurrence derived from the raw 10-year climatology. This methodology gives the climatology an advantage over the model, since the climatology is computed from the full 10-year dataset while the logistic regression model is only calibrated with 5 years of randomly resampled data and verified with the remaining 5 years as described in the resampling description of section 2b. The BSS of all the logistic regression models referenced against climatology are significantly greater than zero (Fig. 2.8), suggesting that each model formulation provides better probabilistic skill than climatology. Interestingly, even if the logistic regression model is calibrated using the wildfire climatology of a different domain, the model is still more skillful than climatology. The more complex logistic regression model formulations of LRM2 - LRM4 provide more probabilistic skill than LRM1 in D1, although the differences are not statistically significant. Interestingly, D2 does not benefit from a more complex model. The differences between models are not statistically significant; therefore, LRM1 is selected for developing an FWI, since it represents the simplest formulation.

d. Verification and definition of the FWI

Before defining the FWI from the LRM1 model results, it is important to explore the LRM1 phase space and verify its ability to predict wildfire occurrence. Figure 2.9 shows the relationship between LRM1 fire occurrence probabilities and the size/occurrence of observed wildfires for D1 and D2. Observed wildfires for D1 and D2 tend to fall on days with lower RELH and to a lesser extent, higher TEMP. Since there is no visible relationship between wildfire size and the fire occurrence probabilities, the FWI should be expected to predict wildfire occurrence, not wildfire size.

It is important to examine the seasonality of TEMP and RELH and how that might relate to fire occurrence. Therefore, the climatological mean and standard deviation for TEMP and RELH are plotted by day of year between 1999 and 2008 over D1 (Fig. 2.10). In addition, the evolution of the least active (2000; 23 observed fires) and most active (2006; 104 fires observed) fire occurrence years are plotted for comparison. The climatological RELH experiences a winter and summer maximum along with a spring time minimum coincident with the peak of the fire season. The 2006 season experiences consistently drier RELH than the 2000 season, but the lower RELH values may be most critical before or during the green up period between February and May when most fires are observed. There is little difference in TEMP between the 2000 and

2006 seasons during the late winter and spring, emphasizing that RELH is likely the more important predictor to fire occurrence.

An important characteristic of any probabilistic model is reliability, which describes the relationship between specific values of the forecast and the average observation (Wilks 2011). A common methodology to display reliability is by comparing model forecast probability against observed relative frequency (i.e., what is actually observed) by threshold. Figure 2.10 shows the reliability of LRM1, where the forecasted probability of fire occurrence is binned into increments of 10% and compared to the observed relative frequency. Therefore, if the forecasted fire probability matches the observed relative frequency, the probabilities would fall perfectly on the 1:1 line and the forecast would be considered reliable. The ability of LRM1 to predict wildfire occurrence does not strongly depend upon the domain or the choice of parameter (Fig. 2.11), which is consistent with the BSS analysis in Fig. 2.8. For higher thresholds of probability, LRM1 reliability is above the 1:1 line, indicating under-prediction of wildfire occurrence, although the sample sizes associated with probabilities greater than 50% are limited to 50 cases or less. These results support the conclusion that LRM1 produces a known quantitative output (i.e. probability of wildfire occurrence within a specified domain) that is on the average reliable (Fig. 2.11) and has greater skill than climatology (Fig. 2.8).

Finally, in order to convert the probabilistic output of the LRM1 model into a FWI, four potential FWI categories are defined and categorized as zero, one, two, or three. An FWI of zero corresponds to LRM1 probability of fire occurrence in the domain below 30%, one between 30% and 40%, two between 40% and 50%, and three greater than 50%. For D1 (D2) an FWI of greater than zero, one, and two consists of 10%, 4.3% and 1.4% (26.6%, 8.4%, and 0.4%) of the total days, respectively. These thresholds are chosen based on their rarity and intuitive statistical relationship (i.e. 30%, 40% and 50%) to fire occurrence within each domain. Although the difference between a 30% and 50% chance of fire occurrence may seem minor, the anthropogenic nature of fire occurrence indicates a 50% prediction to be very anomalous. Therefore, the FWI categorizes minimal, low, moderate and high potential for wildfire occurrence as the index increases from zero to three, respectively.

Both the logistic regression model and climatological probabilities are converted to index values as described above to evaluate the benefit of the model using the resampled verification period. The results are presented in the form of the critical success index (CSI; Fig. 2.12a), false alarm ratio (FAR; Fig. 2.12b) and hit rate (HIT; Fig. 2.12c). The climatological FWI (dashed) never produces a category greater than one, and thus only the logistic regression model FWI produces moderate and high categories. When the FWI is one, the logistic regression model produces better CSI and HIT and comparable FAR compared to climatology. It is not surprising that the FAR is high since almost all wildfires in the NEUS have anthropogenic sources (Northeast States Emergency Consortium 2014). For instance, on the average a high FAR would arise even with a high FWI, since the anthropogenic component is not considered in the binomial logistic regression. Incorporating this additional anthropogenic factor into the FWI would be very difficult due to its likely stochastic nature.

Figure 2.13 shows the relationship between FWI value and observed wildfire size/probability. For D1 and D2, both observed wildfire size and wildfire probability increases steadily with increasing FWI, with statistically significant differences for D1 (41.65%; > 99% confidence) and D2 (59.4%; > 99% confidence) wildfire probability between an FWI of one and three. However, all changes in wildfire size with FWI are not statistically significant, emphasizing that the FWI should not be expected to predict wildfire size.

e. FWI climatology

Hitherto, this dissertation has split all LRM models into separate calibration and verification periods using a resampling technique. Since the FWI has been extensively verified, the entire 10-years of wildfire climatology (1999 to 2008) are used as calibration to obtain more precise parameter estimates. These parameter estimates can be used operationally in atmospheric models in a predictive sense or with historical observations to develop a climatology of wildfire potential. In this case, the latter option is chosen by using the 10-year regressed parameter estimates on ASOS observations between 1979 and 2013.

Figure 2.14 shows the 1979 to 2013 FWI climatology using the parameters from the 1999 to 2008 training period stacked by FWI value for D1 and D2. The results are qualitatively similar to Fig. 2.3 with D2 exhibiting weaker seasonal variations and a greater number of FWI days than D1. For instance, 70.2% (56.3%) of all moderate to high FWI days occur between March and April within D1 (D2). This is likely related to low RELH and warm TEMP events occurring before or during the spring green up period, as shown in Fig. 5.10. However, the FWI does not capture the November secondary peak in observed fire occurrence for D1 or D2. This suggests that RELH may not be as significant a predictor in November perhaps resultant from a lower sun angle and weaker sensitivity to diurnal PBL growth. It is possible that other predictors are more critical for November fire occurrence, such as drought or the state of the dry fuels.

FWI value has considerable variability by year (between 33 – 104 events for D1; between 81 – 169 events for D2) for both domains (Fig. 2.15). A linear fit to the D1 (D2) FWI time series shows an increasing trend averaging 0.6 (0.6) wildfires per year at greater than 95% confidence (greater than 90% confidence). This increase in FWI is caused by a general decrease in RELH since 1979 over both domains along with a slight increase in TEMP (not shown). Other studies have found a recent increase in observed wildfire occurrence in northeast Spain (Cardil et al. 2013), the northern Sierra Nevada (Collins 2014), and in southern California (Jin et al. 2014). Clarke et al. (2013) also found an increase in fire weather between 1973 and 2010 for several stations in Australia. However, one should use caution before drawing any conclusions relating increased fire weather potential and climate change without analyzing a longer historical record and considering the influence of low frequency natural variability on wildfire occurrence.

f. Comparing FWD1, FWD2, and the fire occurrence datasets

Despite being developed over slightly different domains, FWD1 and FWD2 are compared to each other and to the fire occurrence data when they overlap between 2007 and 2008. This comparison provides some qualitative insight into FWD1's ability to predict fire occurrence over the NEUS and potential similarities to FWD2 despite FWD1's reliance on indices developed over the western United States. Resampling is not performed since the dates for FWD1 are chosen a priori (i.e. it is not clear how these dates could be resampled and compared to FWD2 dates) and resampling would likely underestimate the true uncertainty regardless due to differences in domain sizes between FWD1 and FWD2. There are 88 observed fire days in the overlap period between 2007 and 2008, with 68 FWD1 days and 148 FWD2 days (of which 68 days have a FWI = 2 and 30 days have a FWI = 3).

Figure 2.16 compares FWD1 and FWD2 for different index values to the observed fire occurrence data for CSI, FAR and HIT. In general, FWD1 is most similar to FWI >= 2. However, an FWI of 2 or greater exhibits higher CSI (by 0.04) and HIT (by 0.06), with a lower

FAR (by 0.06) than FWD1. As mentioned earlier, these comparisons may not be fair due to the differences in domain sizes for FWD1 and FWD2. Various index values for FWD2 are compared to FWD1 in figure 2.17. The HIT is 0.79 when FWI >= 1, suggesting that most of the days captured in FWD1 are also captured in the lowest category of FWD2. However, because FWD2 has considerably more fire weather days than FWD1 (by 80 days), FAR is quite high at 0.67 when FWI > = 1. The highest CSI in Figure 2.16 occurs when the FWI >= 2, further supporting that FWD1 is most comparable to FWD2 at the FWI >= 2 threshold. Overall, these results suggest that there is some similarity between FWD1 and FWD2, with FWD2 capturing more days than FWD1.

2.4 Summary and conclusions

Two different methodologies for subsetting fire weather days (FWDs) over the Northeast United States (NEUS) are discussed; one using preexisting indices that considers meteorological influences (FWD1) and another based on a new statistical fire weather index (FWI) formulated from observations over the NEUS (FWD2). FWD1 uses the Fire Potential Index (FPI) when available and the National Fire Danger Rating System (NFDRS) when the FPI is unavailable to collect a list of fire weather days between 2007 and 2009. Considerable attention in this chapter is devoted to optimizing the predictive ability of the statistical FWI before being used to subset days as FWD2.

The FWI is developed using atmospheric Automated Surface Observing System (ASOS) observations and observed wildfire occurrence days (Pollina et al. 2013) from 1999 to 2008 for two separate domains (D1 and D2). Important weather variables for wildfires are identified by analyzing differences in weather conditions on wildfire occurrence days and the climatological average using a number of 2-m thermodynamic [e.g. temperature (TEMP), specific humidity (SPHU), relative humidity (RELH), dew point (TMPD), and mixing ratio (MIXR)], kinematic [10-m wind speed (WNDS)], and daily accumulated precipitation (PCP). RELH (and by association TMPD) is found to be significantly lower on wildfire occurrence days. Wildfire occurrence days are generally slightly warmer and for larger wildfires (20 to 10,000 acres) slightly windier than climatology, although only the D1 result is statistically significant for WNDS. This suggests that TEMP and WNDS might have secondary importance in the development of wildfires over the NEUS. In addition, unlagged PCP is significantly lower on wildfire days, but this significance decays within two days of the fire event occurring.

Potentially significant weather variables for the FWI are tested in a binomial logistic regression model by splitting the data into separate calibration and verification periods 10,000 times via a resampling method. RELH and TEMP yield the most consistent improvement in model performance within the verification period, although including lagged RELH improved model performance for D1. This suggests that the model could potentially be improved by including a longer term slowly changing metric such as the Palmer Modified Drought Index (PMDI; Heddinghaus and Sabol 1991). For instance, Barbero et al. (2014) shows coincident long and short-term weather variability are critical to the development of very large fires in the eastern United States, suggesting that additional parameters could be added to enhance both probability of fire occurrence and fire size.

Independent reliability plots indicate that the logistic regression model produces accurate probabilistic forecasts of wildfire occurrence, even when the model is calibrated using a different domain. In addition, the logistic regression model produces sharper and more skillful forecasts than climatology. The logistic regression model output is converted to the FWI based on probability thresholds of fire occurrence. Probabilities below 30% are given an index of zero (i.e. minimal wildfire potential), probabilities between 30% and 40% are assigned a value of 1 (i.e. low wildfire potential), probabilities between 40% and 50% are assigned a value of 2 (i.e. moderate wildfire potential) and probabilities greater than 50% are given a value of 3 (i.e. high wildfire potential). Therefore, each category of the FWI has a useful probabilistic meaning that has been verified using independent wildfire data for two separate domains. A fire weather day for FWD2 is defined as having a FWI value of 1 or greater. A comparison between FWD1, FWD2, and fire occurrence days reveals broad similarities, although FWD2 experiences higher skill related to fire occurrence and is less strict than FWD1.

Surface RELH and TEMP can be used to statistically predict the probability of fire occurrence within a designated radius over the NEUS. As designed, applying the logistic regression model of FWD2 to additional data is simple and straightforward. Since the parameter estimates from the regression are generally spatially and temporally consistent within the region studied, only the standardized anomalies of temperature and relative humidity for that location are needed. The model's output contains the probability of wildfire occurrence within the domain selected. For this dissertation, FWD2 is used on ASOS observations between 1979 and 2013 to develop a climatology of wildfire potential. This wildfire potential climatology is very similar to the climatology of observed wildfires, with a noticeable peak in April for both domains, and a much smaller November peak. The FWI also captures the increased potential for wildfires in D2, and the weaker seasonality of wildfire occurrence in that region compared to D1. Although the binomial logistic regression model produces reliable probabilities of fire occurrence, additional predictors representing drought or the underlying dry fuels could improve the independent verification. Chapter 4 details the adaption of FWD2 to a grid for comparison with ensemble model data.

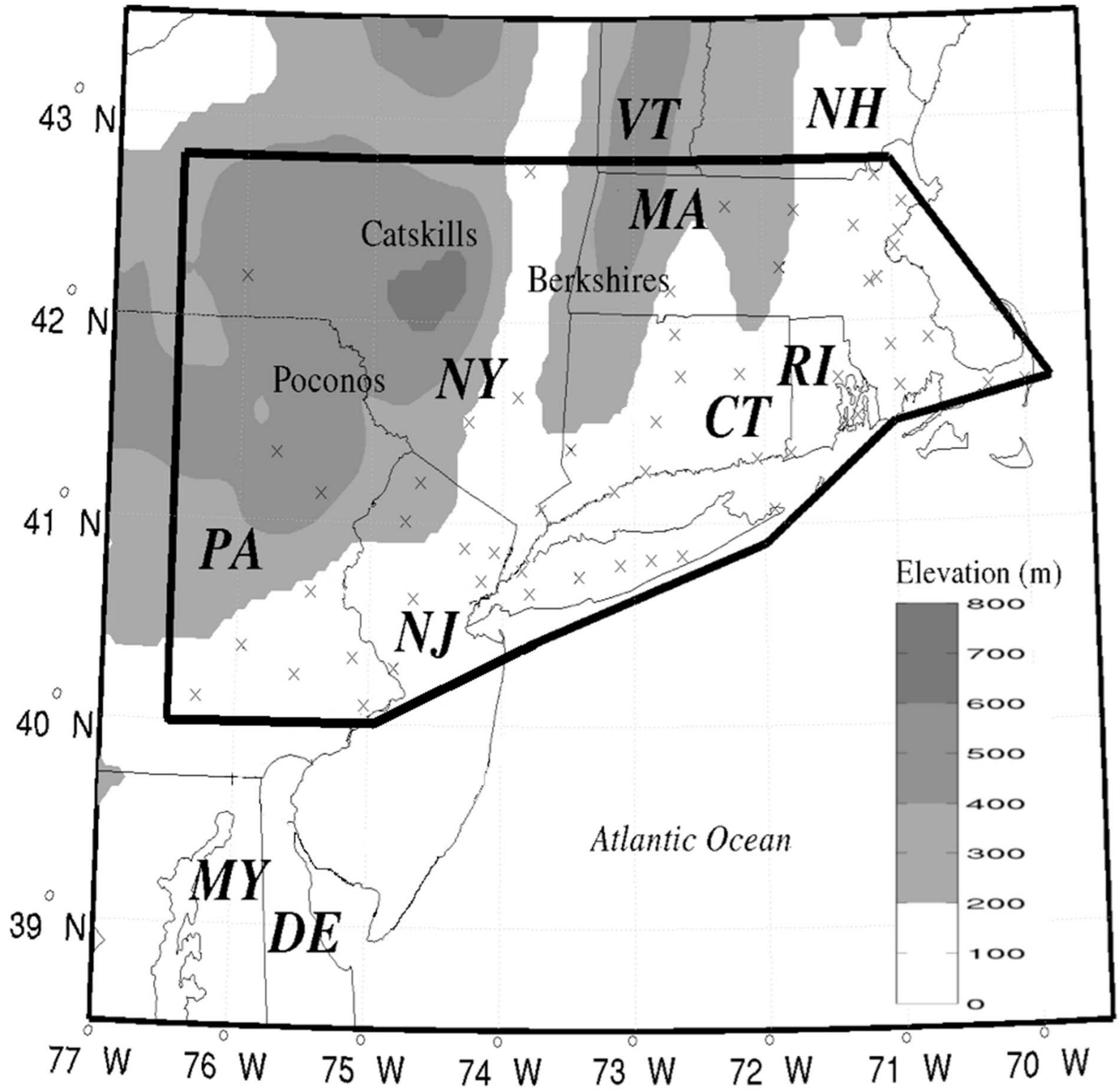

Figure 2.1: Domain used for the verification of FWD1 over the Northeast U.S. (black border) and the ASOS stations (x's). Shaded contours denote elevation while the text labels the mountainous regions of the Poconos, Catskills, and Berkshires. Additional text shows the states of Maryland (MY), Delaware (DE), New Jersey (NJ), Pennsylvania (PA), New York (NY), Connecticut (CT), Rhode Island (RI), Massachusetts (MA), Vermont (VT), and New Hampshire (NH).

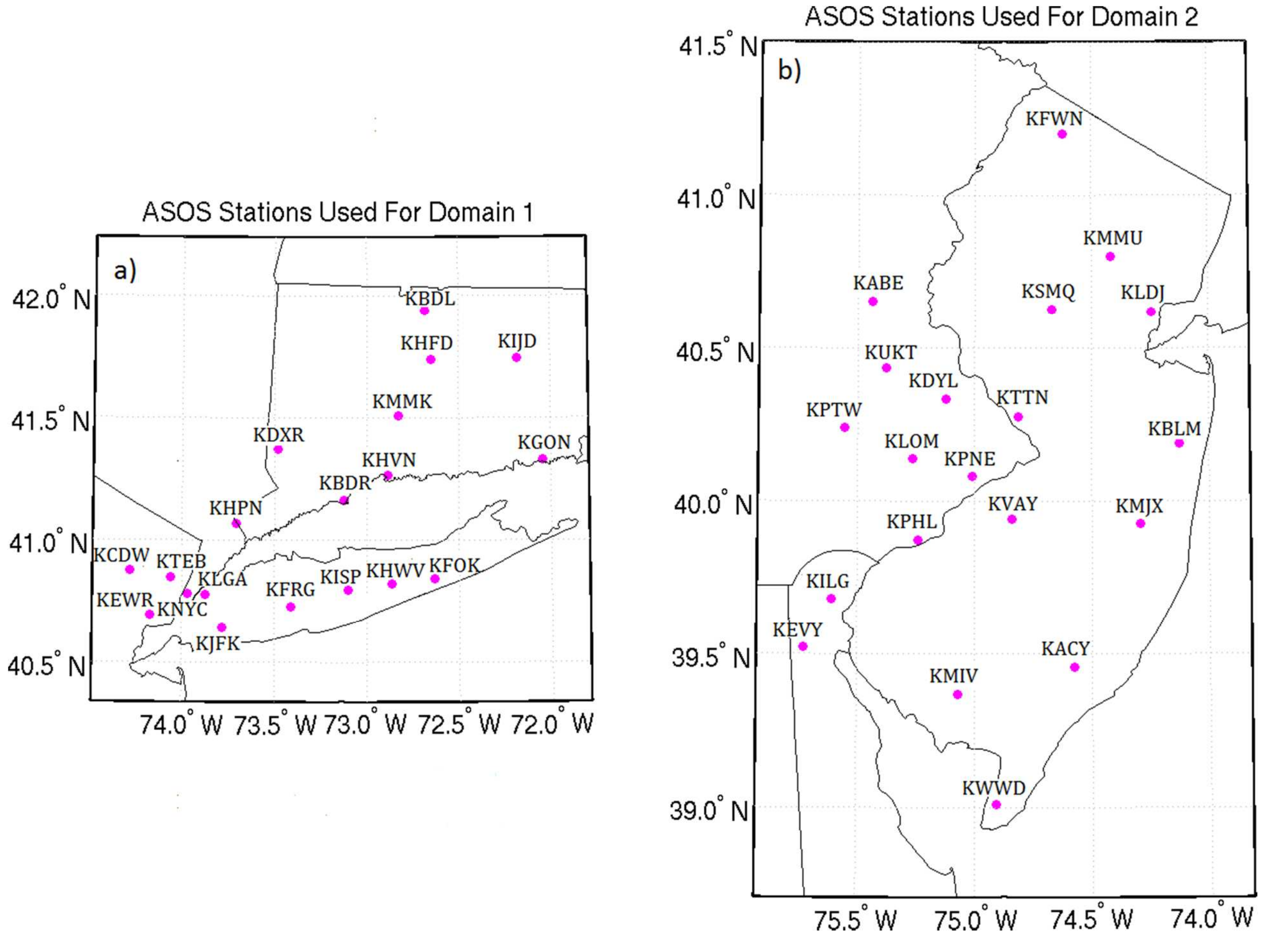

Figure 2.2: ASOS stations and locations used for (a) domain 1 and (b) domain 2.

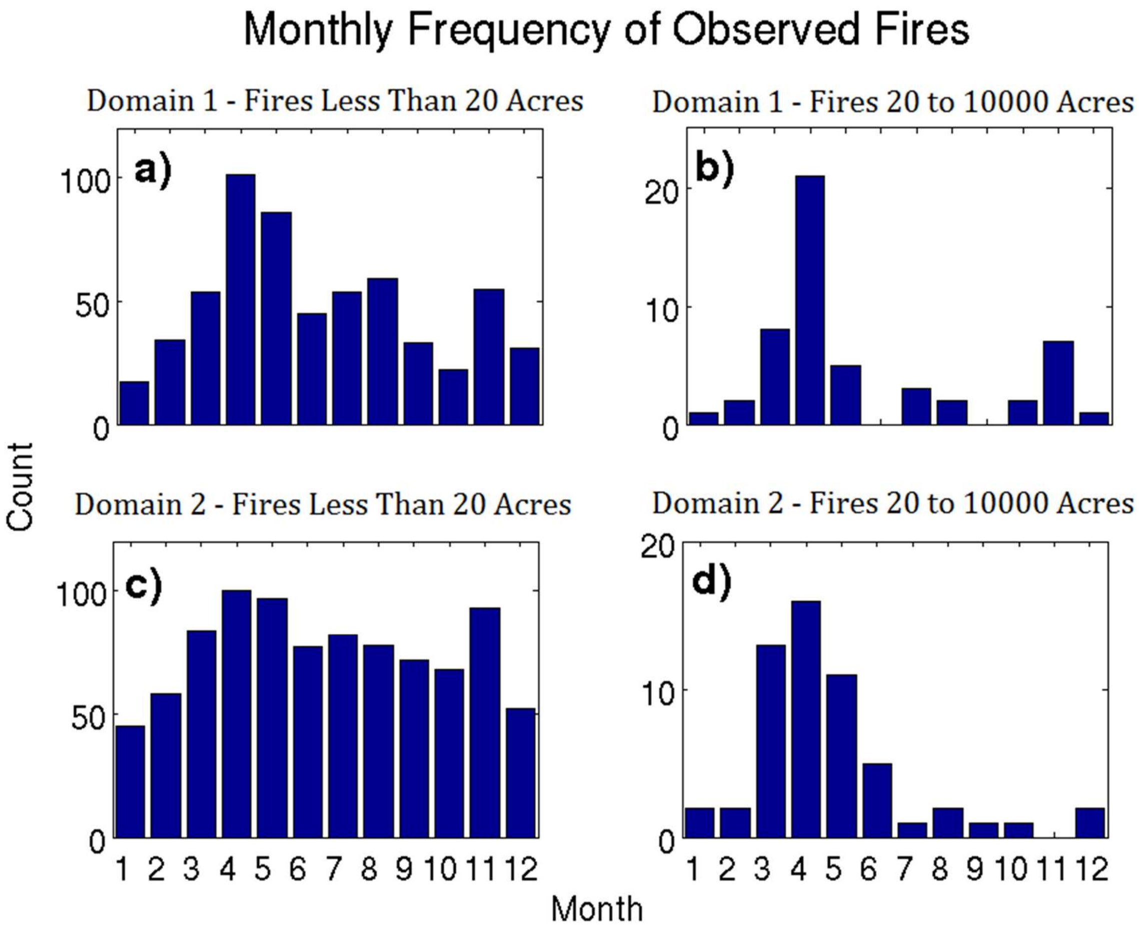

Figure 2.3: Monthly frequency of observed fire size less than 20 acres (a,c) and between 20 and 10,000 acres (b,d) for domain 1 (a,b) and domain 2 (c,d) from 1999 to 2008.

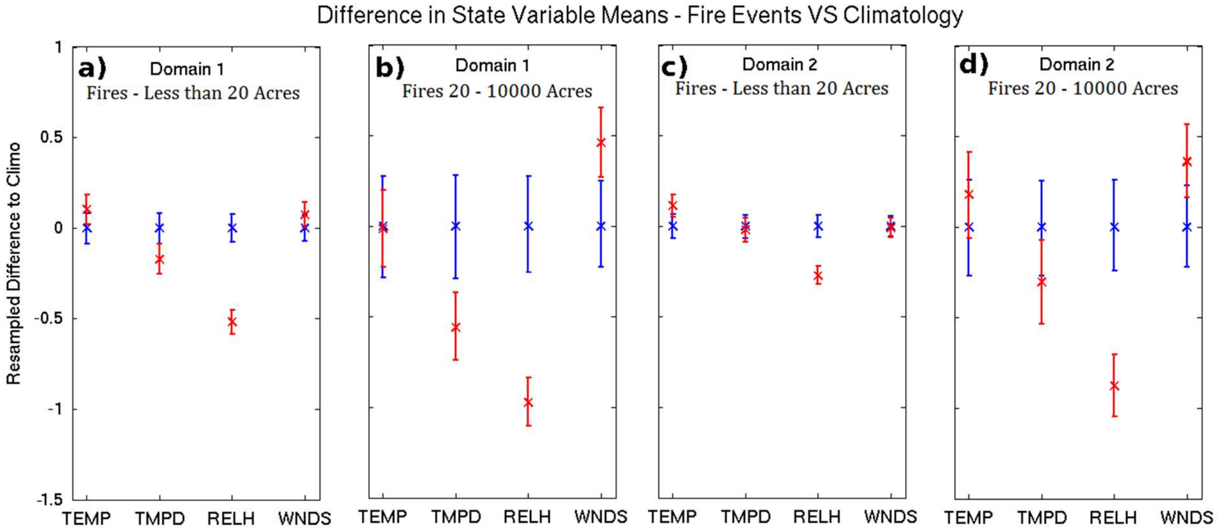

Figure 2.4: Temperature (TEMP), dew point (TMPD), relative humidity (RELH) and wind speed (WNDS) anomalies for the climatological average (blue) and observed fire days (red) for domain 1 fires less than 20 acres (a), domain 1 fires between 20 to 10,000 acres (b), domain 2 fires less than 20 acres (c) and domain 2 fires between 20 and 10,000 acres (d). Bars indicate the 2.5 and 97.5 percentile of the resampled dataset.

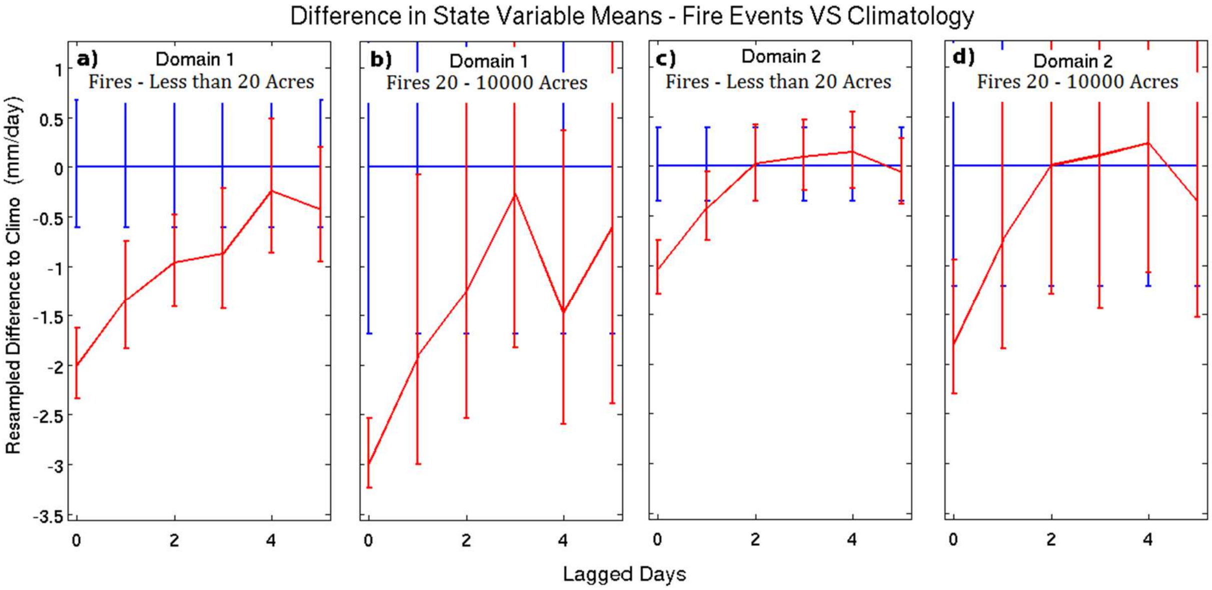

Figure 2.5: Same as figure 2.4, except for lagged precipitation anomalies between zero and five days.

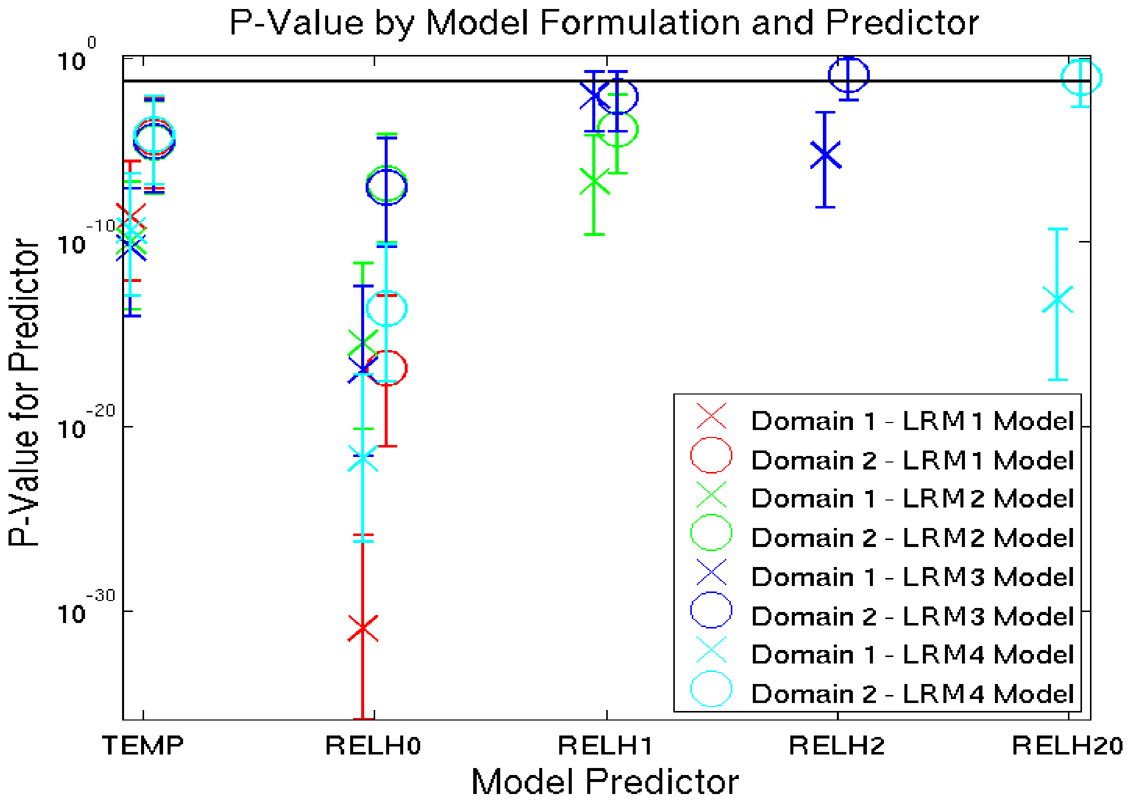

Figure 2.6: Domain 1 (X's) and domain 2 (circles) p-values (significance values) for the predictors (i.e. TEMP, RELH0, RELH1, RELH2 and RELH20) in each logistic regression model formulation (LRM1 – red; LRM2 – green; LRM3 – blue; LRM4 – cyan). Values below the black horizontal line are significant at greater than 95% confidence.

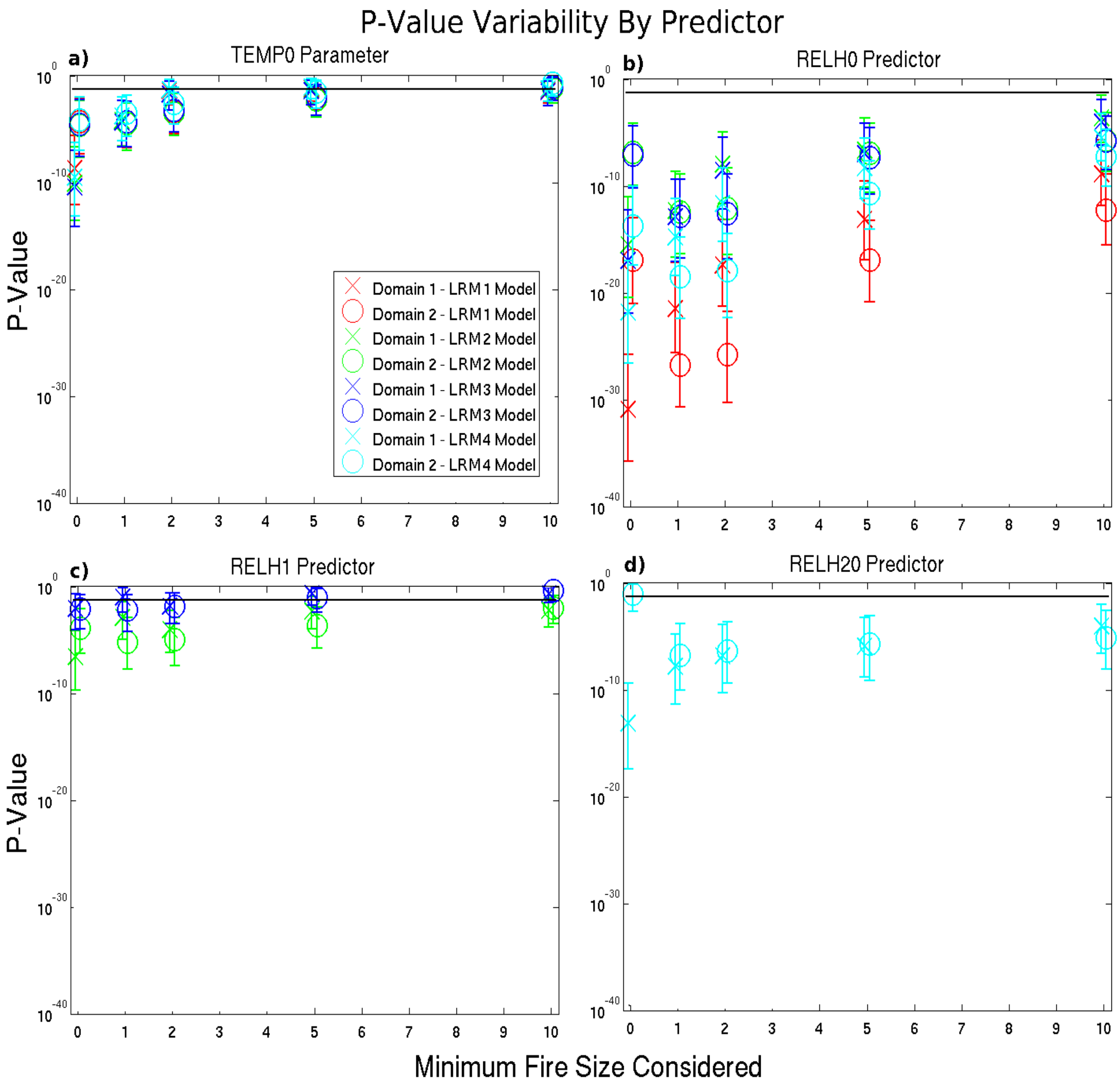

Figure 2.7: Calibration period p-values in LRM 1-4 models by predictor using different minimum fire size thresholds for TEMP (a), RELH0 (b), RELH1 (c) and RELH20 (d). Values below the black horizontal line are significant at greater than 95% confidence.

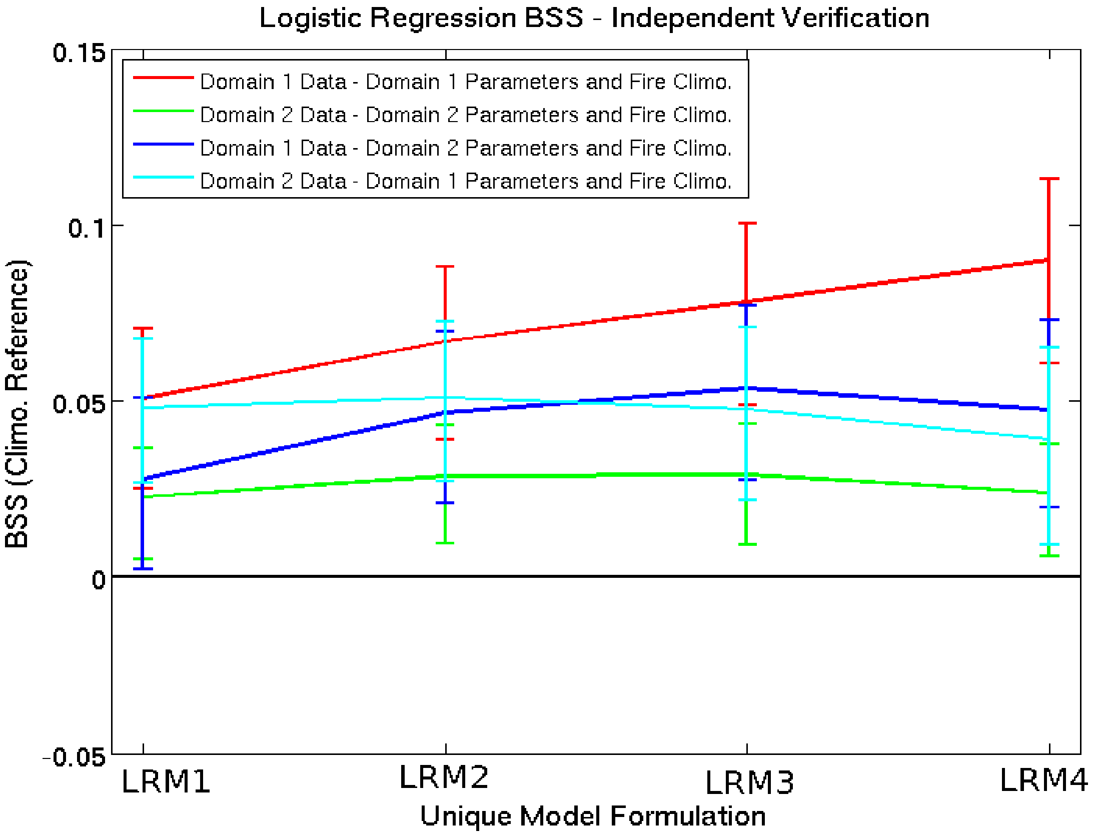

Figure 2.8: Independent logistic regression model Brier Skill Score (BSS) for LRM 1-4 compared to climatology for domain 1 data calibrated with domain 1 parameters (red), domain 2 data calibrated with domain 2 parameters (green), domain 1 data calibrated with domain 2 parameters (blue) and domain 2 data calibrated with domain 1 parameters (cyan). Bars denote 95% confidence intervals.

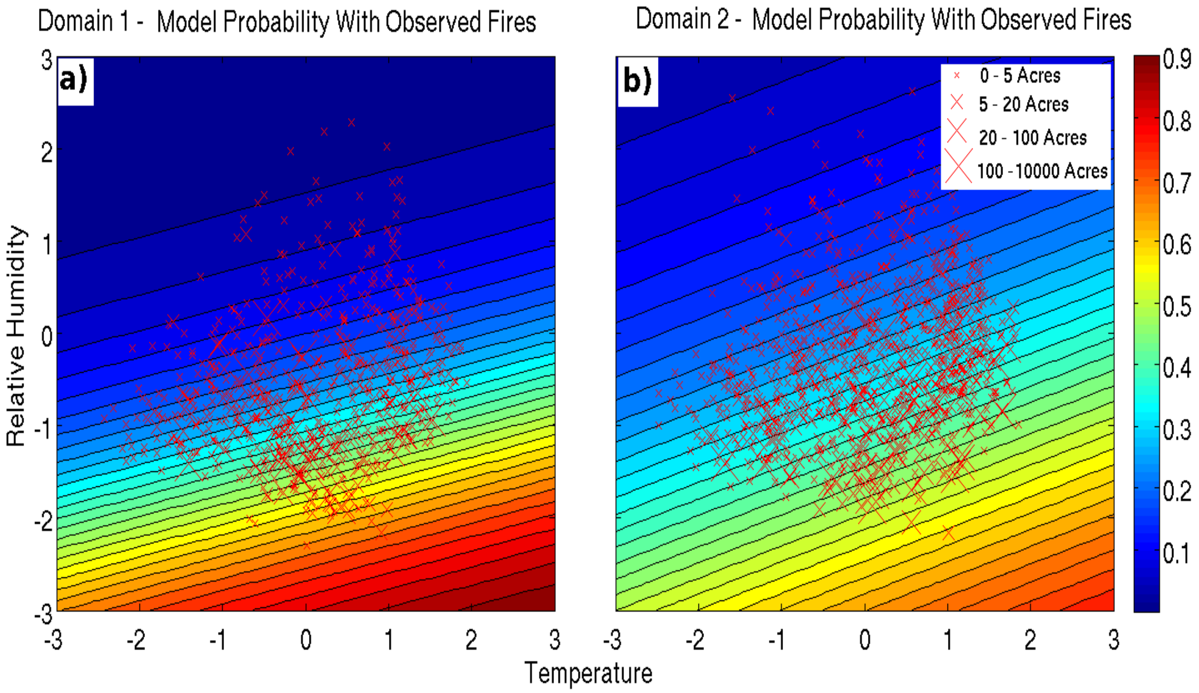

Figure 2.9: LRM1 model probability of fire occurrence (contoured) in model phase space and location of actual fire occurrence (X's) for domain 1 (a) and domain 2 (b).

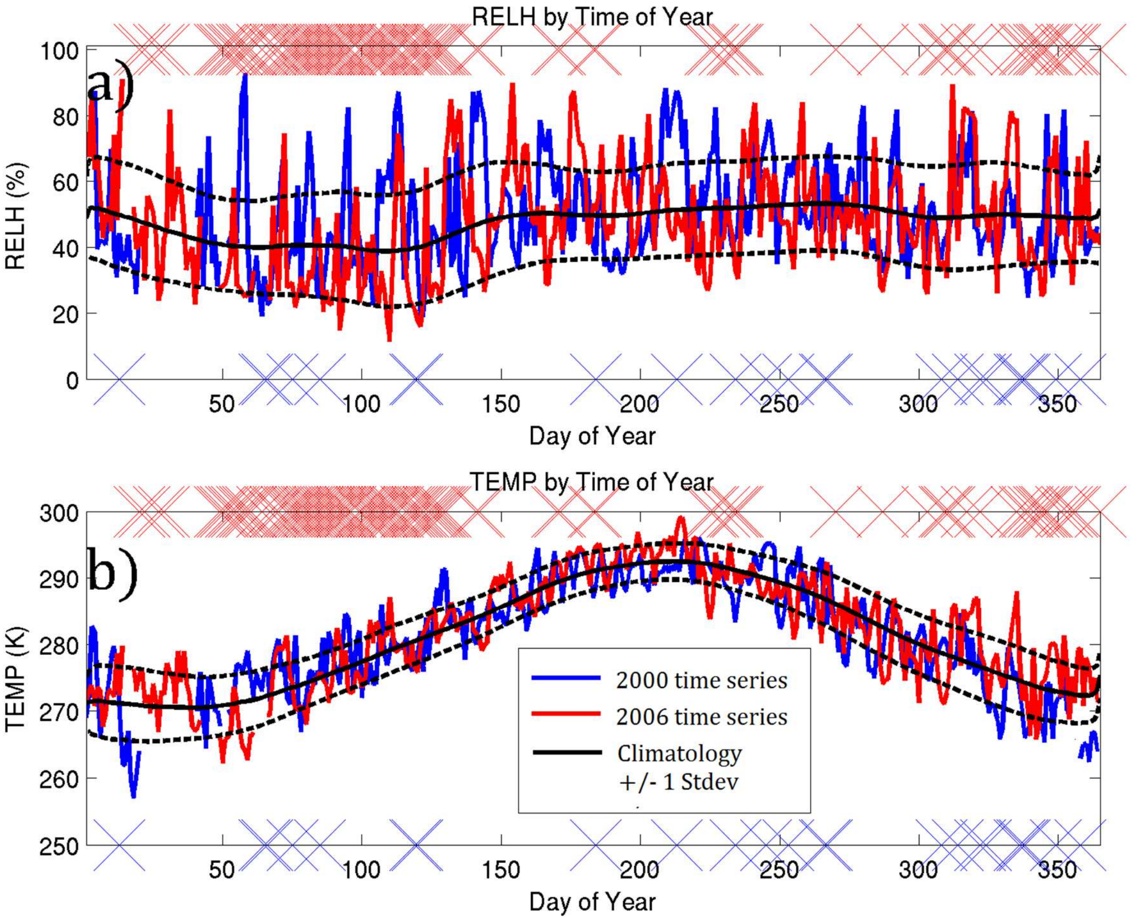

Figure 2.10: Time series for RELH (a) and TEMP (b) climatology (black), 2000 (blue), and 2006 (red). X's denote fire occurrence days for 2000 (blue) and 2006 (red).

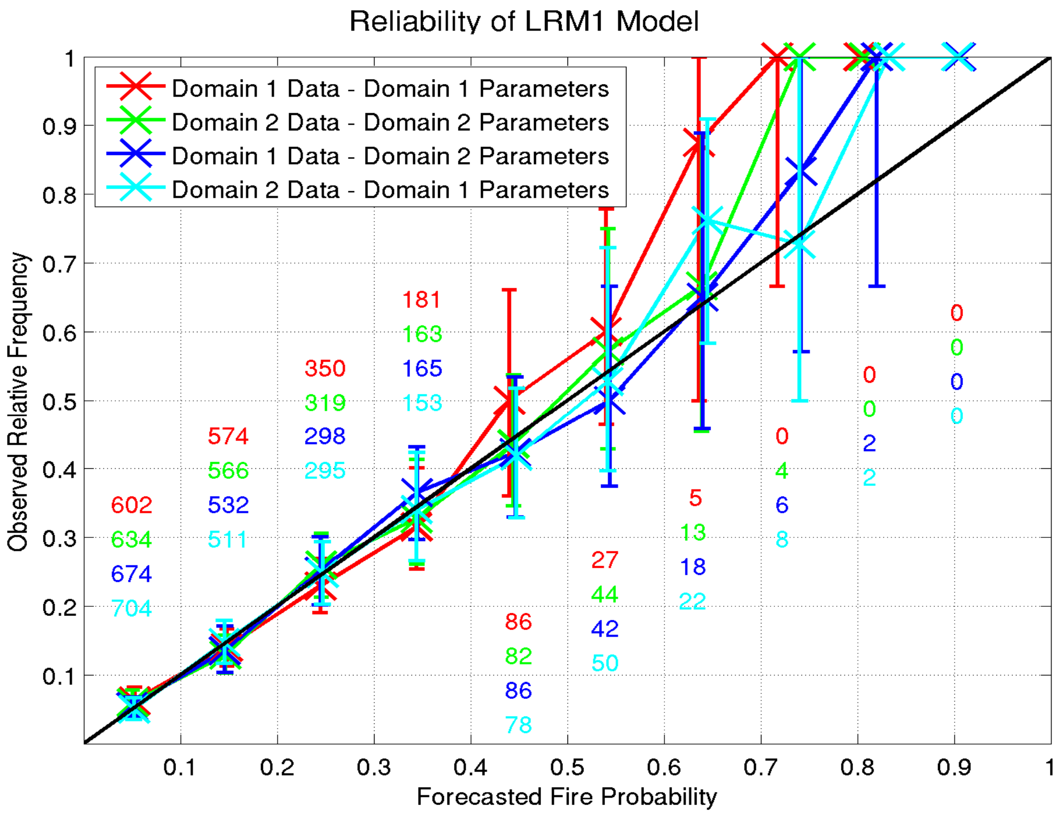

Figure 2.11: LRM1 independent reliability for domain 1 trained with domain 1 parameters (red), domain 2 trained with domain 2 parameters (green), domain 1 trained with domain 2 parameters (blue) and domain 2 trained with domain 1 parameters (cyan). Numbers are sample size for each probability bin. Error bars denote 95% confidence intervals.

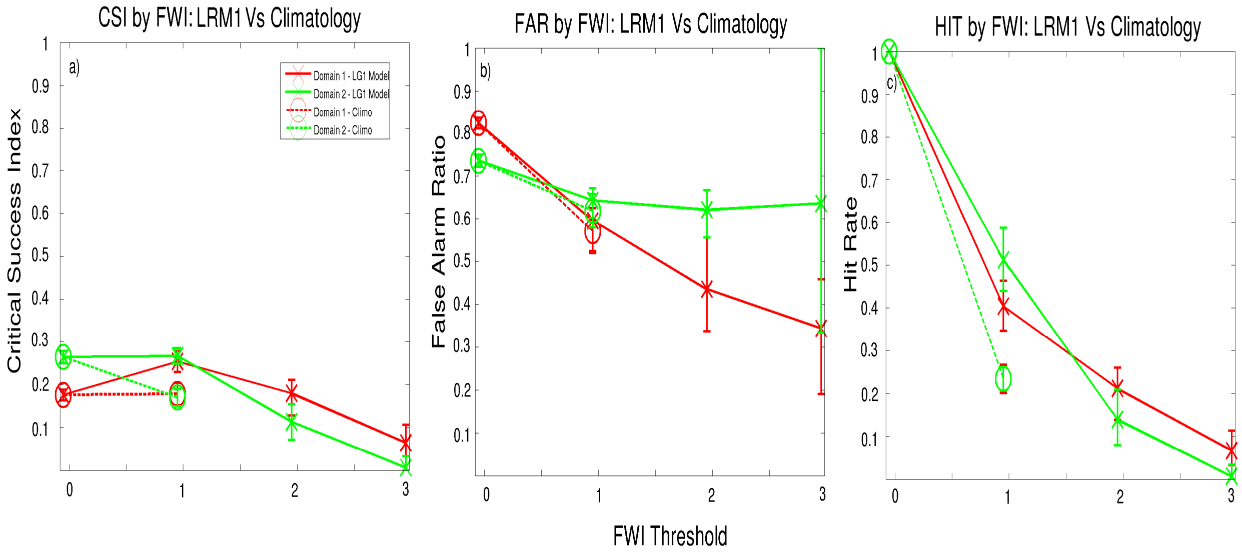

Figure 2.12: Independently verified CSI (a), FAR (b), and CSI (c) by HIT index value for domain 1 (red) and domain 2 (green) using the logistic regression model (solid) and climatology (dashed). Error bars denote 95% confidence intervals.

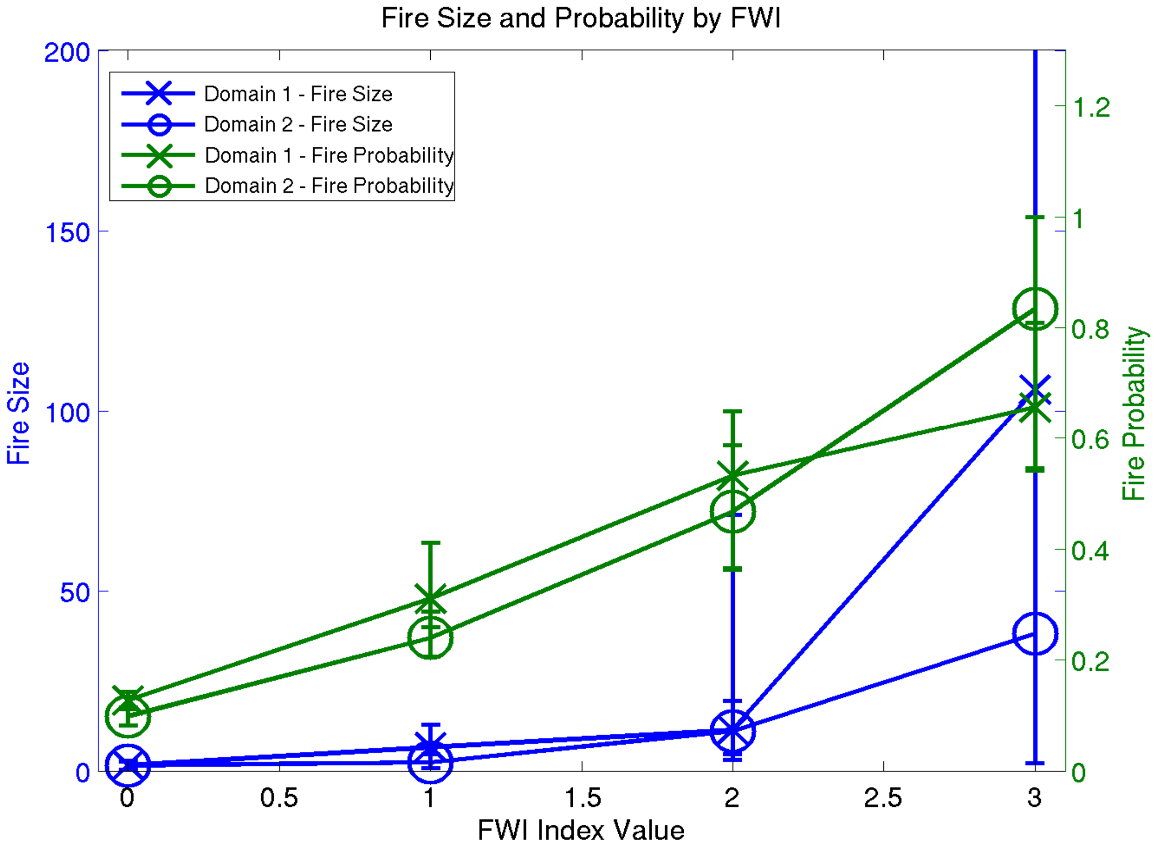

Figure 2.13: Average fire size (blue) and probability (green) by FWI value for domain 1 (X's) and domain 2 (circles). Error bars denote 95% confidence intervals.

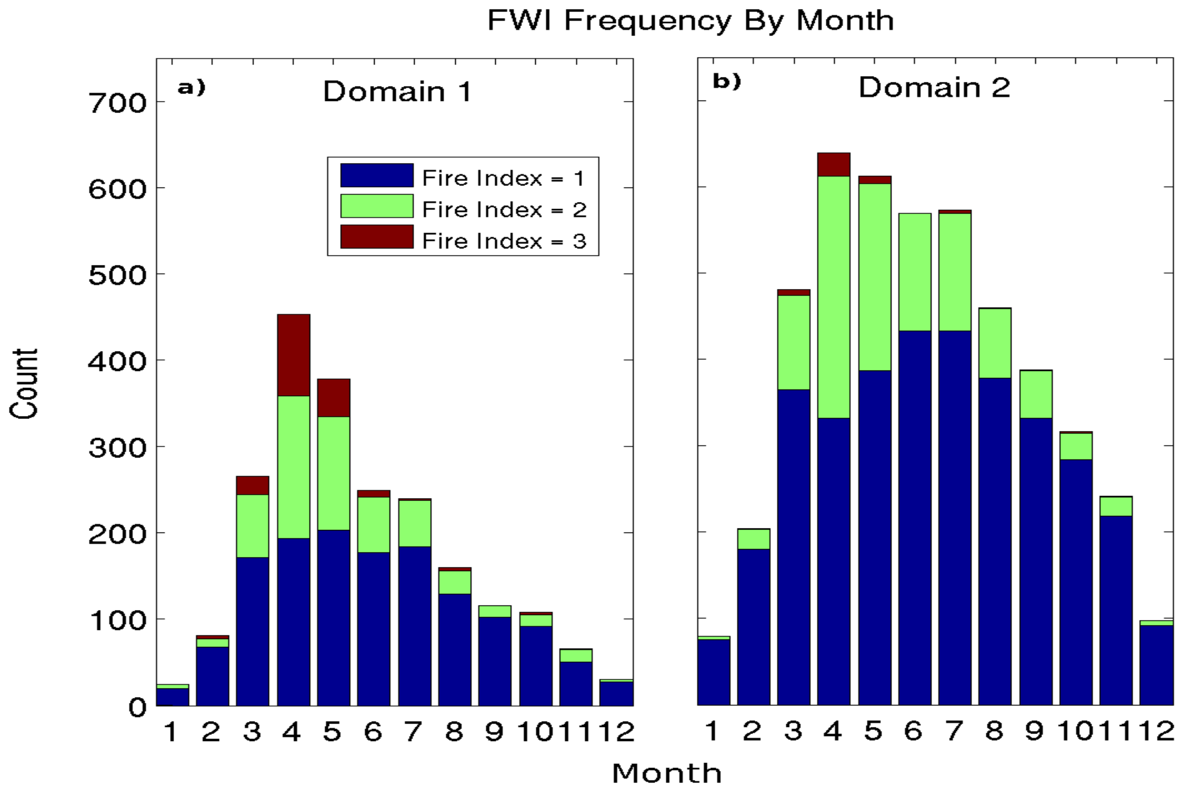

Figure 2.14: Monthly FWI by index value derived from LRM1 between 1979 and 2013 for domain 1 (a) and domain 2 (b).

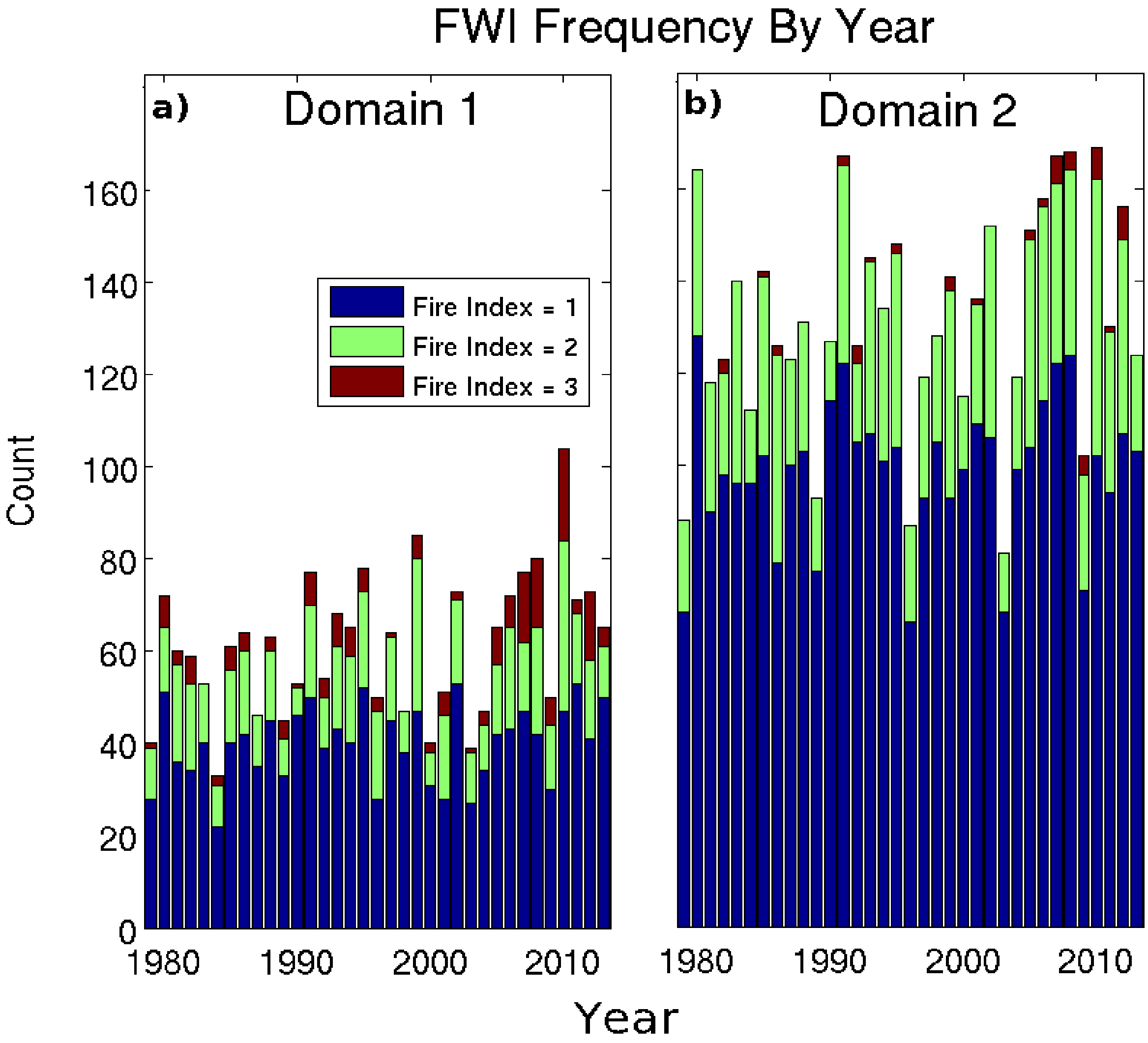

Figure 2.15: Same as figure 2.14, but for annual FWI frequency.

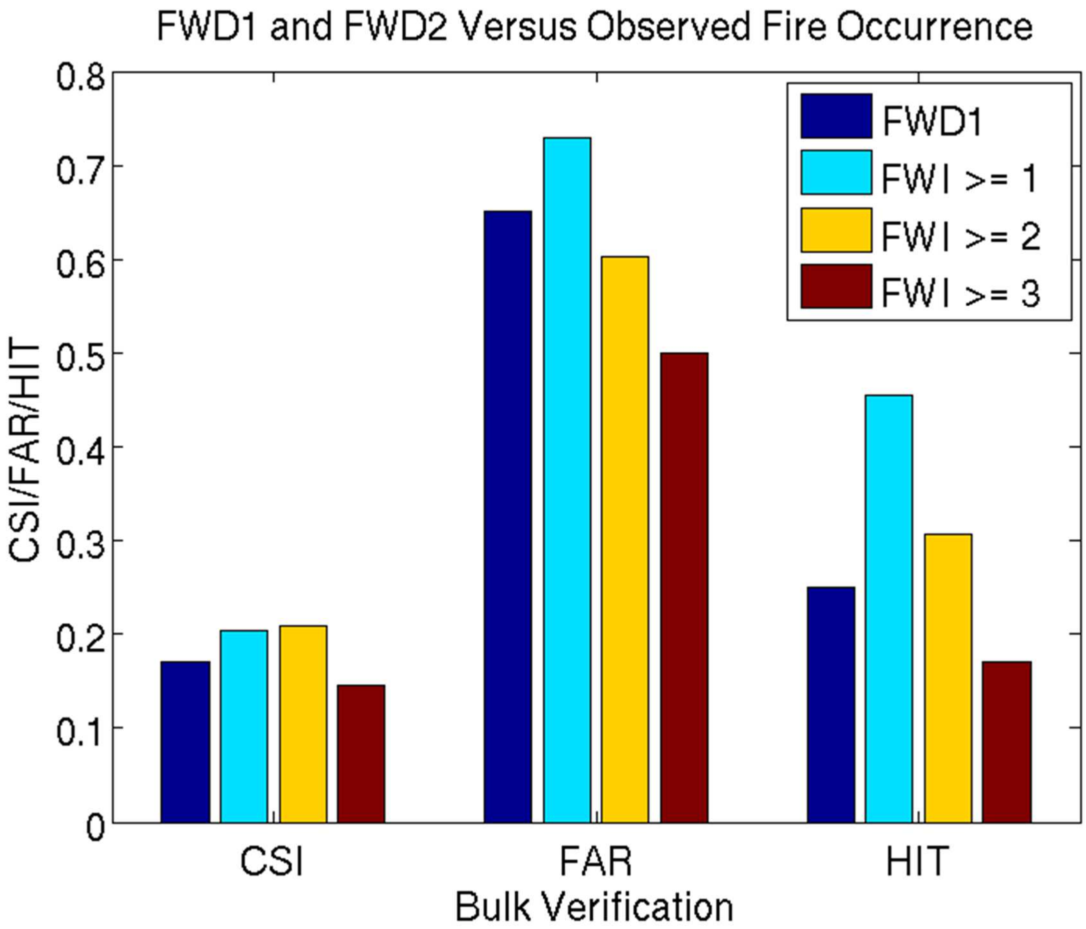

Figure 2.16: FWD1 (derived from existing fire threat indices) and FWD2 (derived from LRM1) verified against fire occurrence observations for CSI, FAR, and HIT between 2007 and 2008.

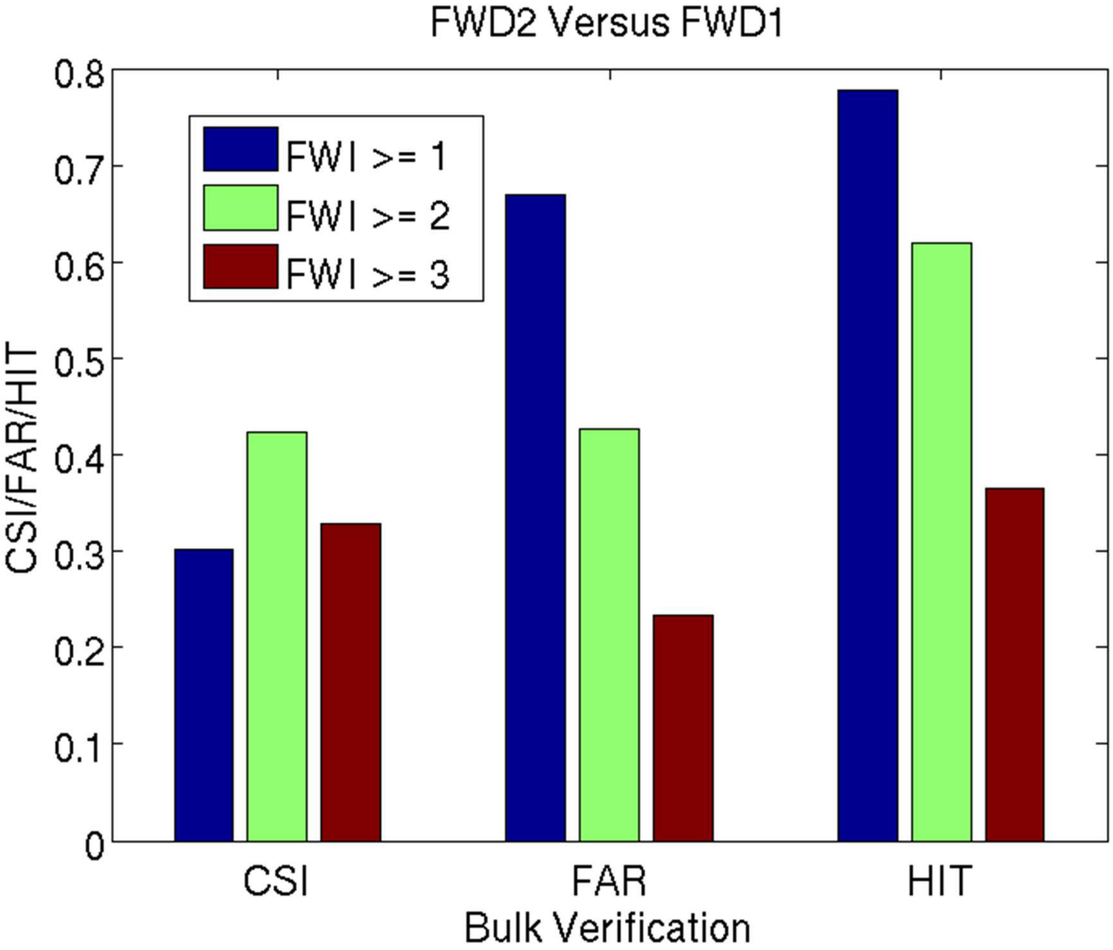

Figure 2.17: The index thresholds of FWD2 verified against FWD1 for CSI, FAR, and HIT between 2007 and 2008.

Chapter 3:

Impact of Bias Correction Type and Conditional Training on Bayesian Model Averaging Using FWD1 over the Northeast United States

3.1 Introduction

There have been a limited number of atmospheric ensemble model verification and post-processing studies over the Northeast United States (NEUS; Jones et al. 2007) and no NEUS verification studies on fire weather days (FWDs). Furthermore, the performance of Bayesian Model Averaging (BMA) has never been tested over the NEUS. This is concerning given the NEUS's high population density, complex topography, varying land use characteristics, and air mass source regions (Green and Kalkstein 1996). Additionally, the NEUS is affected by mesoscale features; including the sea breeze circulation (Miller and Keim 2003), convection (Lombardo and Colle 2010; Murray and Colle 2011), and terrain-forced flows (Gopalakrishnan et al. 2000; Carrera et al. 2009). As a result, attention is given to conditional model biases and post-processing, particularly on FWD1 events.

This chapter focuses on site-specific verification and post-processing methods at the surface. The multi-model ensemble used in this chapter considers several model cores, different model physics, and different initial conditions. This could be problematic since BMA has mostly been tested on ensembles ranging between 8 members (Raftery et al. 2005) and 18 members (Wilson et al. 2007). Fraley et al. (2010) ran BMA with 86 members; although 80 members are exchangeable (i.e. members that only differ in some random perturbation of their initial conditions). There is no clearly established best method for running BMA with a large number of members, which can result in overfitting since BMA estimates one weight per ensemble member. In this case, it is desirable to select a smaller subset of members for BMA, which is discussed further in this chapter.

Section 3.2 details the multi-model ensemble and post-processing methods used on FWDs and non-FWDs. This chapter defines an FWD using the FWD1 methodology from chapter 2. The results are presented in section 3.3, including the effectiveness of post-processing on FWDs, the differences in model performance by model hour and ensemble member, and the sensitivity of post-processing to the selected ensemble. Section 3.4 concludes and discusses possible extensions to operational forecasting.

3.2 Data and methods

a. Multi-model ensemble

The Stony Brook University (SBU) and National Center for Environmental Prediction (NCEP) Short Range Ensemble Forecast (SREF) ensembles are evaluated over the NEUS domain shown in Fig. 2.1 for the 2007 to 2009 warm seasons (April to September). The SBU ensemble consists of 7 Penn State-National Center for Atmospheric Research (NCAR)

Mesoscale Model (MM5; Grell et al. 1994) members and 6 Weather Research and Forecasting (WRF; Skamarock et al. 2008) members at 36- and 12-km grid spacing domains covering the eastern two-thirds of the U.S. and NEUS, respectively (see Fig. 1 in Jones et al. 2007). The SBU ensemble combines different initial conditions and physical parameterizations (convective parameterization, boundary layer, and microphysics) to increase forecast diversity (Table 3.1). The sea surface temperature (SST) initialization uses the U.S. Navy Optimum Thermal Interpolation System (OTIS) SST. Soil moisture for all WRF members and MM5 member four (Table 3.1) is initialized with the North American Mesoscale (NAM) model at 32-km grid spacing, while the remaining MM5 members uses the GFS soil moisture at 0.5 degree spacing.

The NCEP SREF is run at 32-45 km grid spacing at 0300, 0900, 1500, and 2100 UTC (Du et al. 2006). There are four model cores [Eta, Regional Spectral Model (RSM), WRF Nonhydrostatic Mesoscale Model (NMM) and WRF Advanced Research WRF (ARW)]. Initial condition perturbations for the SREF implement a breeding technique similar to Toth and Kalnay (1997), and each core uses different physics (Table 3.1). Except for the Eta and RSM, SREF members sharing the same core can be considered exchangeable.

The SBU and SREF ensembles are verified for the 0000 UTC and 2100 UTC run cycles, respectively, for the region shown in Fig. 2.1. A subset of NEUS is selected based on its high population density and neighboring woodland area (i.e. Catskills, Poconos, Berkshires; Fig 2.1). The predicted 2-m temperature (TEMP) and 10-m wind speed (WNDS) are verified with Automated Surface Observing System (ASOS) observations. 2-m specific humidity (SPHU) is verified for the SBU ensemble only.

Except when specified, the verification and post-processing of FWDs focus on the daytime period (1200-0000 UTC), since this is when atmospheric optimal conditions likely peak for fire initiation. Two adjacent forecast cycles are used; the SBU (SREF) 36 to 48 hour (39 to 51 hour) forecast one day prior to the FWD and the 12 to 24 hour (15 to 27 hour) forecast on the FWD. Model forecast data at the four grid points surrounding the observation sites are bilinearly interpolated to each ASOS location (Jones et al. 2007). Verification is performed using mean error (ME) and mean absolute error (MAE) by threshold in addition to occasionally employing contingency table approaches (Wilks 2011), such as contingency bias (CB) and equitable threat score (ETS).

b. Bias correction methods

Three different bias correction techniques are compared using the ensemble forecasts: linear regression, additive, and cumulative distribution function (CDF). The linear regression uses the forecast variable as the only predictor and does not consider spatial variations [Wilson et al. 2007, their Eq. (4)]. The additive bias correction determines the ME in the training period for each forecast hour and station separately [Wilson et al. 2007, their Eq. (3)] before subtracting it from the most recent forecast. Considering bias by station improves the additive bias correction MAE by 5 to 10 percent. The CDF method comprises two steps; the first adjusts the CDF of the model to the observations over the domain simultaneously (Hamill and Whitaker 2006), and a second step bins the data by terrain elevation (at thresholds between 0 to 50 meters, 50 to 100 meters, 100 to 200 meters, and greater than 200 meters) and dominant land use categories (i.e. Urban, Mixed Forest, Water, etc.) to find residual spatial bias. Residual bias is calculated as the sum of the CDF-corrected forecasts of a particular bin divided by the sum of the observational values. The final bias corrected forecast is the inverse of the bias multiplied by the CDF-corrected forecast for that elevation or dominant land use. This extra step of removing spatial bias improves MAE by an additional 2 to 3 percent.

The sensitivity of model state variable to bias correction method is explored for TEMP and WNDS using a training window length of 14 days. The training window is always kept separate from verification by sorting the entire dataset in chronological order and employing a sliding window approach. This approach begins by selecting the first 14 sorted days to correct the 15th day. The sliding window then proceeds to verify the remainder of the dataset one day at a time by incrementally dropping the oldest day from the training period while adding the most recent event. The comparison of training lengths between 7 and 21 days shows improvement up to, but not beyond 14 days in length (not shown). Additionally, two different methods for training the statistics are examined. The first uses a training window of the most recent consecutive 14 days (sequential training) while the second uses the most recent similar 14 days (conditional training).

There is a discrepancy between ASOS stations that report all sustained winds less than 1.5 m s^{-1} as calm (NWS 1998), and the modeled WNDS. For consistency, all modeled WNDS values below 1.5 m s^{-1} are set to zero. Without this adjustment, the simulated winds less than 1.5 m s^{-1} would result in an artificial high model bias.

c. Bayesian Model Averaging

BMA creates a posterior probability density function (PDF) for each model variable [Raftery et al. 2005 their Eq. (1)] as given by:

$$p(y \mid f_1, \ldots f_k) = \sum_{k=1}^{K} w_k g_k(y \mid f_k) \qquad 3.1$$

where f is the forecast, y is the observation, w_k is the probability of each ensemble member k being the best, and g_k is the conditional PDF of observation y being correct given that member k is best. The frequency distribution g varies depending on the state variable being considered. This dissertation applies BMA similar to that in Raftery et al. (2005); including the assumption that TEMP is normally distributed. However, this dissertation separates bias correction from BMA, and chooses the best bias correction method prior to BMA for each model state variable.

BMA for WNDS is implemented similar to Sloughter et al. (2007) for precipitation and Sloughter et al. (2010) for maximum WNDS, but with a few adjustments. Sloughter et al. (2007, 2010) assumes the variance parameter is linearly related to the model forecast. As with Schmeits and Kok (2010), directly estimating the variance parameter gives similar results, and is used for this dissertation. Sloughter et al. (2010) assumes the gamma distribution's mean follows a linear relationship with the forecasted WNDS. This dissertation uses the bias corrected modeled WNDS for the gamma mean, since running BMA with the linear assumption does not add any additional skill. However, this assumption does not hold when the WNDS is calm, since it results in unacceptable zero valued gamma means. Therefore, all zero valued gamma means are replaced with the average observed WNDS for each ensemble member when the model forecasted WNDS is calm. In this manner, BMA can still calculate an unbiased posterior PDF.

As in Sloughter et al. (2007), the WNDS BMA posterior PDF consists of a mixture model that combines a point mass at zero and the gamma distribution for all other values. The point mass is described by a logistic regression that calculates the probability of the observed

WNDS being calm given the forecast. A power transformation is used to make the WNDS data more Gaussian. The BMA fit to observations over NEUS is optimized during the daytime hours when the modeled WNDS data is transformed to the 3/4th power (not shown).

The model weights and variance for BMA cannot be solved analytically (Raftery et al. 2005), and are estimated using the maximum likelihood technique (Fisher 1922). This method seeks a set of parameter values under which the observed data are most likely to have happened. In this case the DiffeRential Evolution Adaptive Metropolis (DREAM) Markov Chain Monte Carlo (MCMC) algorithm developed by Vrugt et al. (2008) is used. DREAM MCMC is designed to work in a high dimensional space and is more likely to find the global maximum likelihood estimate than the Expectation-Maximization method used in Raftery et al. (2005).

As with bias correction, BMA separates the calibration window from verification, but is applied on a 28-day sliding window of bias corrected model forecasts. This is within the 25 to 30 day training period recommended by Raftery et al. (2005) and Sloughter et al (2007, 2010). BMA is evaluated using conditional and sequential training with two consecutive cycles of ensemble model runs between 1200 UTC and 0000 UTC (SBU 12-24 h and 36-48 h forecasts).

The DREAM MCMC algorithm did not converge when considering all SBU and SREF members, even after 15,000 iterations. Since this is likely caused by overfitting, 5 members are selected from each of the SBU and SREF ensembles. The 5 best SBU members are selected in terms of lowest seasonally averaged MAE within each PBL scheme (Table 3.1). The SREF control members are used, since the perturbed members had higher MAE for TEMP and WNDS (not shown). Here 11 unknown parameters are estimated by iterating DREAM MCMC 2500 times; 10 weights for each member and one variance parameter.

Statistical significance is determined via bootstrapping (Wilks 2011), where resampling with replacement is used to create a larger dataset. 1000 samples from the dataset are randomly chosen in order to calculate the 95% confidence intervals. Additionally, the ensemble mean and smaller ensemble subsets for all cases are calculated by averaging through all hours and stations, before averaging across all members.

3.3 Results

a. Bias correction methodology

The warm season performance of different bias correction methods averaged across all ensemble members are verified for ME (Fig. 3.1a), MAE (Fig. 3.1b), CB (Fig. 3.2a) and ETS (Fig 3.2b). Ensemble mean 2-m TEMP has an increasingly negative mean error (ME) from -1.19 K to -1.96 K between the 289 K and 304 K thresholds (Fig. 3.1a). All bias correction techniques improve the cool bias for all thresholds. However, the linear regression overcorrects for the warmest TEMP by giving little weight to the forecast predictor and too much weight to the intercept parameter. This adjusts the model data closer to the average value (i.e. reduces the forecast variance) as is found in Hamill (2007). As a result, the linear regression has an increasing negative bias for higher TEMP thresholds (ME = -0.8 K and CB = 0.71 at 304 K). The additive bias correction has lower average MAE (2.18 K) and higher ETS (0.39) than the CDF (MAE = 2.22 K; ETS = 0.37) and linear techniques (MAE = 2.51 K; ETS = 0.34) at the 304 K threshold. The differences in MAE and ETS between the linear, CDF and additive techniques are all statistically significant (not shown).

The same verification metrics for ensemble mean 10-m WNDS has a positive bias that increases from 0.64 m s^{-1} to 1.00 m s^{-1} for thresholds between 1.5 m s^{-1} and 5 m s^{-1} (Fig. 3.3a).

The WRF MYJ PBL, SREF-NMM and SREF-ETA members have the largest positive bias (averaging 1.90 m s^{-1} at 5 m s^{-1}, not shown). Excluding these members gives an ensemble bias of ~0.24 m s^{-1} at the 5 m s^{-1} threshold (not shown). The linear bias removal has a growing negative bias with increasing threshold, reaching -0.99 m s^{-1} by 5 m s^{-1} (Fig. 3.3a), while the additive (0.10 m s^{-1}) and CDF (0.20 m s^{-1}) methods have less bias. The linear bias correction is lower (1.76 m s^{-1} Fig. 3.3b) than the CDF (1.79 m s^{-1}) or additive (1.93 m s^{-1}) methods for average MAE. However, the linear method suffers from the same reduction of variance problem as TEMP, causing high MAE (2.28 m s^{-1}) at 5 m s^{-1}. This deficiency with the linear bias correction is better reflected with CB and ETS (Fig. 3.4). The linear method has higher CB than the raw WNDS below the 3.5 m s^{-1} threshold (by 0.1 at 3 m s^{-1}) with lower CB than the raw WNDS above 3.5 m s^{-1} (by 0.92 at 5 m s^{-1}). The linear method has consistently worse ETS than the raw field (by 0.1 at 5 m s^{-1}) while the additive and CDF methods improve upon the raw model ETS (by 0.02 and 0.04, respectively). Although the linear method performs best in terms of MAE for most thresholds, the CDF method is preferred for WNDS due to better performance at higher thresholds and overall better performance when employing contingency verification metrics.

Considerable weight is given to effective bias correction methods at higher WNDS thresholds, since strong winds can contribute to the generation and spread of wildfires (Charney and Keyser 2010). Given the results above, the remainder of this chapter will bias correct TEMP (WNDS) with the additive (CDF) method.

b. Bias correction applied to FWDs

TEMP ME and MAE are presented for FWDs (Fig. 3.5). Results are grouped according to PBL scheme within the SBU ensemble or SREF model core and averaged for several different periods: the warm season, FWDs, FWDs with sequential bias correction (SBC), and FWDs with conditional bias correction (CBC). The ensemble mean has a warm season cool bias of -1.09 K (Fig. 3.5a), with a bias of -2.45 K on FWDs. The SBU's MM5 (Mellor-Yamada) MY and WRF MYJ members have the largest cold bias on FWDs (-3.29 K and -2.77 K, respectively), while the SBU's MM5 MRF and MM5 Blackadar (BLK) members have cool biases of -1.18 K and -1.05 K, respectively. The SBC on FWDs has an ensemble mean cool bias (-0.85 K), while CBC has almost no bias (-0.02 K). The TEMP MAE for FWDs is reduced for CBC (1.85 K) compared to sequential training (2.06 K; Fig. 3.5b). The improvement in the ensemble mean ME and MAE for CBC is statistically significant at greater than 95% confidence compared to SBC.

Ensemble averaged TEMP ME is plotted spatially for the warm season average (Fig. 3.6a), FWDs (Fig. 3.6b), sequentially corrected FWDs (Fig. 3.6c) and conditionally corrected FWDs (Fig. 33.6d). This spatial plot is created by interpolating model biases for each station to a 0.25° latitude by 0.25° longitude grid. Both model and observation must exceed the 298 K threshold for the ME to be calculated, although results for other thresholds are similar (not shown). Warm season biases are generally negative (-4.0 K to 0.4 K), but cool biases are larger on FWDs (-7.2 K to -1.5 K). SBC reduces the spread of potential biases (-4.1 K to -0.2 K), but CBC removes the magnitude and spread of the negative bias more effectively (-2 K to 1.2 K).

SBU and SREF ensemble mean TEMP ME is plotted by forecast hour for the warm season raw, FWD raw, FWD with SBC and FWD with CBC (Fig 3.7). The SBU (SREF) ensemble exhibits almost no TEMP bias around forecast hour 9 (6) and forecast hour 33 (33), which coincides with the early morning hours before sunrise. Conversely, SBU (SREF) cool TEMP biases peak around forecast hour 23 (20) and forecast hour 47 (44), which is concurrent with the time of maximum heating. Most critical, SBU (SREF) model biases for TEMP are 2 K (2.4 K) cooler on FWDs compared to the warm season average. Therefore, the significantly greater negative TEMP model biases found on FWDs are sensitive to the diurnal cycle and peak during the afternoon when atmospheric conditions are most favorable to fire initiation and spread.

Warm season biases with 2-m SPHU (Fig. 3.8) vary between 0.04 g kg^{-1} for the WRF YSU PBL members to 4.40 g kg^{-1} with the MM5 MY PBL (Fig. 3.8a). Schwitalla et al. (2008) noted a strong daytime wet bias in SPHU using the MM5 MY PBL, and attributed this bias to weaker vertical mixing in the boundary layer. FWDs have a larger positive moisture bias than the warm season average, with a 0.74 g kg^{-1} (0.06 g kg^{-1}) bias in the ensemble mean after SBC (CBC). The difference between SBC and CBC ensemble mean ME and MAE is statistically significant exceeding 95% confidence levels.

Warm season model biases vary between ensemble members for 10-m WNDS (Fig. 3.9). For instance, the MM5 members have a small negative bias (~-0.12 m s^{-1}), while the WRF-MYJ, SREF-NMM and SREF-ETA members have significant positive biases (~1.22 m s^{-1}). The ensemble mean on FWDs with SBC has a ME of -0.55 m s^{-1}, which is corrected with CBC (~0 m s^{-1}). The improvement in MAE between CBC (1.49 K) and SBC (1.53 K) is not statistically significant, but the improvement in ME exceeds 95% confidence. Nonetheless, improving ME is important even without benefit to MAE, since model bias can affect BMA's ability to calibrate an ensemble (section 3.3c).

WNDS biases for the ensemble mean are presented spatially on FWDs (Fig. 3.10). The warm season average ensemble mean overestimates WNDS (-0.7 m s^{-1} to 2.5 m s^{-1}), and to a lesser extent for FWDs (-0.9 m s^{-1} to 2 m s^{-1}). SBC overcorrects WNDS bias, while CBC performs best. The lingering spatial model biases after CBC appear to be related to geographical location (Fig. 8d). For instance, the regions around New York City (KNYC) and Providence (KPVD) to Worchester (KORH) exhibit negative biases (-0.7 m s^{-1} to -0.4 m s^{-1}), while the mountainous areas of the Poconos and Berkshires have positive biases (0.5 m s^{-1} to 1 m s^{-1}). Although the CDF method bins together dominant landuse type and elevation, this correction could be improved by considering other spatial features (Mass et al. 2008; Kleiber et al. 2011).

c. Bayesian Model Averaging

BMA is applied to FWDs after bias correction using a 28-day sliding window. BMA with sequential training is called BMA-ST and BMA implemented with conditional training will be referred to as BMA-CT.

Probability Integral Transform (PIT, Raftery et al. 2005) of FWDs with BMA-ST and BMA-CT are compared to bias correction only rank histograms for 2-m TEMP (Fig. 3.11a,b) and 10-m WNDS (Fig. 3.11c,d). Additionally, the reliability index (RI) of Delle Monache et al. [2006 their Eq. (2)] is calculated to quantify the impact of bias and underdispersion, where lower values represent a flatter rank histogram. Bias correction without BMA results in severely underdipsersed (U-shaped) TEMP and WNDS forecasts for all FWDs (grey bars in Fig 3.11). Additionally, SBC has a backwards "L" shaped histogram, indicative of a negative bias for both TEMP and WNDS (Fig. 3.11b,d). Applying BMA after bias correction removes most of the underdispersion for TEMP and WNDS (black bars in Fig 3.11) and also greatly lowers RI index values (by 29.5% for TEMP and 57.4% for WNDS on average). However, BMA-ST is still negatively biased (Figs. 3.11b,d) compared to BMA-CT (Figs. 3.11a,c) for all FWDs with TEMP and WNDS. This is reflected in the lower RI values for conditional versus sequential training (by 23% for TEMP and 3.2% for WNDS), suggesting that BMA cannot correct for lingering model biases.

The influence of BMA is shown probabilistically using reliability plots for FWDs exceeding the 4 m s^{-1} and 6 m s^{-1} thresholds (Fig. 3.12). Ensemble underdispersion is evident for all cases using bias correction and no BMA, with lower (higher) forecasted probabilities being too low (high) compared to the observed relative frequency. BMA improves ensemble reliability for all thresholds (black versus grey lines). However, the negative bias caused by sequential training on FWDs (Fig. 3.9a; Fig. 3.10c) results in less accurate probabilistic forecasts for BMA-ST (Figs. 3.12b,d) compared to BMA-CT (Figs. 3.12a,c).

The dimensionless and positively oriented Brier Skill Score [BSS; Wilks 2011, his Eq. (8.37)] is used to assess probabilistic skill of BMA-CT and BMA-ST referenced against SBC with no BMA for multiple thresholds. FWDs are evaluated between 286 K and 300 K for TEMP and 1.5 m s^{-1} to 5.2 m s^{-1} for WNDS (Fig 3.13). The stations, models, and days considered are identical for bias correction, BMA-ST and BMA-CT, allowing for a direct comparison among all three post-processing methods. All thresholds for TEMP (Fig. 3.13a) and WNDS (Fig. 3.13b) have BMA BSS values greater than zero at 95% confidence; demonstrating that BMA improves probabilistic scores regardless of training period. Results for TEMP on FWDs reveal statistically significant improvement in BMA-CT compared to BMA-ST for all thresholds (Fig. 3.13a). However, there is little difference for WNDS between BMA-CT and BMA-ST on FWDs (Fig. 3.13b).

d. Sensitivity of BMA to forecast hour

The impact of forecast hour on BMA is explored on FWDs by rerunning BMA for 2-m TEMP between SBU forecast hours 3 and 48 (6 to 51 hour SREF). In order to obtain two diurnal cycles for FWDs, two subsequent model runs are post-processed and combined. BSS is used to assess BMA performance for 2-m TEMP by hour with SBC as a reference (Fig. 3.14). BMA improves all probabilistic results regardless of forecast hour, with statistically significant results for all times except hour 30. The BSS for BMA-CT on FWDs varies diurnally, and are nearly out of phase with BMA-ST. As a result, the probabilistic results with BMA-CT on FWDs are significantly better than those with BMA-ST between 1500 and 0000 UTC, but degrade BMA-CT compared to BMA-ST between SBU forecast hours 15 and 24 (both exceeding 95% confidence). The better performance of BMA-CT compared to BMA-ST between hours 15-24 and 39-48 is a result of better performance with the CBC during the day (Fig. 3.7). Regardless, both BMA-ST and BMA-CT generally improve upon SBC.

e. Sensitivity of BMA to ensemble member selection

Unfortunately, the DREAM MCMC algorithm does not converge on confident parameter estimates for the entire 34-member ensemble, so a subset of 10 members is selected. However, the best way to construct a deterministically and probabilistically skillful ensemble using BMA is not immediately obvious. Raftery et al. (2005) showed that BMA weights each member based on its uniqueness and skill using different model initial conditions. However, unique model cores with different parameterized physics may also have useful information. Therefore, members not included in section 3.3c-d may have useful information for BMA.

In order to test BMA's sensitivity to member selection, 5 SBU and 5 SREF members are chosen from the total ensemble as described in section 3.2 (B10). This ensemble is compared to 5 SBU and 5 SREF members that are randomly selected from the total ensemble 1000 times (R10). R10 and B10 are compared for 2-m TEMP and 10-m WNDS after bias correction for the 2007 to 2009 warm seasons.

The difference between B10 and R10 in terms of MAE, variance, and BSS (referenced against R10) is calculated by threshold for TEMP and WNDS (Fig. 3.15). B10 has significantly lower MAE for both TEMP and WNDS (Fig. 3.15a,b), but is more underdispersed (Fig. 3.15c,d). This affects the probabilistic results (Fig. 3.15e,f), particularly for WNDS, where the R10 ensemble has better probabilistic skill. Comparing B10 and R10 is an interesting test case for ensemble calibration on FWDs, since BMA should correct for underdispersion while preserving the ensemble's skill.

The sensitivity of BMA to member selection for FWDs is tested by comparing 10 randomly drawn members from the total ensemble (R10-BMA) to the 10 best members (B10-BMA). The R10-BMA and B10-BMA results are presented in terms of BSS using 2-m TEMP for FWDs (Fig. 3.16) referenced against the SBC of R10-BMA. B10-BMA is significantly better than R10-BMA on FWDs for thresholds between 287 and 289 K and 293 K (Fig. 3.16). Therefore, the B10 ensemble should be used over R10 only when BMA is applied to correct for ensemble underdispersion.

Previous studies have shown the benefits of combining members from different ensembles (Cartwright and Krishnamurti 2007; Candille 2009; Zsoter et al. 2009) since increased forecast diversity captures uncertainties in both the initial conditions and model physics. Three additional ensembles derived from B10-BMA are used to test the benefits of combining the SBU and SREF. These ensembles include the 5 SBU members (B5-SBU-BMA), the 5 SREF members (B5-SREF-BMA) and a combined ensemble (B5-ALL-BMA) that for each FWD has a random draw of two SBU members, two SREF members, and one member that have an equal chance of being either SBU or SREF. Although the ensemble size is reduced in this setup, the random drawing will eventually sample all 10 members. This is confirmed by rerunning B5-ALL-BMA several times with similar probabilistic results.

BSS by threshold for FWDs (Fig. 3.17) are plotted by threshold for 2-m TEMP B5-SBU-BMA and B5-SREF-BMA, with B5-ALL-BMA as the reference. Negative BSS values correspond to better probabilistic scores for the combined SBU/SREF ensemble. FWDs benefit from a combined SBU/SREF ensemble, except at higher thresholds (greater than 298 K), where sample size is more limited.

In general, selecting skillful and unique members from both the SBU and SREF ensembles for use in BMA improves probabilistic results. This suggests that model diversity is important, even with a calibration method like BMA. Given previous results, the members selected for B10 is a good choice, but perhaps not the best one. Fraley et al. (2010) noted that exchangeable members receive similar BMA weights, suggesting that the overfitting problem can be partially solved by fixing the weights across all SREF members that share the same model core (except for the Eta and RSM). More study is required to test the applicability of running BMA on a 34-member ensemble with the exchangeable member constraint.

3.4 Summary and conclusions

The post-processing of 2-m temperature (TEMP) and 10-m wind speed (WNDS) for the combined 13-member Stony Brook University (SBU) and 21-member National Centers for Environmental Prediction (NCEP) Short Range Ensemble Forecast (SREF) ensemble is evaluated using different bias correction methods [additive, linear and cumulative distribution function (CDF)] and Bayesian Model Averaging (BMA). Furthermore, the sensitivity of model biases and post-processing are explored for fire weather days (FWDs) over a subset of the Northeast United States. Post-processing with 2-m SPHU (SBU ensemble only) is also explored.

FWDs exhibit different model biases for 2-m TEMP and 10-m WNDS compared to the warm season average, which can degrade attempts to calibrate an ensemble. Therefore, bias correction and Bayesian Model Averaging (BMA) are implemented using a training window of most recent 14 similar days (conditional training) and compared to training using the most recent 14 consecutive days (sequential training). Only daytime forecast hours are used (1200 to 0000 UTC) for two subsequent model runs initialized at 0000 UTC (2100 UTC for the SREF) on the day of and the day before the FWD. Additionally, BMA uses a smaller ensemble consisting of the 5 best performing SBU members (in terms of mean absolute error) with a unique PBL and the 5 SREF control members.

The optimal bias correction depends on the model state variable. The CDF (additive) bias correction performs best for 10-m WNDS (2-m TEMP). On average, simulated FWDs have biases that are cooler, moister and less windy than the typical warm season average bias. As a result, conditional bias correction (CBC) results in significant improvements in mean error (ME) over sequential bias correction (SBC) for TEMP, WNDS, and SPHU on FWDs. There are also statistically significant improvements in mean absolute error (MAE) on FWDs for 2-m TEMP and 2-m SPHU.

BMA is used to evaluate the benefit of conditional training (BMA-CT) compared to sequential training (BMA-ST). Statistically significant improvement in Brier Skill Scores (BSS) with BMA-CT is most evident on FWDs for 2-m TEMP. WNDS results with BMA-CT and BMA-ST are similar probabilistically for FWDs. Regardless of the model state variable; BMA-ST and BMA-CT almost always improve ensemble probabilistic value compared to SBC without BMA. This suggests that BMA can still provide probabilistic value even when the training period is not entirely representative of the day being validated.

BMA performance on FWDs varies as a function of forecast hour. For FWDs, this is likely caused by a more effective CBC during the daytime hours. This suggests a potential physics problem in the lower level of the modeled atmosphere in the presence of short wave radiation; perhaps caused by the parameterized planetary boundary layer, land surface model, or radiation scheme. Given the sensitivity of bias correction to forecast hour, it may be beneficial to run BMA separately for the daytime and nighttime hours.

The sensitivity of BMA to the members selected is analyzed by comparing the 10-member ensemble used earlier (B10) to 10 randomly selected members (R10). Over the warm season average, B10 performs better in terms of MAE for WNDS and TEMP, but is consistently more underdispersed than R10. This negatively affects the probabilistic value of B10 compared to R10. A similar experiment is run for BMA, where 10 members are randomly selected for each FWD (R10-BMA) and compared to the best ensemble (B10-BMA). In this case, B10-BMA usually has more probabilistic value than R10-BMA, since the underdispersion is corrected. This implies the members selected for BMA are important, and that picking skillful and unique members can benefit post-processing.

The SBU and SREF ensembles are combined in this chapter, since additional model cores are expected to bring about greater probabilistic accuracy. This hypothesis is tested by comparing two ensembles consisting of 5 SBU and 5 SREF members each (B5-SBU-BMA and B5-SREF-BMA, respectively), to a third ensemble created by randomly drawing from the first two (B5-ALL-BMA). With TEMP, there is a benefit associated with post-processing the combined SBU and SREF ensemble on FWDs.

The presence of conditional model biases on fire weather days underscores the importance of similar day post-processing. It is critical to to predict FWDs operationally and understand why model biases vary temporally. Event-based post-processing has been developed (Hamill and Whitaker 2006; Mass et al. 2008; Delle Monache et al. 2011), and it might be beneficial to experiment with RELH, TEMP, or drought analog bias corrections for future use in operations. FWDs represent a unique challenge in analog bias correction, since multiple model state variables should be employed to gather the ideal training dataset for post-processing. However, it is not intuitive how such an analog post-processing method should be constructed. Although more complex, cluster analysis could also be useful operationally in identifying the synoptic flow patterns common to FWDs a priori.

Unfortunately, little research has been done with the predictability of fire weather parameters over the Northeast United States. However, Mölders (2010) has shown that WRF can be used to create a reasonably well forecast National Fire Danger Rating System (NFDRS) over Interior Alaska. Therefore it might be feasible to create an operational FWD bias correction method that “turns on” when FWDs conditions (i.e. defined by widespread low humidity and strong winds) are predicted within the next 48 hours. This threshold would have to be adjusted based on the original model biases.

Since FWDs are generally dry, they provide a good test case to explore potential sources of structural model error in the parameterized model physics. A next step would be to examine the regimes associated with FWDs, and relate these flow patterns to the model biases. Although this would not isolate the bias source in the model physics, it can pinpoint a difficult to forecast regime. This topic is discussed further in section 6.3.

Table 3.1: Description of the SBU and SREF ensembles, including model used, microphysical schemes, PBL schemes, radiation schemes, cumulus schemes and initial conditions. *Includes those members for BMA only experiments; SREF uses the control member only.

Stony Brook University (SBU) and Short Range Ensemble Forecast Members (SREF)						
Members	**Model**	**Microphysics**	**PBL Scheme**	**Radiation**	**Cumulus**	**Initial Condition**
1	MM5	Simple Ice	MY	CCM2	Betts Miller	GFS
2	MM5	Reisner	MY	Cloud Radiation	Grell	WRF-NMM
3*	MM5	Simple Ice	MY	CCM2	Kain Fritsch	CMC
4	MM5	Simple Ice	MRF	Cloud Radiation	Grell	WRF-NMM
5*	MM5	Reisner	MRF	Cloud Radiation	Kain Fritsch	GFS
6	MM5	Simple Ice	Blackadar	CCM2	Grell	NOGAPS
7*	MM5	Simple Ice	Blackadar	CCM2	Grell	GFS
8	WRF-ARW	WSM3	MYJ	RRTM	BMJ	WRF-NMM
9	WRF-ARW	WSM3	MYJ	RRTM	Kain Fritsch	NOGAPS
10*	WRF-ARW	WSM3	MYJ	RRTM	Kain Fritsch	GFS
11	WRF-ARW	Ferrier	YSU	RRTM	Kain Fritsch	WRF-NMM
12	WRF-ARW	Ferrier	YSU	RRTM	Grell	GFS
13*	WRF-ARW	WSM3	YSU	RRTM	BMJ	NOGAPS
14-16*	WRF-ARW	Ferrier	YSU	RRTM	Kain Fritsch	GFS
17-19*	WRF-NMM	Ferrier	MYJ	GFDL	BMJ	GFS
20-24*	ETA 1	Ferrier	MY	GFS Radiation	BMJ	NAM
25-29*	ETA 2	Ferrier	MY	GFS Radiation	Kain Fritsch	NAM
30-34*	RSM	Zhao	MRF	GFS Radiation	SAS/RAS	GFS

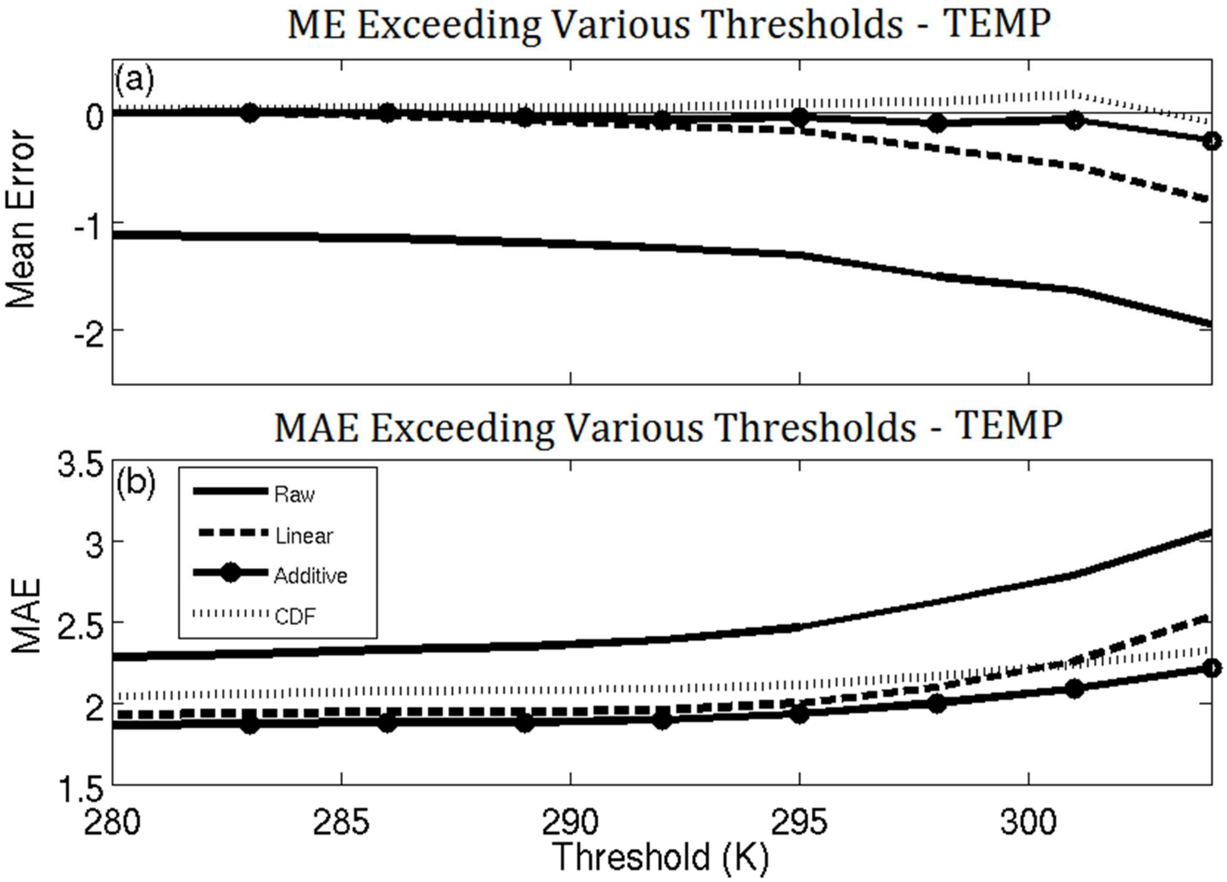

Figure 3.1: Surface 2-m temperature (TEMP; degrees K) (a) mean error (ME) and (b) mean absolute error (MAE) by threshold for raw (solid), linear bias correction (dashed), additive bias correction (filled circles) and CDF bias correction (dotted). Thin black horizontal line in (a) denotes zero bias.

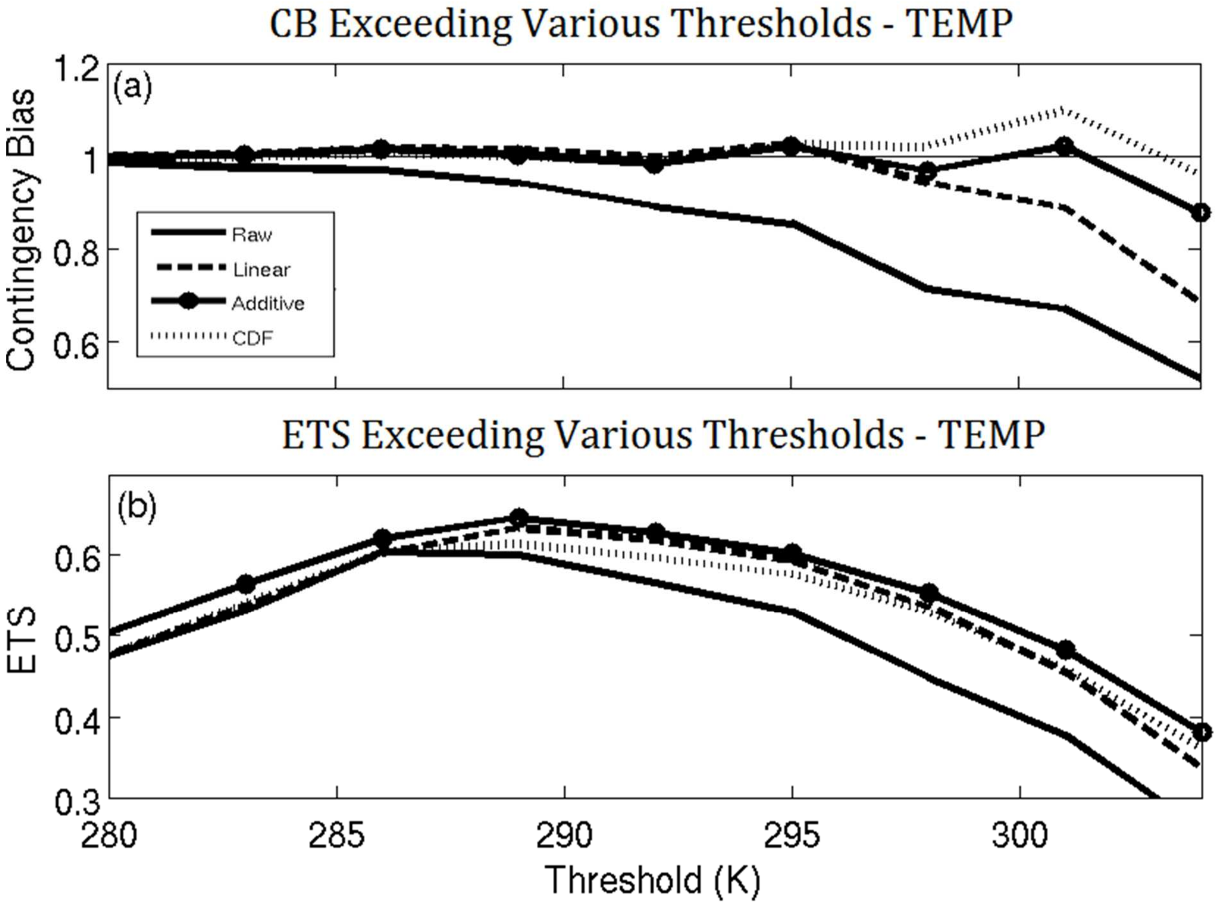

Figure 3.2: Same as figure 3.1, but for contingency bias (CB; a) and equitable threat score (ETS; b).

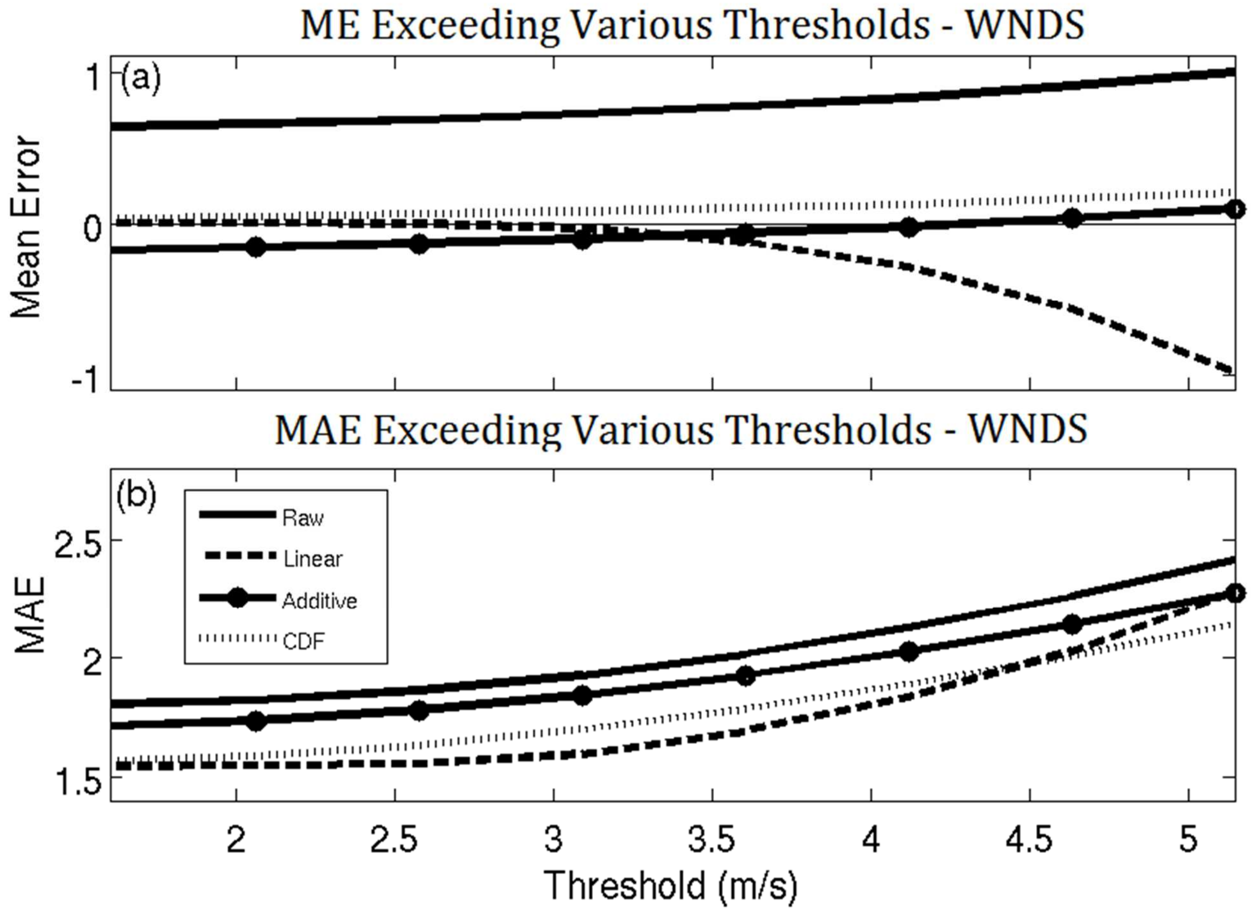

Figure 3.3: Same as figure 3.1, but for 10-m wind speed (WNDS; m s^{-1}) ME (a) and MAE (b).

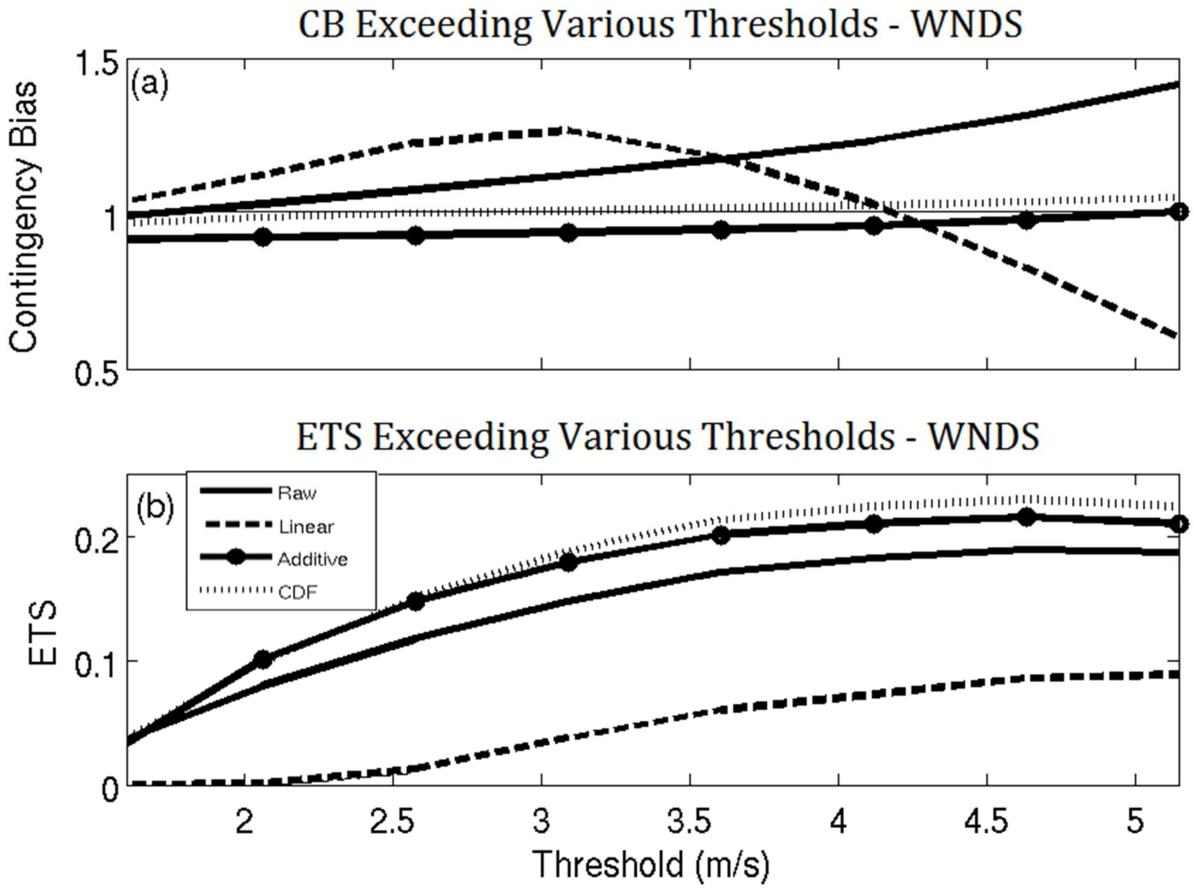

Figure 3.4: Same as figure 3.1, but for WNDS CB (a) and ETS (b).

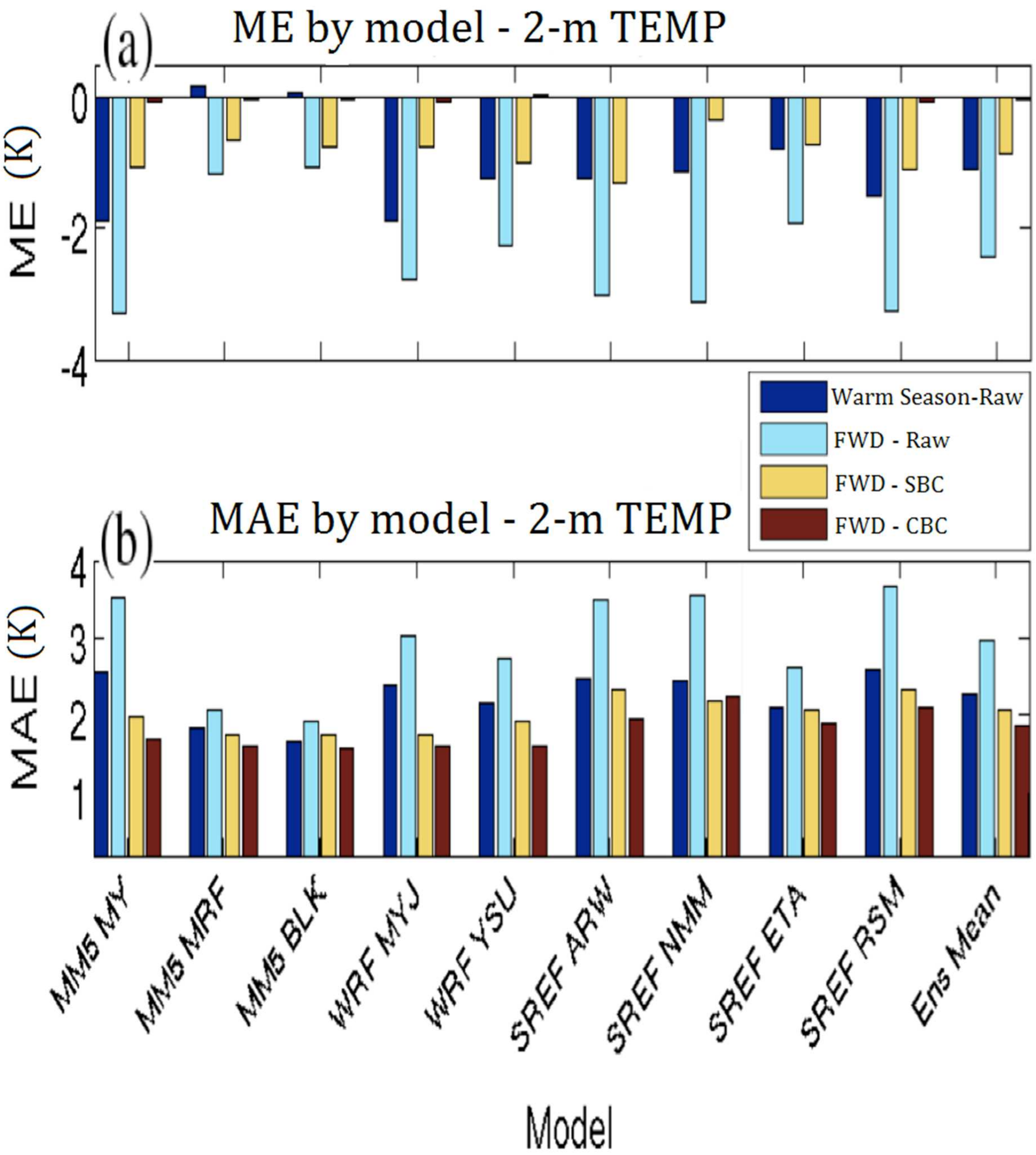

Figure 3.5: Bar plots of 2-m TEMP ME (a) and MAE (b) within each ensemble member subgroup that share the same PBL scheme in MM5/WRF and model core in SREF showing warm season raw (dark blue), FWDs (light blue), FWDs with sequential bias correction (SBC; yellow), and FWDs with conditional bias correction (CBC; red).

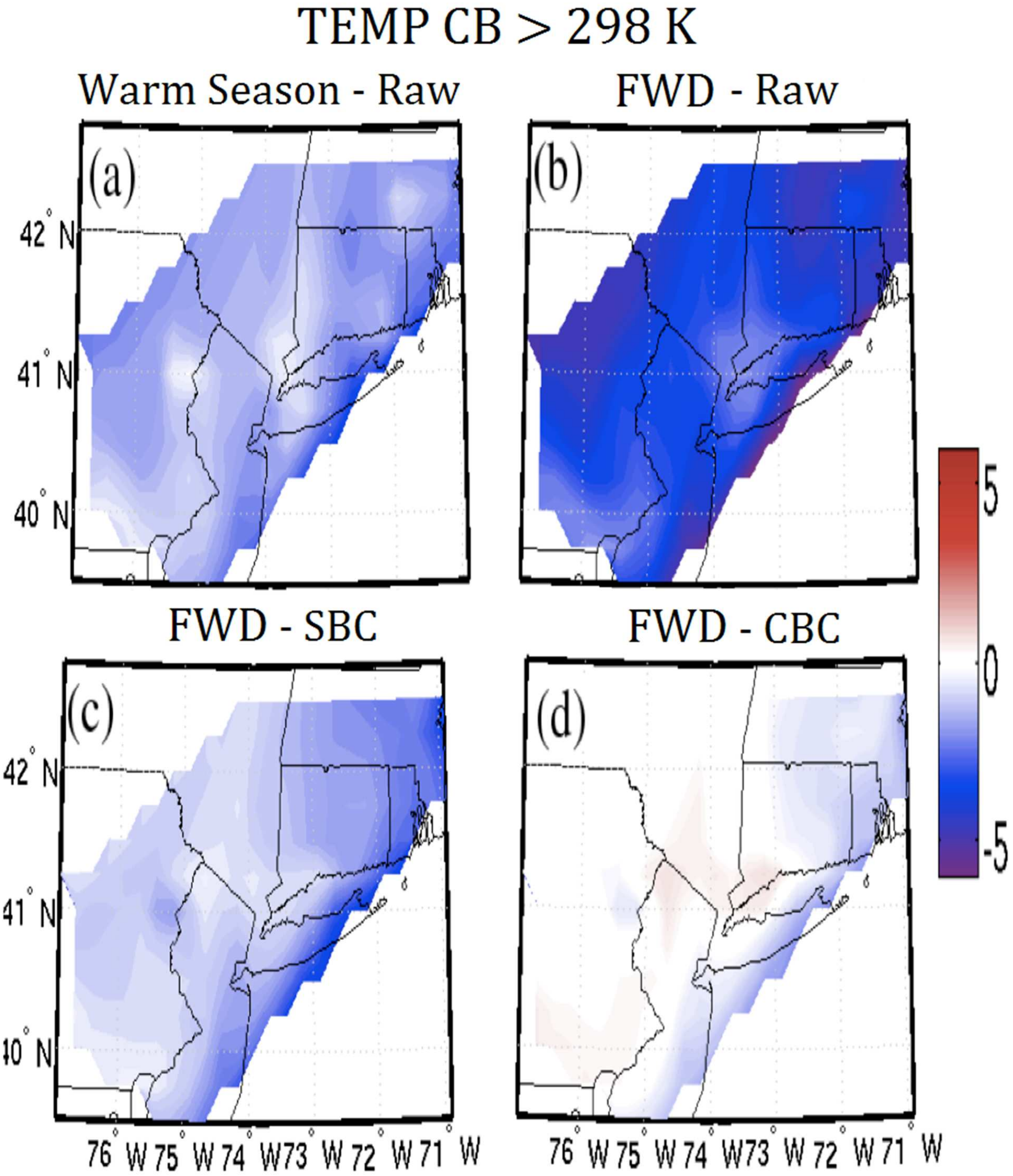

Figure 3.6: Spatial 2-m TEMP ensemble CB for the threshold exceeding 298 K for the (a) raw warm season, (b) FWDs, (c) FWDs with SBC, and (d) FWDs with CBC.

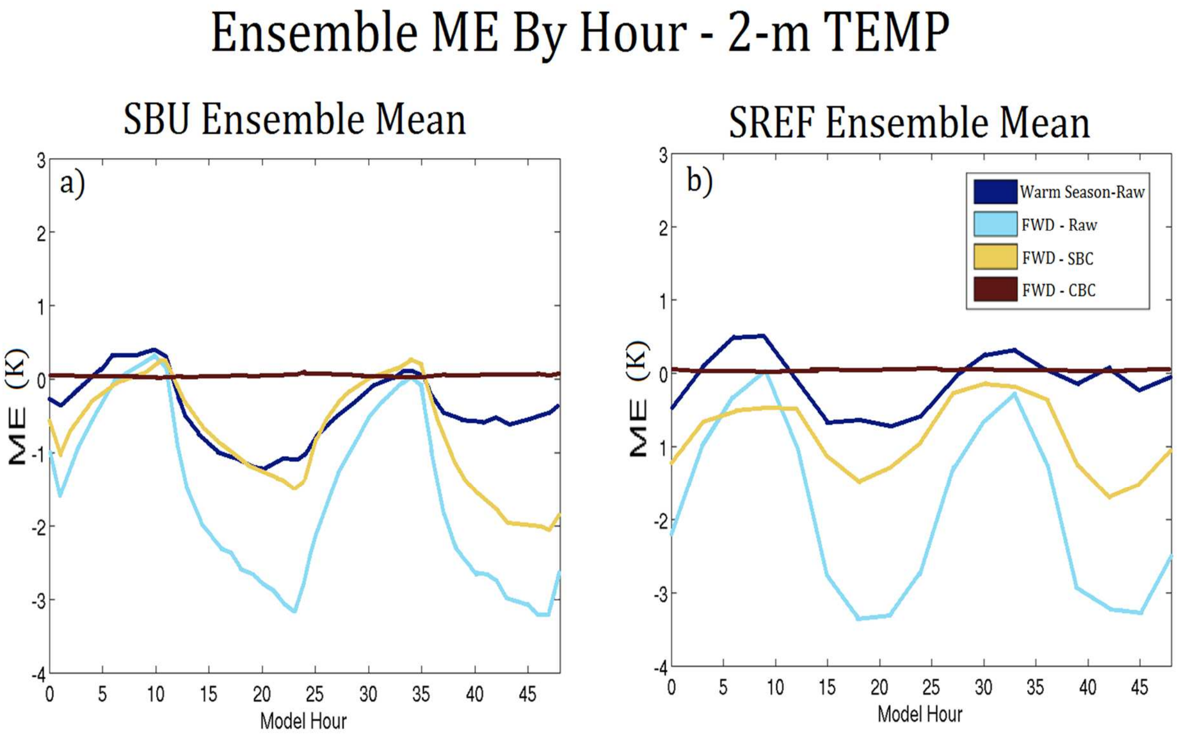

Figure 3.7: 2-m TEMP ensemble ME by hour for the SBU ensemble mean (a) and SREF ensemble mean (b) warm season raw (blue), FWDs raw (cyan), SBC (yellow) and CBC (red).

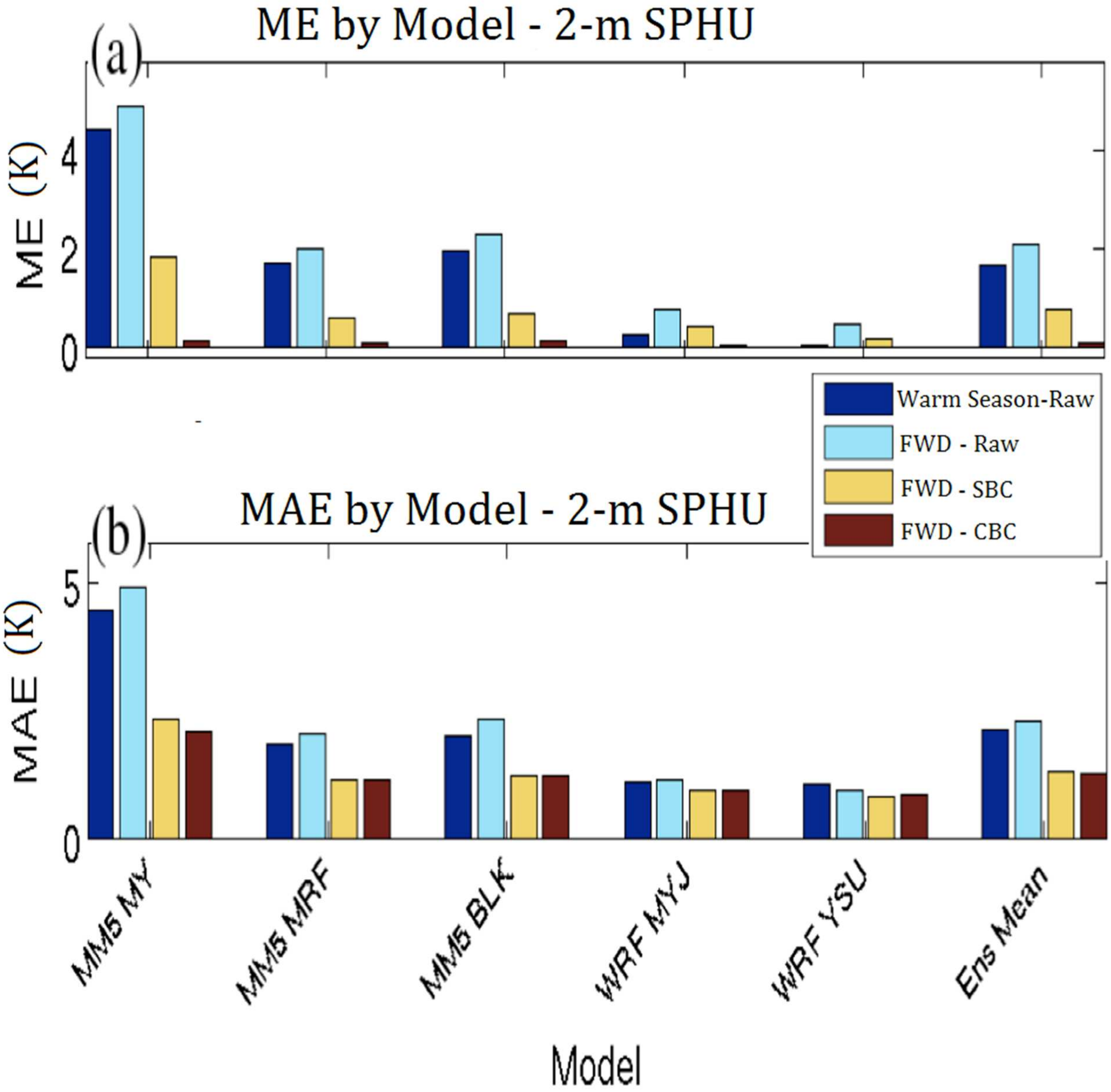

Figure 3.8: Same as figure 3.5, but for 2-m SPHU using only the SBU ensemble.

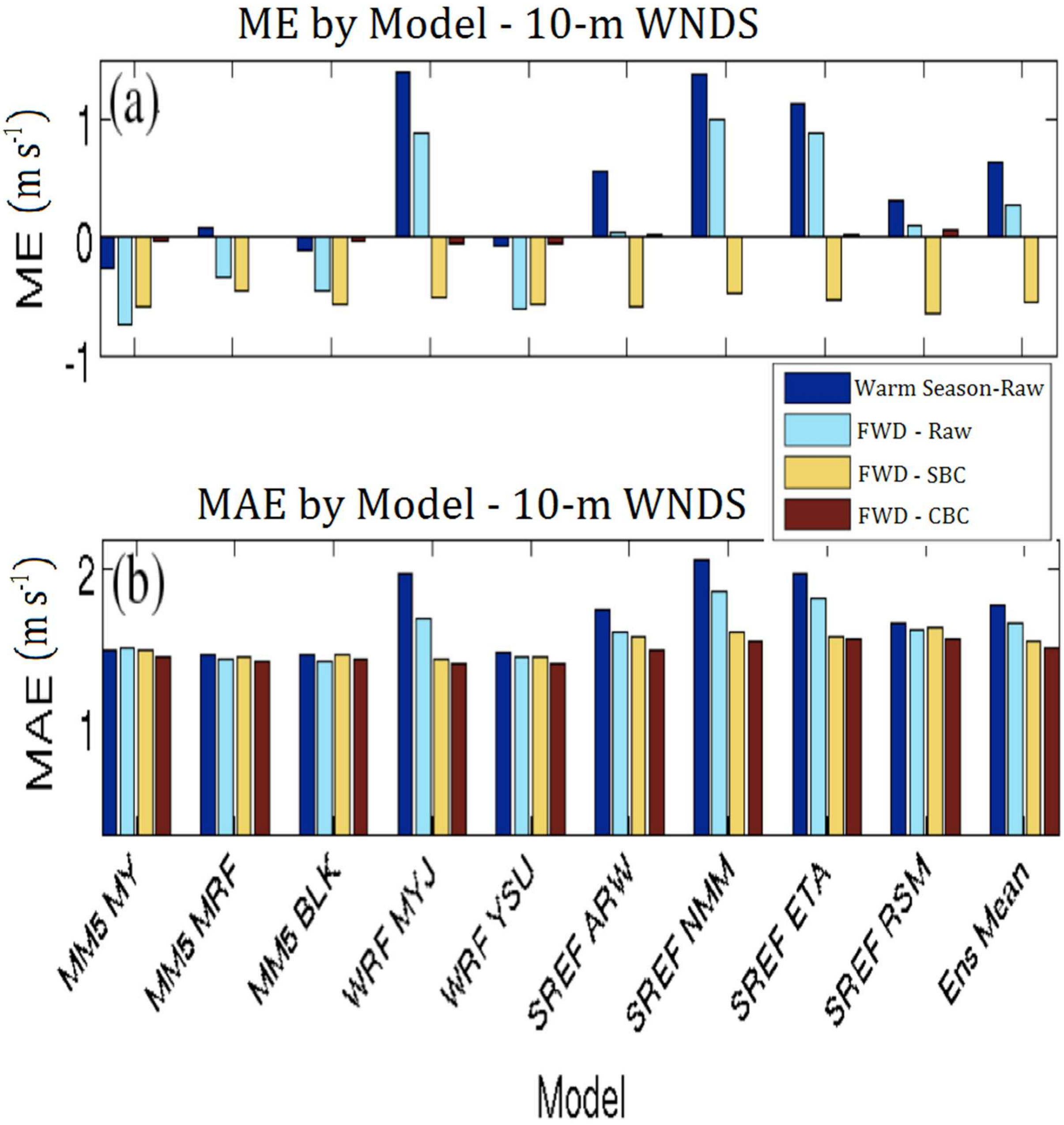

Figure 3.9: Same as figure 3.5, but for 10-m WNDS.

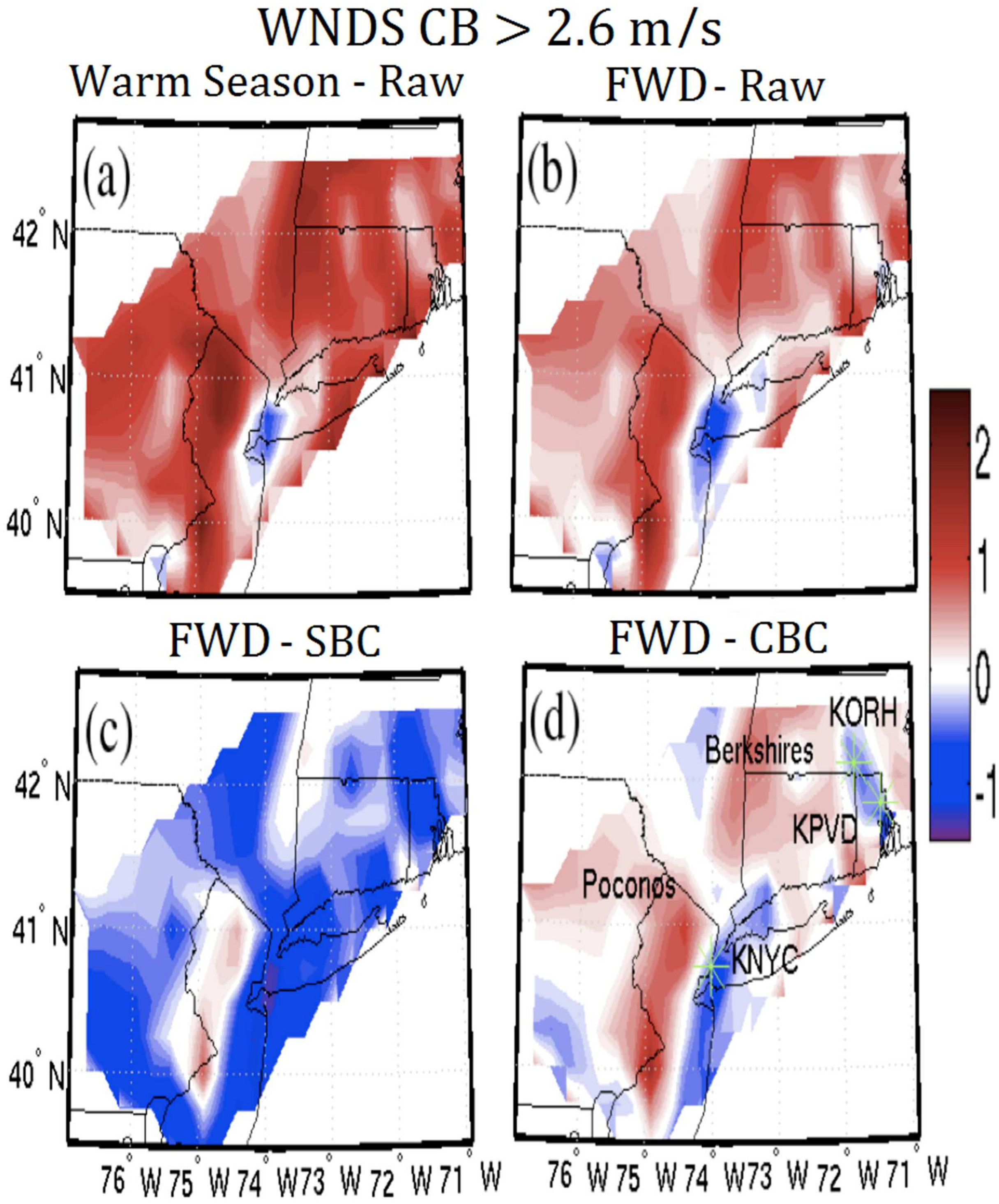

Figure 3.10: Same as figure 3.6, but for 10-m WNDS. In (d), the green asterisks (*) indicate the location of New York City, NY (KNYC), Providence, RI (KPVD) and Worchester, MA (KORH). The Poconos and Berkshire Mountains are also labeled.

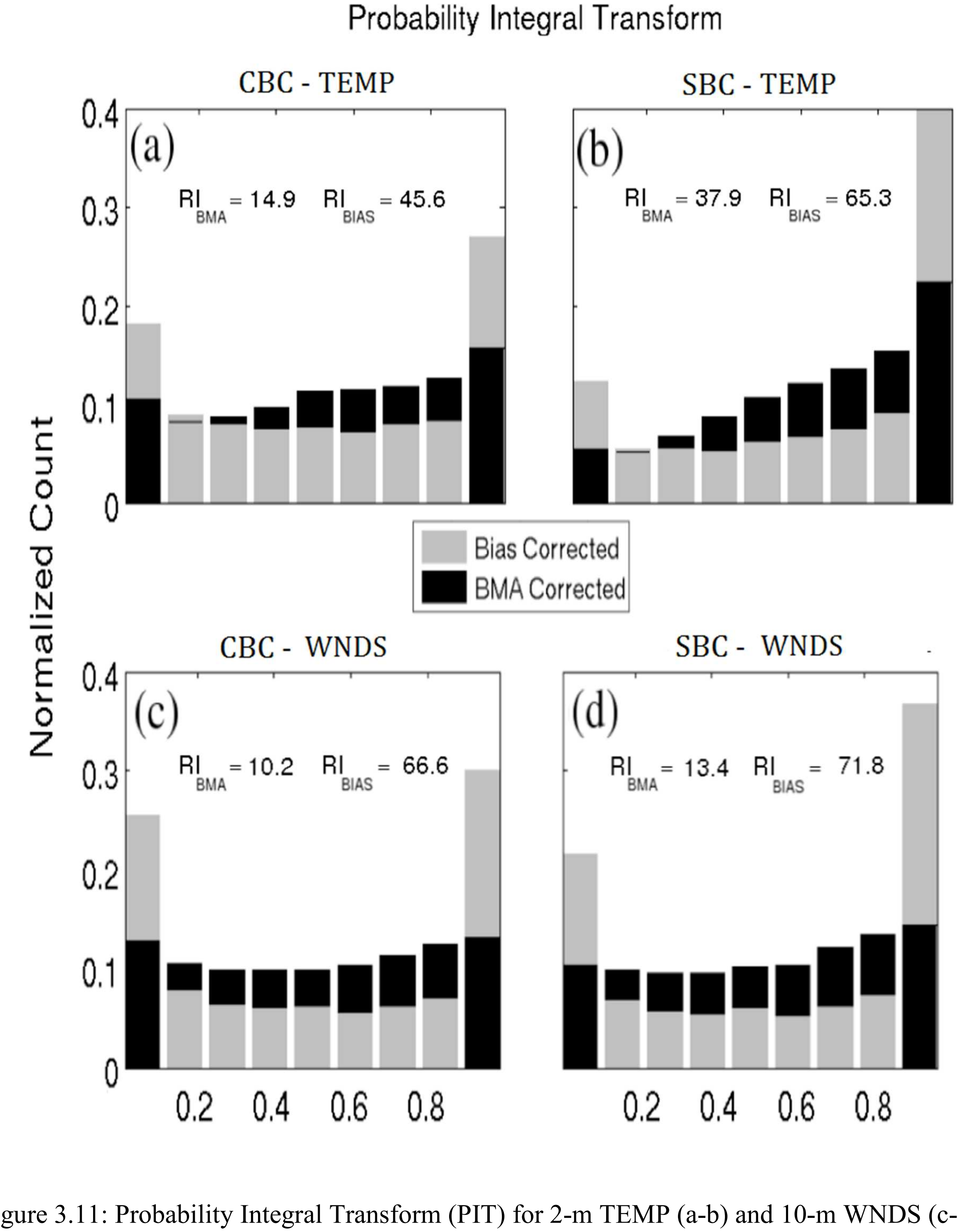

Figure 3.11: Probability Integral Transform (PIT) for 2-m TEMP (a-b) and 10-m WNDS (c-d) showing bias (gray) and bias+BMA (black) corrected prediction during FWDs using conditional (a,c) and sequential (b,d) training. Reliability index for each rank histogram (see text for full description) is also included.

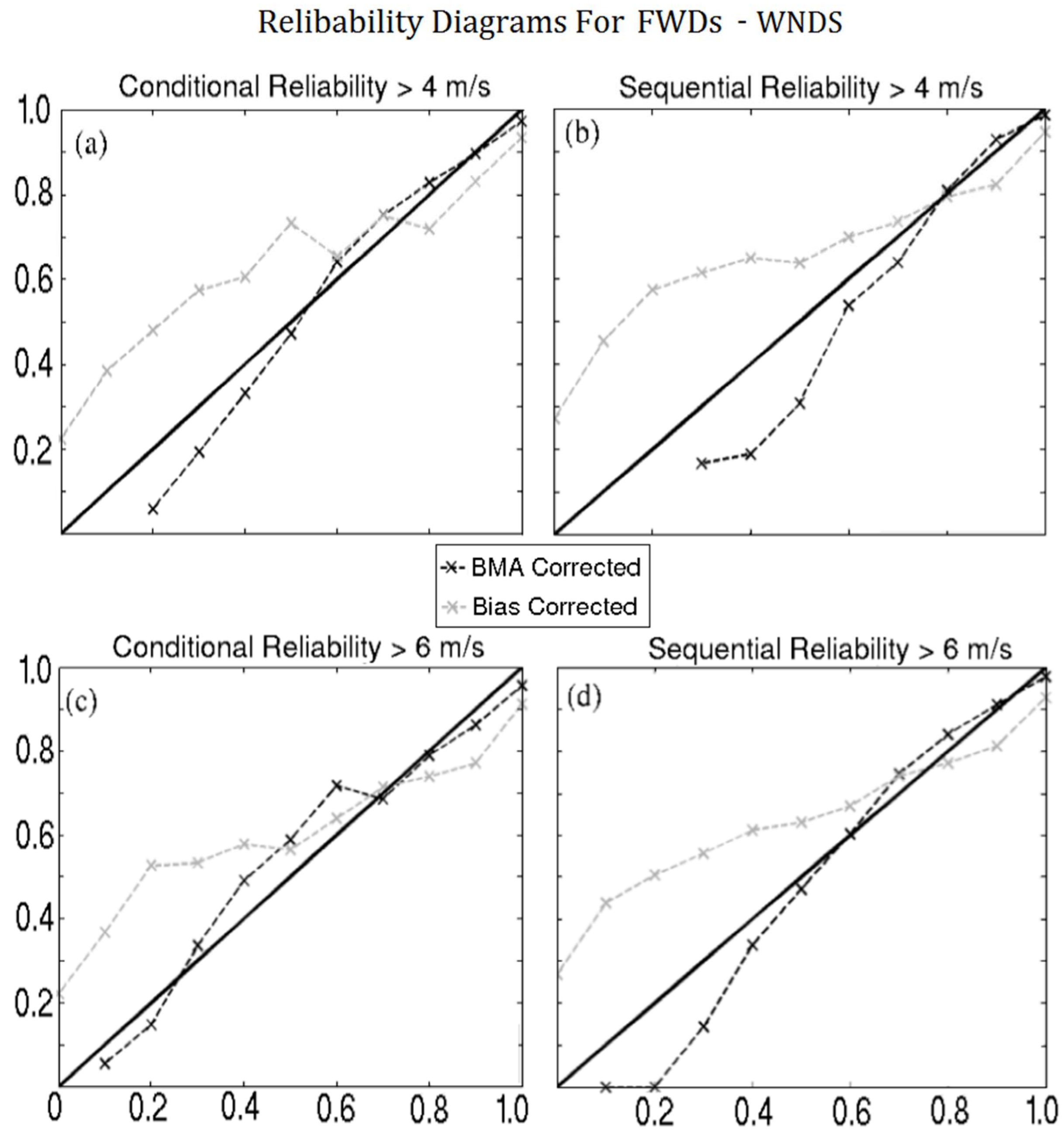

Figure 3.12: Reliability diagrams for 10-m WNDS exceeding 4 m s^{-1} for (a) conditional and (b) sequential training and 6 m s^{-1} for (c) conditional and (d) sequential training. The BMA corrected (black dashed) and bias corrected (gray dashed) is compared with the 1:1 line (solid black).

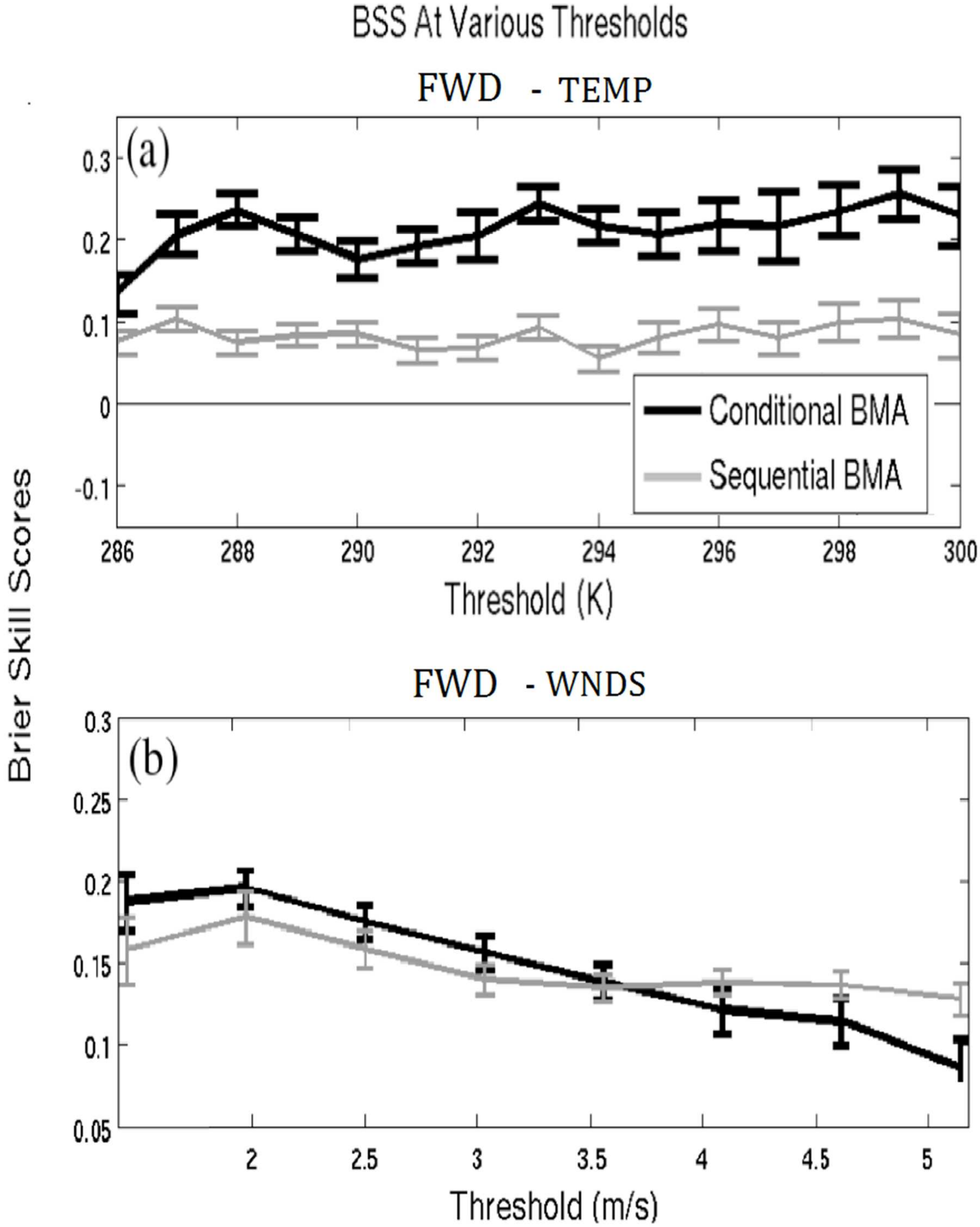

Figure 3.13: FWD brier skill scores for 2-m TEMP (a) and 10-m WNDS (b) using a sequential (gray) and conditional (black) BMA referenced against SBC. Thin black horizontal line denotes zero BSS for (a).

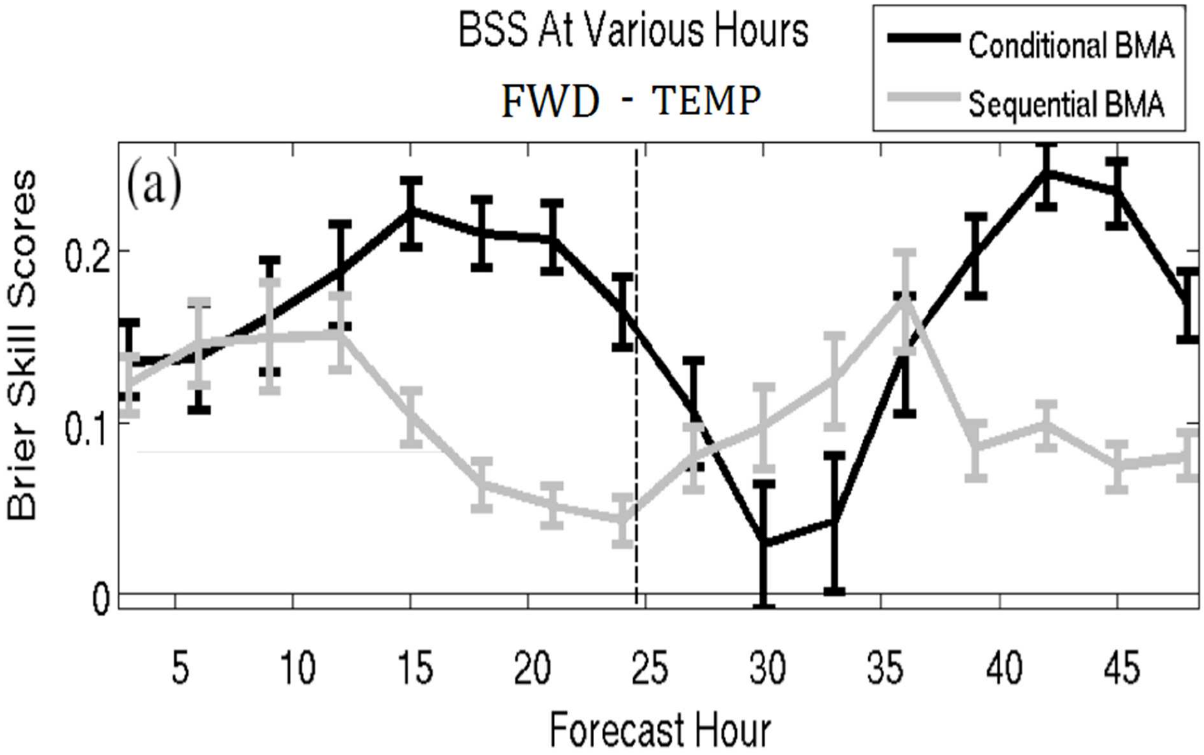

Figure 3.14: Brier skill scores for 2-m TEMP by forecast hour for FWDs using sequential (gray) and conditional (black) BMA referenced against SBC. The thin black horizontal line denotes zero BSS, while the dashed black vertical line separates the day zero forecast from the day one forecast.

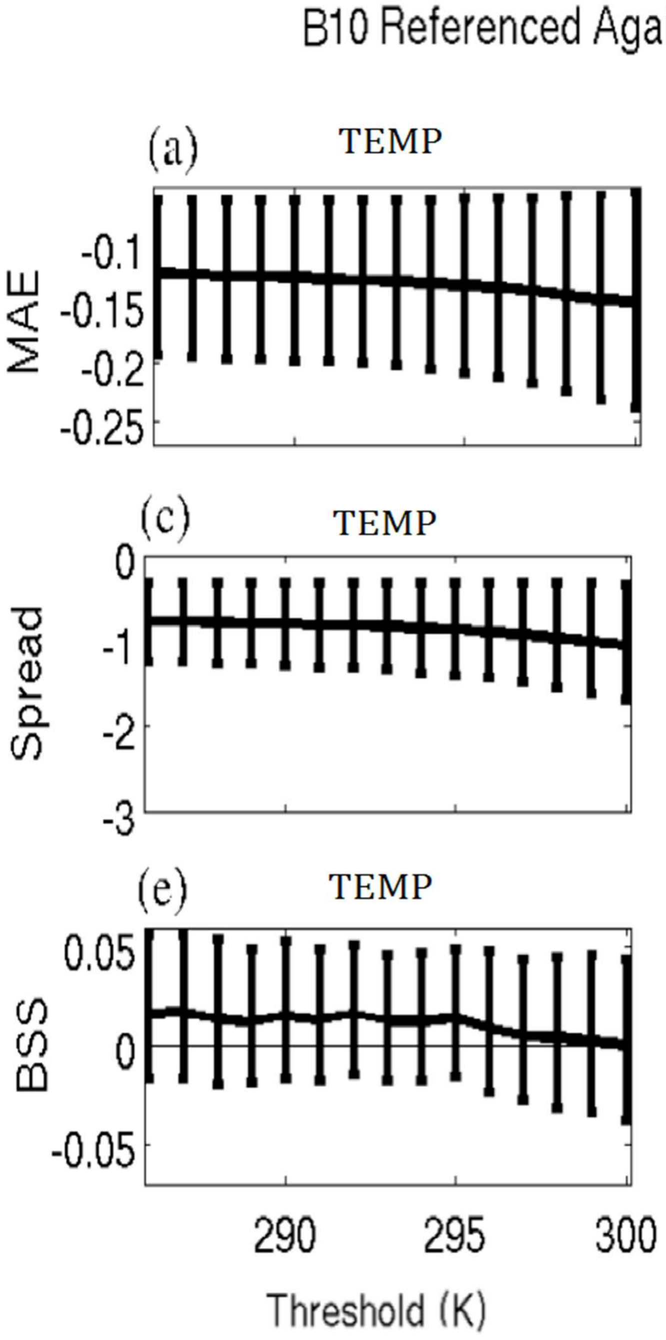

Figure 3.15: Difference between a randomly generated 10-member ensemble (R10) and the best 10 member ensemble (individual members with the lowest MAEs; B10) for (a) MAE, (b) ensemble spread, and (c) BSS (referenced against the R10) for warm season 2-m TEMP. Figures (b), (d), and (f) same as (a), (c), and (e) except for 10-m WNDS.

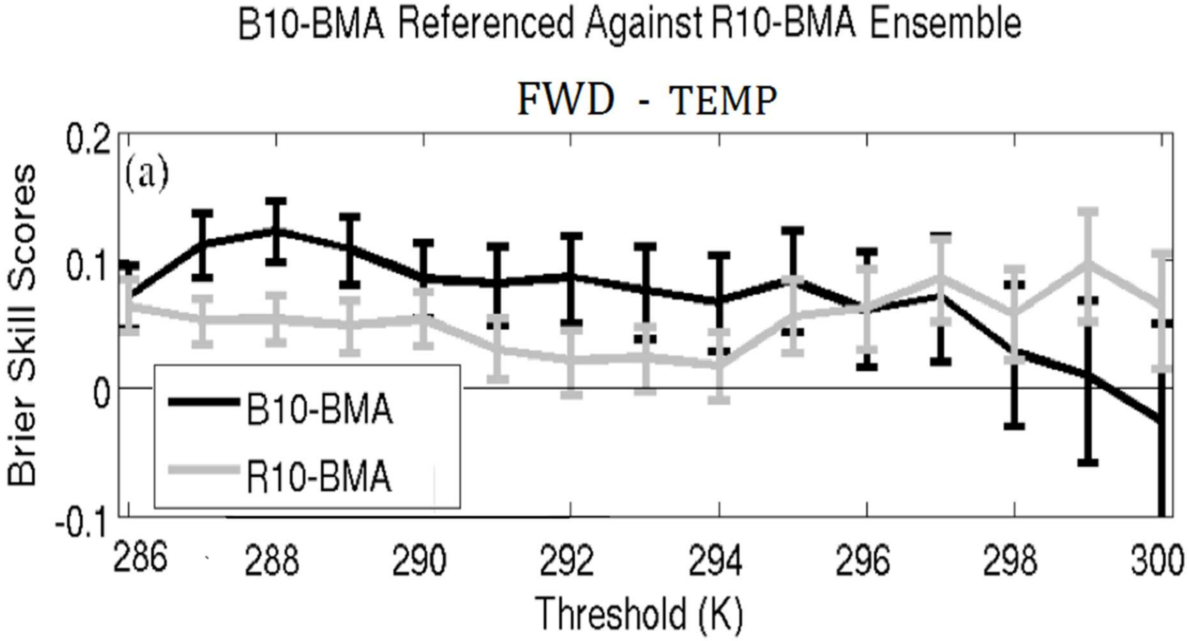

Figure 3.16: Brier skill scores for warm season 2-m TEMP by threshold for FWDs. The 10 best members with BMA (B10-BMA, black) and 10 randomly selected members with BMA (R10-BMA, gray) are referenced against R10 with SBC. The thin black horizontal line denotes zero BSS.

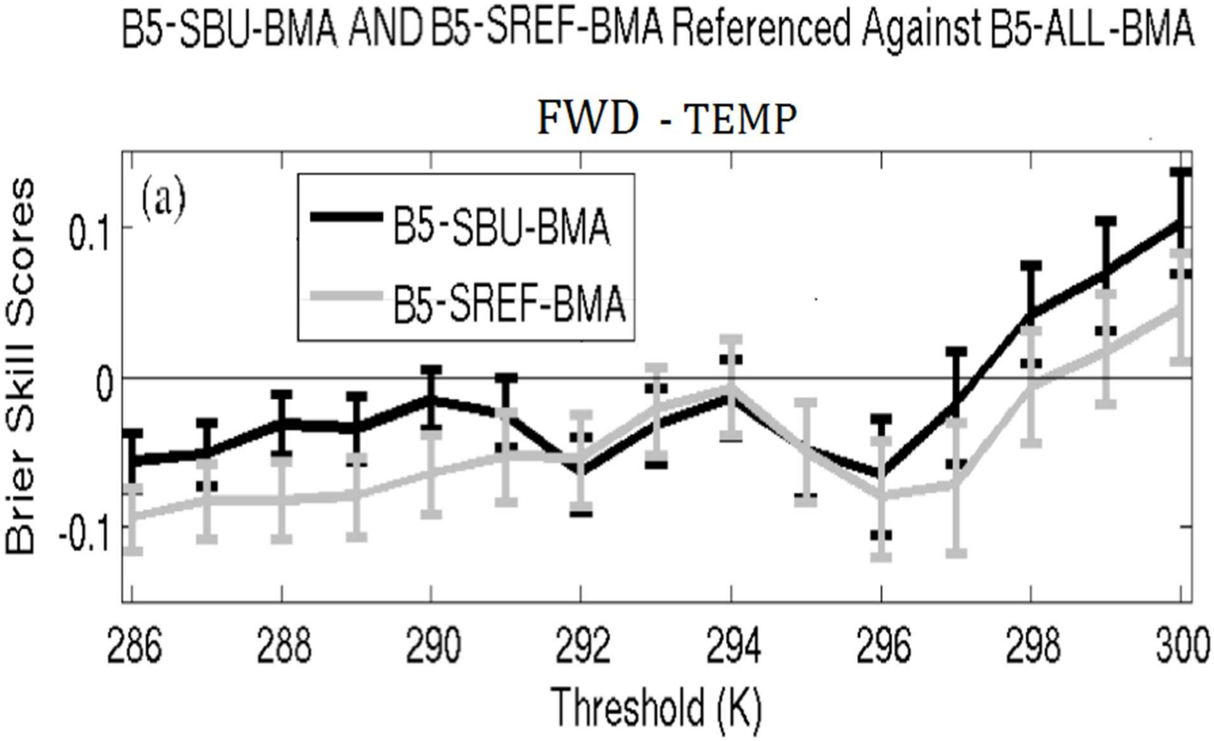

Figure 3.17: Brier skill scores for warm season 2-m TEMP by threshold for FWDs. The 5 best SBU members (B5-SBU-BMA, black) run with BMA and 5 best SREF members (B5-SREF-BMA) run with BMA are referenced against a BMA run 5-member ensemble of randomly selected SBU and SREF members (B5-ALL-BMA). The thin black horizontal line denotes zero BSS.

Chapter 4:

Gridded Verification and Post-processing of the NCEP-SREF Using a Statistical Fire Weather Index (FWD2)

4.1 Introduction

Chapter 3 presents evidence for unique model biases on fire weather days (FWDs) compared to the warm season average, albeit by verifying an older version (2007-2009) of the Short Range Ensemble Forecast (SREF) system. There are three caveats specific to chapter 3's verification; FWD1 (see chapter 2) is based on indices that are not easily relatable to fire occurrence over the Northeast United States (NEUS), the SREF has undergone extensive upgrades since 2009 (Du et al. 2012) and near-surface model output [i.e. 2-m temperature [TEMP], 2-m relative humidity (RELH), and 10-m wind speed (WNDS)] is very sensitive to the planetary boundary layer (PBL) scheme (Pleim 2007). Therefore, gridded 3-D (i.e. horizontal and vertical) verification with post-processing is warranted using a more modern ensemble system for the FWD2 (see chapter 2) methodology.

Mass et al. (2008) note that model bias is not a true source of forecast uncertainty since bias can be removed using a variety of post-processing methods. The challenge with this approach comes in identifying and correcting the many different forms of model bias, particularly those associated with season or synoptic flow regime. The goals of this chapter are to determine if model biases are consistent on FWDs using the FWD2 methodology, and to evaluate how correctable FWD model biases are vertically, spatially and temporally. This approach builds off of chapter 3, which only focuses on the near-surface model variables using the FWD1 method.

In order to perform gridded verification, the fire weather index (FWI) developed in chapter 2 must be adapted to gridded data. This requires each point on the grid to have a unique climatology so that the data can be standardized. Likewise, Bayesian Model Averaging (BMA) in its current form is only designed for site-specific verification. While previous studies have adapted BMA to a grid (Berrocal et al 2007), this chapter focuses only on verification and bias correction rather than ensemble calibration. Section 4.2 details the FWI adaption to gridded format, the analysis used, and the SREF ensemble to be verified. Section 4.3 presents the verification and post-processing results with comparisons to section 3.3 where applicable. Section 4.4 concludes and discusses an operational implementation of the FWI index for the SREF.

4.2 Methods

a. Extending the fire weather index to a grid

The analysis dataset for this chapter consists of the Rapid Update Cycle (RUC; Benjamin et al. 2004) initial condition field prior to 01 May 2012 and the Weather Research and Forecasting (WRF) based Rapid Refresh (RAP; Benjamin et al. 2007, Brundage et al. 2014) thereafter. The RUC and RAP models are run hourly with assimilation to create a high resolution short-term simulation at 13-km grid spacing. The RUC/RAP system has been gradually adapted over time to include a variety of observations (Weygandt et al. 2014) including Tropospheric Airborne Meteorological Data Reporting (TAMDAR), radar reflectivity, profiler data, rawinsondes, surface mesonet stations, offshore buoys, Automated Surface Observing System (ASOS) stations, and a variety of satellite data including precipitable water estimates and cloud top pressure and temperature.

Extending the FWI to gridded format requires a representative TEMP and RELH climatology of the mean and standard deviation (stdev) for each grid point. The climatology is used to create standardized anomalies of the predictors (TEMP and RELH) for the binomial logistic regression model described in chapter 2. This climatology must be consistent (i.e. not biased) with respect to the RUC/RAP analysis data. One approach is to use the RUC/RAP analysis as the climatology directly, although several years of data is generally shorter than the typical 20 to 30 years used in computing climatologies. A second approach is to use a reanalysis dataset that is available for over 30 years. This chapter explores both approaches by comparing the RUC/RAP analysis and the North American Regional Reanalysis (NARR; Mesinger et al. 2006) dataset between the overlapping years of 2007 and 2014. Although the NARR is available for a longer period of time, significant differences in the daily maximum TEMP or daily minimum RELH would result in different climatologies between the two analyses. For instance, if the NARR has a high RELH bias compared to the RUC/RAP, then the biased gridded FWI climatology would result in the over prediction of FWI. Only the mean and stdev biases of daily maximum TEMP and daily minimum RELH are analyzed. The grid used in comparing the NARR and RUC/RAP climatologies extends from roughly 28 N^{o} to 45^{o} N and -84^{o} W to -68^{o} W. Although this is well beyond the boundaries of the NEUS, analyzing the spatial variations in the climatology of these analyses are desired.

After determining the appropriate climatology, the ASOS based FWI (see chapter 2) is compared to another FWI formulation using gridded RUC/RAP analyses. The domain selected for model verification with accompanying ASOS stations extends from 40.5^{o} N to 42^{o} N and -74.5^{o} W to -71.5^{o} W (Fig. 4.1) and is generally consistent with domain 1 (D1) from chapter 2 (Fig 2.2a). This dissertation assumes the binomial logistic regression parameters determined with D1 (see chapter 2) are spatially invariant and can be applied to each point on the model or observational grid. This assumption is based on the small variations in parameter estimates between D1 and D2 and the small variations in parameter estimations for each D1 station (not shown). Although this assumption is not likely to hold over long distances, it likely approximates reality over the Northeast United States (NEUS) region.

The binomial logistic regression predictors are the spatial median of each grid point's (rather than site specific) daily maximum 2-m TEMP and daily minimum 2-m RELH. In other words, the same statistical model as chapter 2 is applied here with one TEMP and one RELH input to calculate the probability of fire occurrence over the domain. The domain representative FWI is automatically set to zero if snow cover is present anywhere within the domain. The presence of snow cover is determined from the Multisensor Snow and Ice Mapping System (IMS) Northern Hemisphere Snow and Ice Analysis (National Ice Center, 2008) to find and exclude these days.

b. Description of the gridded analysis and ensemble data

During verification, the RUC and RAP analyses are separated into two unique periods since the models are run using two distinct model cores. There are some additional physics changes and upgrades within the RUC and RAP periods (Weygandt et al. 2014), but this dissertation assumes the impact of these changes to the analysis field is minor. The two separate analysis periods are:

1) The 13-km RUC analysis for every hour between 01 April 2007 and 01 May 2012.
2) The 13-km RAP analysis for every hour between 21 August 2012 and 30 June 2014.

The National Centers for Environmental Prediction (NCEP) Short Range Ensemble Forecast (SREF) system is verified between 2007 and 2014 (Du et al. 2012). The SREF had three upgrades between 2007 and 2014 as detailed below:

1) SREF2007 – Run between 01 April 2007 and 26 October 2009 with 4 unique cores (3 WRF-ARW, 3 WRF-NMM, 10 ETA and 5 RSM) at 32 to 45 km grid spacing.
2) SREF2009 – Run between 26 October 2009 and 21 August 2012) with 4 unique cores (5 WRF-ARW, 5 WRF-NMM, 6 ETA and 5 RSM) at 32 to 35 km grid spacing.
3) SREF2012 – Run between 21 August 2012 and 30 June 2014 with 3 unique cores (7 WRF-ARW, 7 WRF-NMM, 7 WRF-NMMB) at 16 km grid spacing.

For additional details on the SREF physics and set up, see Table 2.1 and Du et al. (2012). For simplicity, SREF 2007 and SREF 2009 are combined into one verification period since the number of model cores remains consistent while the number of members within each core changes slightly. Although there are some minor physics changes between SREF2007 and SREF2009 within some model cores (particularly within the ETA model), these changes are assumed to have a minor effect on the verification results. For consistency, verification and post-processing results are averaged by each separate model core within the SREF for two different periods.

1) SREF1 consists of the SREF2007 and SREF2009 between 01 April 2007 and 01 May 2012 verified with the RUC analysis.
2) SREF2 consists of the SREF2012 between 21 August 2012 and 30 June 2014 verified with the RAP analysis.

The time period between 02 May 2012 and 20 August 2012 is not considered to keep the newer (older) SREF version verification consistent with the RAP (RUC) analysis. All RUC/RAP analyses are bi-linearly interpolated to the SREF grid before verification or calculating FWI.

c. Ensemble verification and post-processing

The ensembles are verified against the RUC/RAP analysis by analyzing systematic bias and non-systematic error for all FWI values greater than zero (i.e. FWD2 events). In addition, model biases and error of potentially important state variables are assessed for FWD2 events to quantify their 3D structure. As with earlier verification (chapter 3), ensemble biases and error are assessed by looking at contingency bias (CB) and equitable threat score (ETS) by threshold. When appropriate, mean error (ME) and mean absolute error (MAE) are also used.

A simplified bias correction is applied and evaluated for the SREF ensemble using a spatially invariant additive approach (similar to section 3.2b). The additive bias correction generally performs best for normally distributed variables like TEMP (similar to Fig. 3.1 and Fig 3.2) and will likely perform well for daily minimum RELH as well. Two types of post-processing are explored; one with sequential bias correction (SBC; uses the previous 14 days) and another with conditional bias correction (CBC; uses 14 previous FWD2 days to correct future FWD2 days), similar to section 3.2c. When FWI is being verified with CBC or SBC, RELH and TEMP are post-processed before the index is calculated.

Training and verification periods are created by bootstrapping the original dataset with replacement 1000 times to reconstruct a new dataset the same size as the original number of FWD2 events. In the case of CBC, all FWI days are resampled, which rearranges the chronological order of the original data. In the case of SBC, the FWI days are resampled but the chronological order of the previous 14 days (which may or may not be a FWD2 event) are preserved. The resampling with replacement still separates the training and verification windows, although duplicate data may result in the calibration and verification windows. Post-processing is then applied using a sliding window approach described in section 3.2b. All results with the raw, SBC and CBC are presented for data in the independent verification window only. Error bars in all plots represent the 2.5th and 97.5th percentile of the resampled dataset. Furthermore, all references to "statistically significant" indicate confidence exceeding the 95% threshold using the bootstrapped data sets.

4.3 Results

a. Determining the FWI climatology

To determine the appropriate climatology for the gridded FWI, the mean and stdev of daily maximum 2-m TEMP are compared for the NARR minus RUC mean (Fig 4.2a), NARR minus RAP mean Fig 4.2b), NARR minus RUC stdev (Fig 4.2c), and NARR minus RAP stdev (Fig 4.2d). All NARR data is interpolated to the RAP grid. Some differences in maximum TEMP between the NARR and the RUC/RAP over the NEUS may be topographical and related to interpolation errors with the NARR being colder compared to the RUC/RAP by up to 3^{o}C in the Hudson and Connecticut valleys (Fig 4.2 a,b). More importantly, the NARR uses a different forward model (ETA) than the RUC/RAP and only assimilates surface observations over land using 2DVAR (Mesinger et al. 2006). As a result, the NARR analysis at the surface may be heavily impacted by the forward iterations from the ETA model rather than the data assimilation technique. Differences in stdev over the NEUS are minor but exceed 1 stdev (K) in the piedmont of NC and SC (Fig 4.2 c,d). The NARR averages greater than 3^{o}C cooler in daily maximum

TEMP compared to the RUC/RAP over the Gulf Stream, which could be related to differences in sea surface temperature, model core, or a lack of marine surface data assimilation in the NARR.

Figure 4.3 shows the NARR minus RUC/RAP mean and stdev for minimum daily 2-m RELH. There are large positive biases in NARR RELH compared to the RUC/RAP over the considered domain (Fig. 4.3) for both the RUC (~10%) and RAP (~15%). Biases are even greater over the ocean, particularly over the Gulf Stream in the RAP (> 20%). In addition, stdev is notably lower in the NARR compared to the RUC (averaging -2 stdev) and RAP (averaging -2 stdev) over the NEUS domain. This indicates that the NARR has greater RELH with less variability than the RUC/RAP. The high RELH bias over the NEUS is largely the result of the NARR's inability to detect RELH events that are below 40% in the RUC/RAP (not shown). As a result, the NARR is not a suitable climatological dataset for the RUC/RAP. Instead the RUC and RAP analysis between 2007 and 2014 is used to develop climatological values.

Developing a climatology from the RUC/RAP analysis is not intuitive since the RUC and RAP may be different as a result of their unique model core, model physics adjustments, and changes to data assimilation (Weygandt et al. 2014). For instance the RUC uses its own unique core while the RAP uses the WRF-ARW core. In addition, the RUC uses a 3DVAR data assimilation technique while the RAP uses a hybrid 3DVAR-EnKF data assimilation partially borrowed from the Global Ensemble Forecast System (GEFS). Therefore, the RUC and RAP climatologies should first be compared. If the RUC and RAP climatologies are similar, standardization may benefit from the increased sample size of combining both the RUC and RAP initial condition fields.

The RUC/RAP based FWI is verified against the ASOS FWI (chapter 2) between 2007 and 2014 over the NEUS in Fig. 4.4 for FAR (a), HIT (b) and CSI (c). The verification is performed using differing RUC/RAP climatologies; the RUC climatology for the entire period (RUC_ALL; blue), the RAP climatology for the entire period (RAP_ALL; cyan), the combined RUC/RAP for the entire period (RR_ALL; yellow), and the RUC (RAP) climatology used separately for the RUC (RAP) (RR_SEP; magenta). RR_SEP is the most logical approach for developing the FWI climatology, but others are tested due to the short length of the RAP analysis in operations. In general, RAP_ALL has statistically significantly lower HIT and CSI (by -0.22; and -0.17, respectively for FWI >=1) compared to all other climatologies at all FWI thresholds. The poor performance of RAP_ALL makes sense given the short length of the RAP climatology (~2 years) and differences between the RUC and RAP models. RR_SEP is not significantly different than the optimally performing climatology, which is defined as the metric with the lowest FAR, highest HIT and highest CSI for any FWI threshold.

The gridded FWI is compared to the ASOS FWI (chapter 2) for CB (Fig. 4.5a) and ETS (Fig. 4.5b) for all thresholds. There is consistent underprediction of the gridded FWI for all thresholds using RUCALL (CB = 0.84), RAPALL (CB = 0.46), RR_ALL (CB = 0.72), and RR_SEP (CB = 0.71), respectively at FWI >= 1. Consistent with Fig. 4.4, RAP_ALL has significantly lower CB (by -0.29 at FWI >= 1) and ETS (by -0.16 at FWI >= 1) than all other climatologies. RUCALL, RR_ALL and RR_SEP are never significantly different from each other using CB or ETS at any FWI threshold. This suggests that although the gridded FWI might have a slight underprediction compared to ASOS FWI, all climatologies perform equally well, with the exception of RAPALL. Therefore, RR_SEP is selected as the default climatology used to normalize the gridded FWI for this dissertation.

To evaluate spatial consistency, gridded FWI is interpolated to each ASOS station and compared to the ASOS FWI for CB and ETS (Fig. 4.6). With these spatial comparisons, the final step of computing the spatial median for FWI values is skipped. CB values at FWI >=1 are close to one with the exception of KOXC (CB = 2.16; see Fig 4.1 for location) and KNYC (CB = 2.25), which might be related to the quality of the station observation rather than the RUC grid. For instance, KOXC is the only Automated Weather Observing System (AWOS) station, although the distribution of daily TEMP and RELH is not significantly different for KOXC compared to any other station (not shown). Likewise KNYC does not exhibit different behavior compared to other ASOS locations with TEMP or RELH, although this station has more data missing (75%) compared to the average station (averaging 19.4%). Ignoring KOXC and KNYC, ETS values for coastal locations (KFOK, KHWV, KISP, KFRG, KJFK, KLGA, KHPN, KBDR, KHVN, and KGON) are significantly lower than slightly inland locations (KBDL, KHFD, KIJD, KMMK, KDXR, KEWR, KTEB, and KCDW) at FWI >= 1 (difference averaging 0.14). The lower ETS values at the coastal boundary are likely related to interpolation errors associated with the lower resolution RUC/RAP grid compared to ASOS stations. Considering regions with more uniform land use or using higher resolution gridded data would likely improve the consistency between the gridded FWI and ASOS FWI and reduce representativeness errors.

The monthly climatology of gridded FWI and ASOS FWI are compared in Fig. 4.7 and stacked by index value. The FWI climatologies are qualitatively similar with a primary peak in April and a smaller secondary peak in July. These climatologies are also similar to the 1999 to 2008 fire occurrence monthly frequency observations (Fig. 2.3), but with less FWD2 events observed in Fig. 4.7 particularly during the late autumn and winter months. Caution should be used when comparing the monthly fire occurrence observations of Fig. 2.3 to the monthly frequency of FWI in Fig. 4.7 since the former is between 1999 and 2008 while the latter is between 2007 and 2014. Since the gridded and ASOS FWI are comparable, the gridded technique is used identify FWDs in the SREF and RUC/RAP analysis.

b. SREF gridded verification and post-processing by FWI

SREF CB is computed for all FWI thresholds and presented by model core for the SREF1 and SREF2 (Fig. 4.8). FWI values greater than one are severely underpredicted for the SREF1 (ensemble average of 0.26) and SREF2 (ensemble average of 0.05). This underprediction results in the SREF rarely predicting FWI values of three, with the exception of the SREF1 WRF-NMM core (CB = 0.48). Ensemble average CB at FWI >= 1 is significantly improved with SREF1 SBC (improvement averaging 0.27) and SREF2 SBC (averaging 0.49). There is an additional significant improvement at FWI >= 1 with CBC over SBC for SREF1 (by 0.25) and SREF2 CBC (by 0.23). However, there are instances where the CBC overcorrects the bias in FWI, particularly at the FWI >= 3 threshold with the SREF1 (averaging 1.65) and SREF2 (averaging 2.25). Nonetheless, the optimal performance of the CBC in computing FWI suggests that near-surface atmospheric biases significantly differ for FWD2 events compared to the annual average. These results are consistent with findings from chapter 3, since greater cool (Fig. 3.5) and moist (Fig. 3.8) biases found for FWD1 would likely result in the underestimation of FWI. The impact of bias correction on specific model state variable biases for FWD2 is examined in section 4.3c.

As with CB, ETS is computed by FWI threshold for SREF1 and SREF2 in Figure 4.9. ETS values at FWI >= 1 are very low or zero for SREF1 (averaging 0.001) and SREF2 (averaging 0.000). The very low ETS values at FWI >= 1 are likely due to a lack of false alarms (i.e. model forecasted an event but it did not occur) with the contingency table approach (Wilks

2011), since all Figure 4.9 data is subset on FWD2 events. Hence, there are only hits (both the model and forecast are correct) and misses (model forecasted the event but it did not occur), resulting in a zero ETS when the observed hits are equal to the expected hits by chance (Wilks 2011). Hits, misses and false alarms all occur at higher thresholds, resulting in more representative skill values. On average, the SBC significantly improves ETS for both the RUC (by 0.12) and RAP (by 0.13) at FWI >= 2. Additional improvements with CBC over SBC are generally mixed or negligible.

To avoid the problem with ETS, MAE by threshold is also calculated (Fig. 4.10) for all thresholds and presented by SREF core. MAE is significantly reduced with the SBC versus raw data for the SREF1 (averaging 0.36) and SREF2 (averaging 0.49) at a FWI >= 1. However, CBC improvement is mixed when compared to SBC, with a statistically significant improvement only found for SREF1 WRF-ARW (averaging 0.20) and SREF1 WRF-NMM (averaging 0.20). This suggests that although post-processing is generally effective with removing bias, there is still considerable day to day variability that is difficult to correct with CBC. Since FWDs over the NEUS are associated with multiple synoptic flow patterns (Pollina et al. 2013), it is possible that a bias correction conditional on fire weather pattern could further improve the CBC.

The 2012 SREF upgrade (i.e. SREF2) invoked a newer WRF version (from v2.2 to v.3.3), increased the model resolution (from ~32 km to 16 km), added the NMMB core, removed the RSM and ETA cores, and increased the physics and initial condition diversity (Du et al. 2012). To check for potential improvements between SREF1 and SREF2, differences in MAE are discussed between Fig. 4.10a and Fig. 4.10b. The raw MAE in SREF1 is significantly better than SREF2 at FWI >= 1 for the WRF-NMM (0.23) only. After CBC is applied, the MAE is slightly worse for the WRF-ARW (0.14) and WRF NMM (0.06) in SREF2 compared to SREF1, although these differences are not significant. The less effective CBC for SREF2 might be related to the limited training period available (total of 85 FWD2 events), compared to SREF1 (total of 235 FWD2 events). Nonetheless, these results suggest that bias correction for FWDs is important regardless of the SREF upgrade. In addition, the TEMP MAE within all cores of the SREF2 are very similar to each other with or without post-processing. This is not true for the SREF1, where the RSM core is significantly worse than the other model cores at all FWI thresholds.

CBC spatial statistics at FWI >= 1 are presented for the SREF1 (Fig. 4.11) and SREF2 (Fig. 4.12). Although CBC removes the average model bias, spatial CB values range from 0.41 to 2.20 in SREF1 (Fig. 4.11 a,c,e,g) and from 0.13 to 1.57 in SREF2 (Fig. 4.12 c,f,i). Anomalously high spatial CB (exceeding 1.5; Fig. 4.11 a,c,e,g) occurs over Long Island sound and in the New York Bight region for SREF1, which degrades the ETS for these locations (Fig. 4.11 b,d,f,h). These regions of high CB are likely caused by representativeness errors between the higher resolution RUC/RAP analysis and coarser SREF model grid. The instantaneous RELH and TEMP gradients between the ocean and atmosphere during FWD2 events are significant and can vary by as much as 70% and 20 K, respectively (not shown). As a result, the increased resolution of the SREF2 likely improved the CB and ETS (Fig 4.12) over Long Island Sound and the New York Bight.

Despite some interpolation errors, the lingering spatial bias after CBC (Fig. 4.11 a,c,e,g and Fig. 4.12 a,c,e) appears to exhibit some land use dependence. Neglecting the Long Island Sound and New York Bight regions, SREF1 WRF-ARW (Fig. 4.11a), ETA (Fig. 4.11c), and WRF-NMM (Fig. 4.11e) exhibit a slight negative bias over the ocean (~ 0.72) and little bias inland (~ 0.94), which is reversed in the RSM (1.29 and 0.84, respectively; Fig. 4.11g).

Similarly, the SREF2 WRF-ARW (Fig 4.12c) has a negative bias over the ocean (CB averaging 0.51) and little bias over land (CB averaging 0.97). There are also some slight differences along the coastal boundary for SREF2 WRF-NMM (Fig. 4.12 c) and WRF-NMMB (Fig. 4.12 e), although this might be related to representativeness errors mentioned in the last paragraph. A spatially varying additive bias correction that considers land use characteristics may further improve the results for SBC and CBC.

c. SREF gridded verification and post-processing by model state variable

From section 4.3b, the use of raw SREF data results in a large FWI underprediction. Therefore, it is important to understand how model biases in TEMP and RELH affect the FWI and explore model performance for additional state variables of interest to the fire weather community (such as WNDS). TEMP mean error by height for the raw, SBC and CBC is shown for the SREF1 core (Fig. 4.13a) and SREF2 core (Fig. 4.13b). TEMP exhibits a significant cool bias maximized at 1000 hPa for SREF1 (averaging 2.41 K) and SREF2 (averaging 1.32 K) that slowly decays to a ME of zero above 800 hPa. SBC results in significantly improved ME compared to the raw TEMP for SREF1 (SREF2) below 825 hPa (875 hPa). Likewise, CBC significantly improves on TEMP ME over SBC for all cores in the SREF1 (SREF2) below 875 hPa (850 hPa). The SREF2 has significantly improved raw ME below 700 hPa compared to the SREF1 for the WRF-ARW and WRF-NMM, although there are no appreciable differences after CBC is applied. The cool TEMP bias and effectiveness of CBC on bias removal are consistent with the findings from chapter 3 (Fig. 3.5) but Fig. 4.13 suggests that the cool surface biases are propagated upward through most of the PBL. The ETA and RSM cores of SREF1 have greater raw negative TEMP bias than the WRF cores (by 0.5 K), although CBC effectively removes this bias.

The vertical structure of SREF MAE is also analyzed by model core for the SREF1 and SREF2 (Fig. 4.14). SBC results in a significant improvement when comparing SBC to the raw data below 850 hPa for all SREF1 cores (at 1000 hPa averaging 1.10 K), but this improvement is only statistically significant for the SREF2 core below 975 hPa. Likewise, there are no statistically significant differences between SBC and CBC for either SREF1 or SREF2, which is similar to the FWI MAE results of Fig. 4.10 and earlier results from Fig 3.5. SREF2 MAE is generally better than SREF1 MAE, with statistically significant results below 900 hPa for the WRF-ARW and WRF-NMM cores. Both the raw ensemble ME and MAE are improved in the SREF2 with the removal of the high error RSM and ETA cores (not shown).

As with TEMP, RELH ME (Fig. 4.15) and MAE (Fig. 4.16) by height and model core are presented for the SREF1 and SREF2. A statistically significant positive bias exists for the SREF1 (SREF2) raw data below 800 hPa (850 hPa) that is maximized at 975 hPa (1000 hPa) for all model cores. As with TEMP, the raw ETA and RSM cores exhibit the highest RELH biases in the PBL (peaking at 16% and 18.5% respectively). SREF1 (SREF2) SBC significantly improves ME below 850 hPa (900 hPa) with additional significant improvements for CBC below 875 hPa (925 hPa). As with Fig. 4.14, MAE improvements with RELH are most beneficial with SBC in the lower troposphere (below 875 hPa with SREF1 and 900 hPa with SREF2), with no significant additional improvements with CBC. The SREF upgrade in 2012 reduces the RELH MAE, with the greatest impact between 825 hPa and 975 hPa for the WRF-ARW and WRF-NMM. Interestingly, the MAE is very similar between SREF1 and SREF2 WRF-ARW and

WRF-NMM cores. These results are consistent with the surface positive SPHU biases found in Fig. 3.8. Therefore, the consistent underprediction of FWI (Fig. 4.8) in the raw SREF model data makes sense given the cool TEMP (Fig. 4.13) and positive RELH (Fig. 4.15) biases found in the PBL. Bias correction can effectively remove these PBL biases and marginally improve model error metrics like MAE. Chapter 6 details how post-processing can be adapted to further improve metrics like ETS and MAE.

There is a potential inconsistency between the lower PBL improvement in raw SREF2 TEMP (Fig. 4.13) and RELH (Fig. 4.15) and the slightly worse raw FWI forecasts for SREF2 (Fig. 4.10b) versus SREF1 (Fig. 4.10a). Since near-surface variables are used to compute the FWI, biases in 2-m TEMP and 2-m RELH for the SREF1 and SREF2 are examined in Table 4.1. In general, near-surface TEMP and RELH are very similar to the verification performed at 1000 hPa. There are statistically significant improvements in ME for CBC over SBC and the raw model data for 2-m TEMP and 2-m RELH. In addition, SREF2 2-m TEMP has a statistically significant improvement in the WRF-ARW (0.7 K) with no change in the WRF-NMM. However, there is a statistically significant increase in SREF2 2-m RELH bias for the WRF-ARW (by 6.6%) and WRF-NMM (by 7.0%) when compared to the SREF1. This large positive surface RELH bias in SREF2 negatively impacts the raw FWI and emphasizes the importance of careful post-processing, even when ensembles experience critical upgrades.

Another potentially important model state variable to the rapid spread of wildfires is WNDS (Charney and Keyser 2010), even though it is not used in calculating FWI. Since other indices may consider using ensemble WNDS forecasts in the future, the impact of SBC and CBC are analyzed. SREF1 biases in the raw modeled WNDS (Fig. 4.17a) are mixed (at 1000 hPa averaging 0.13 m s^{-1}) and SBC degrades all model cores to a negative wind speed bias (at 1000 hPa averaging -0.55 m s^{-1}). Therefore, the raw WNDS bias can vary depending on the model core and likely parameterized model physics within each core. Regardless of model core, FWD2 events exhibit less positive or more negative WNDS bias than the climatological average, which is consistent with Fig. 3.9. CBC improves low level ME over SBC and the raw output for all cores except the WRF-NMM. For SREF2, there is a significant positive raw wind speed model bias (at 1000 hPa ensemble average = 1.06 m s^{-1}), that is overcorrected with SBC (at 1000 hPa ensemble average = -0.41 m s^{-1}), and significantly improved with CBC (at 1000 hPa ensemble average = 0.05 m s^{-1}). SREF1 MAE for WNDS does not significantly differ among model core or bias correction type (Fig. 4.18). However, SREF2 CBC results in a statistically significant improvement for WNDS over the SREF1 within the WRF-ARW core (0.44 m s^{-1}). While SBC generally significantly improves bias for TEMP and RELH, this is not true for WNDS. Therefore, raw model output is preferred over SBC to avoid underestimating fire threat in forecast models.

An additional consideration with model verification and post-processing of FWD2 events are the variability of model biases and error by time of year. For instance, biases may change with season due to synoptic flow regime or variations in land use. Since FWD2 events peak in the early to mid-spring over the NEUS, land use changes associated with the spring bloom may affect surface fluxes with the atmosphere, contributing to changes in model bias. As a result, monthly TEMP bias (Fig. 4.19) are averaged between 1000 hPa and 800 hPa by core for SREF1 raw (a), SREF1 CBC (b), SREF2 raw (c), and SREF2 CBC (d). The climatological average monthly TEMP bias (i.e. FWD2 and non-FWD2 events) is shown in the bold black line of Figure 4.19 a,c and Figure 4.20 a,c. SREF1 ensemble mean TEMP biases on FWD2 events are more negative in the winter and peak in March (ensemble average = -2.73 K), before diminishing in the summer and reaching a minimum bias in August (ensemble average = -0.97). Monthly average raw TEMP biases follow the same behavior as on FWD2 events, but exhibit a smaller negative bias that is significantly different during the peak fire occurrence months of March to May (difference averaging -1.01 K). There are also significantly greater ensemble mean negative model biases on FWDs for the months of June, July, September and October. The CBC ensemble ME for SREF1 (Fig. 4.19b) is significantly improved for all months, but there is still a seasonally varying model bias that is negative from March to May (averaging -0.60 K) and positive from June to October (averaging 0.65 K). The seasonally varying ME after CBC likely indicates that there is additional room for improvement with post-processing model data conditional on season.

SREF2 TEMP biases (Fig. 4.19c and Fig 4.19d) on FWDs are still negative with a March minimum (averaging -1.60 K) and October maximum (averaging -0.10 K) but are significantly reduced compared to SREF1 for April and May and from September to November (improvement averaging 1.11 K). Even with the reduction of SREF2 biases, FWD2 events still exhibit a significantly greater raw negative TEMP bias compared to the monthly average from February to June (Fig. 4.19c; difference averaging -0.63 K). CBC results in a significant improvement of SREF2 ensemble mean bias for February, April, May, June, September, and October (difference averaging 0.98 K). Similar to CBC with SREF1, the SREF2 also exhibits a lingering negative bias from March to June that is not significant (-0.29 K) and a lingering positive bias from August to November that is significant (+1.09 K).

Finally, MAE is presented by month for the SREF1 raw, SREF1 CBC, SREF2 raw, and SREF2 CBC (Fig. 4.20). Raw MAE on FWDs is slightly elevated compared to the climatological average, with statistically significant results for February, April, and August through November for SREF2. As with ME, CBC reduces MAE for both SREF1 and SREF2, particularly during the peak fire occurrence season of March through May. From Fig. 4.19 and Fig. 4.20 it appears that FWDs exhibit different model biases than the climatological average any time of year, but the greatest differences in terms of bias generally occurs between the late winter and early summer. Furthermore, the newer SREF has significantly reduced TEMP biases between 1000 hPa and 800 hPa compared to the older SREF, but a large cool bias on FWD2 events is a lingering problem.

4.4 Discussion

This chapter compares the Short Range Ensemble Forecast (SREF) to the analysis products from the Rapid Update Cycle (RUC) prior to 2012 and the Rapid Refresh (RAP) thereafter. Gridded verification of ensemble model data allows for a comprehensive three-dimensional perspective with regard to model bias, error, and the effectiveness of post-processing. Several metrics are calculated for model bias [defined as CB (contingency bias) or ME (mean error)] and error [defined as MAE (mean error) or ETS (equitable threat score)] on FWD2 days (see chapter 2) during both the Short Range Ensemble Forecasts (SREF) prior to 2012 (SREF1) and after 2012 (SREF2).. Sequential bias correction (SBC) significantly (defined as exceeding 95% confidence using statistical bootstrapping) improves forecasts of the fire weather index (FWI), temperature (TEMP) and relative humidity (RELH) below 800 hPa. However, SBC overcorrects modeled wind speed (WNDS), resulting in a greater negative bias than the raw WNDS. Conditional bias correction (CBC) is more effective at removing bias than SBC for the FWI, in addition to low level TEMP, RELH, and WNDS. Ensemble improvements with conditional post-processing over sequential are generally minor but edge toward slightly beneficial. Overall, bias correction is better than using raw model data when analyzing fire weather events. One important exception is with WNDS, where SBC degrades the error. This is encouraging, since it is simple in practice to implement SBC when predicting FWDs, although CBC is recommended.

The FWI has been applied operationally to the raw SREF ensemble, which updates four times daily at: http://wavy.somas.stonybrook.edu/fire/. As emphasized in this chapter, ensemble post-processing is preferred when calculating fire weather forecasts from operational ensembles and additional modifications to the website will include some form of post-processing. The results from chapter 4 present a general idea of what a conditional bias correction could achieve operationally. Several additions could be made to improve the effectiveness of this bias correction. The most obvious adjustment is to bias correct each model grid point individually. Non-Gaussian variables such as WNDS could be corrected with the cumulative distribution function (CDF) bias correction method from chapter 3. The bias correction could be adapted to consider additional conditional model biases such as season, month or synoptic flow pattern. However, the optimal conditional bias correction is not intuitive and this chapter serves as a good starting point by quantifying some conditional model biases. Nonetheless, care must be taken to insure statistical significance with diminishing sample size. More advanced post-processing methods such as BMA (chapter 3) could be applied to improve model bias in both the 1st and 2nd moments. BMA would have to be extended to model gridded data before this can be applied.

In practice, a FWD2 event would have to be known a priori to implement CBC over SBC. Although this is possible using a synoptic regime capture method or a simple RELH and TEMP threshold method, the results presented in section 4.3 represent the optimal regime capture method, since FWDs are known beforehand. The importance of implementing an operational CBC is discussed in section 3.4, with ideas for future work elucidated in section 6.3.

Finally, it is quite possible that some of the cool TEMP and high RELH biases on FWDs is the result of a lack of evapotranspiration in the observed atmosphere before the spring bloom. For instance, if evapotranspiration is too abundant in the model, this could result in too much evaporational cooling above the ground and an overly shallow, cool and moist planetary boundary layer (PBL). This could explain the increase in the climatological cool TEMP bias (solid black line) in Fig. 4.19a and Fig. 4.19c between March and May. However, given the persistence of a conditional cool TEMP bias exceeding climatology outside of the spring green-up period, evapotranspiration is likely not the only influence contributing to this bias. Additional research is warranted regarding the influence of seasonal land-surface interactions on model bias.

Table 4.1: ME [2.5th percentile; 97.5th percentile] for 2-m TEMP and 2-m RELH using the raw model data, SBC, and CBC on FWDs.

	SREF1	**SREF1**	**SREF1**	**SREF1**	**SREF2**	**SREF2**	**SREF2**
	ARW	**ETA**	**NMM**	**RSM**	**ARW**	**NMM**	**NMMB**
TEMP Raw (K)	[-2.4; -2.2]	[-2.0; -1.8]	[-1.8; -1.5]	[-2.2; -2.0]	[-1.8; -1.4]	[-1.8; -1.4]	[-2.1; -1.7]
TEMP SBC (K)	[-1.1; -0.9]	[-0.9; -0.7]	[-0.7; -0.4]	[-1.0; -0.8]	[-0.8; -0.4]	[-0.8; 0.5]	[-0.9; -0.5]
TEMP CBC (K)	[-0.0; 0.0]	[-0.0; 0.0]	[-0.0; 0.0]	[-0.0; 0.0]	[-0.1; 0.1]	[-0.1; 0.1]	[-0.1; 0.1]
RELH Raw (%)	[8.1; 9.3]	[10.7; 11.8]	[6.3; 8.2]	[12.3;14.3]	[14.1;16.5]	[13.1;15.5]	[3.4; 7.4]
RELH SBC (%)	[4.3; 5.3]	[4.3; 5.3]	[2.6; 4.0]	[4.2; 5.6]	[3.6; 5.5]	[3.1; 5.1]	[6.8; 9.5]
RELH CBC (%)	[-0.1; 0.1]	[-0.3; 0.1]	[-0.2; 0.2]	[-0.4; 0.1]	[-0.5; 0.5]	[-0.5; 0.5]	[-0.8; 0.8]

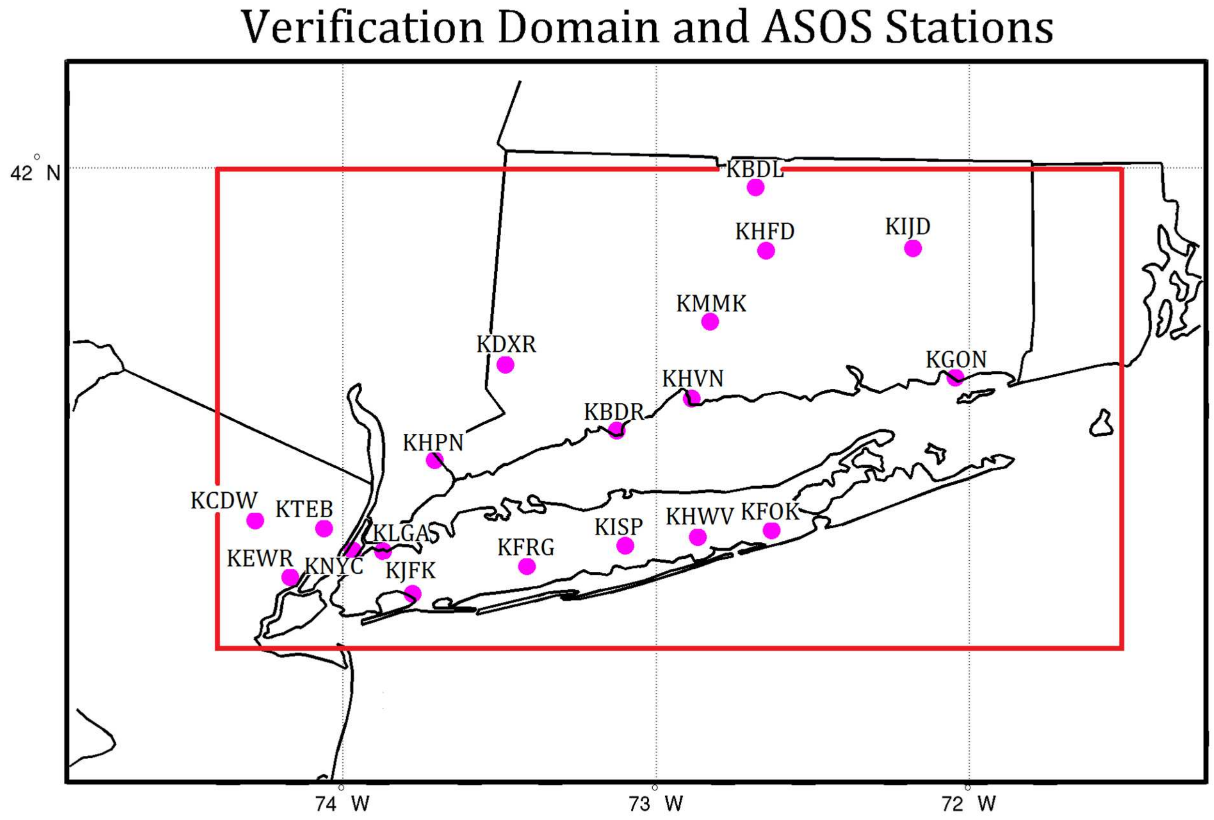

Figure 4.1: Domain used (red box) and ASOS stations used for developing grid-based FWI.

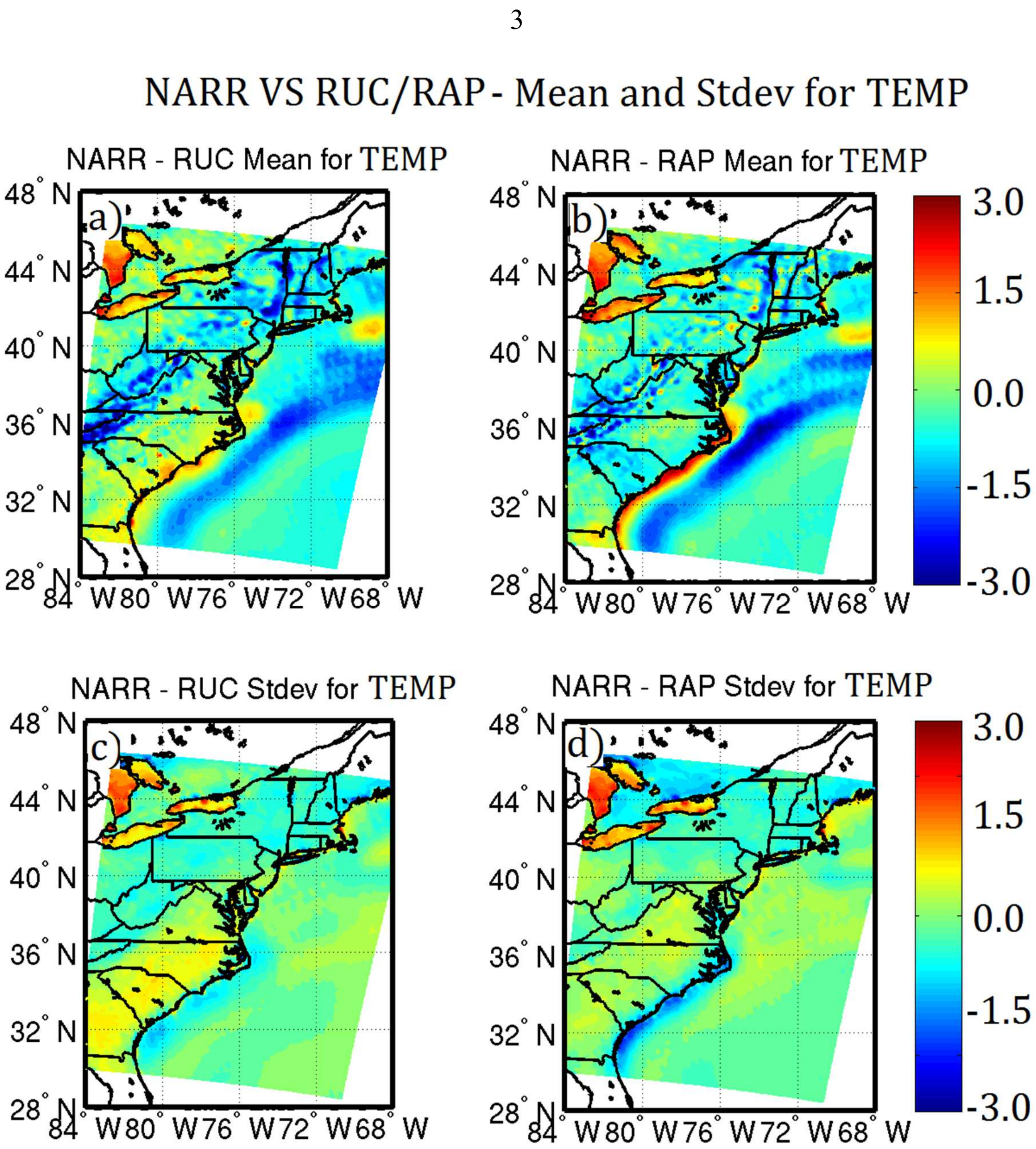

Figure 4.2: NARR reanalysis minus RUC analysis (a,c) and RAP analysis (b,d) for the climatological mean TEMP (a,b) and standard deviation (Stdev) TEMP (c,d).

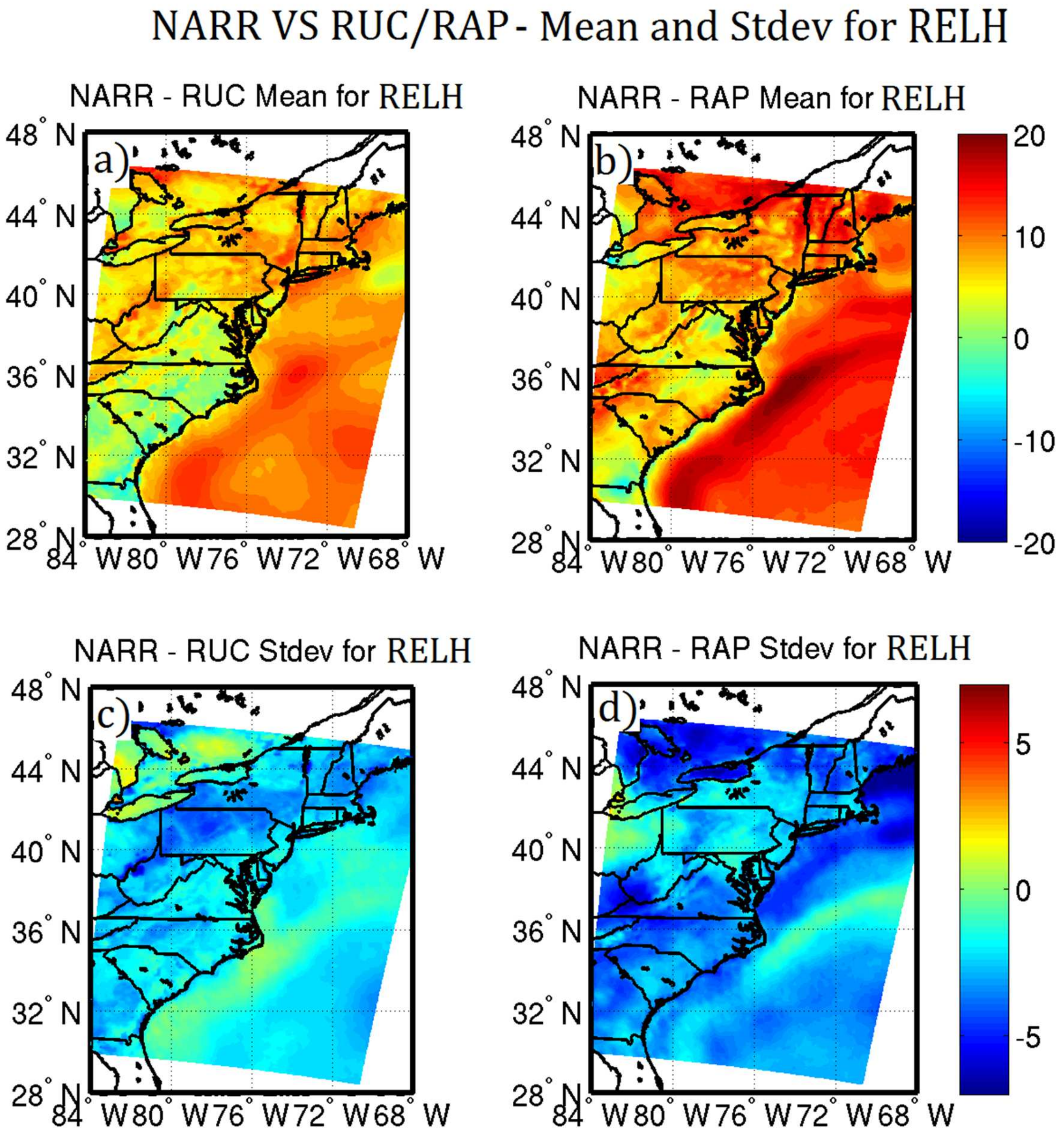

Figure 4.3: Same as figure 4.2, but for RELH.

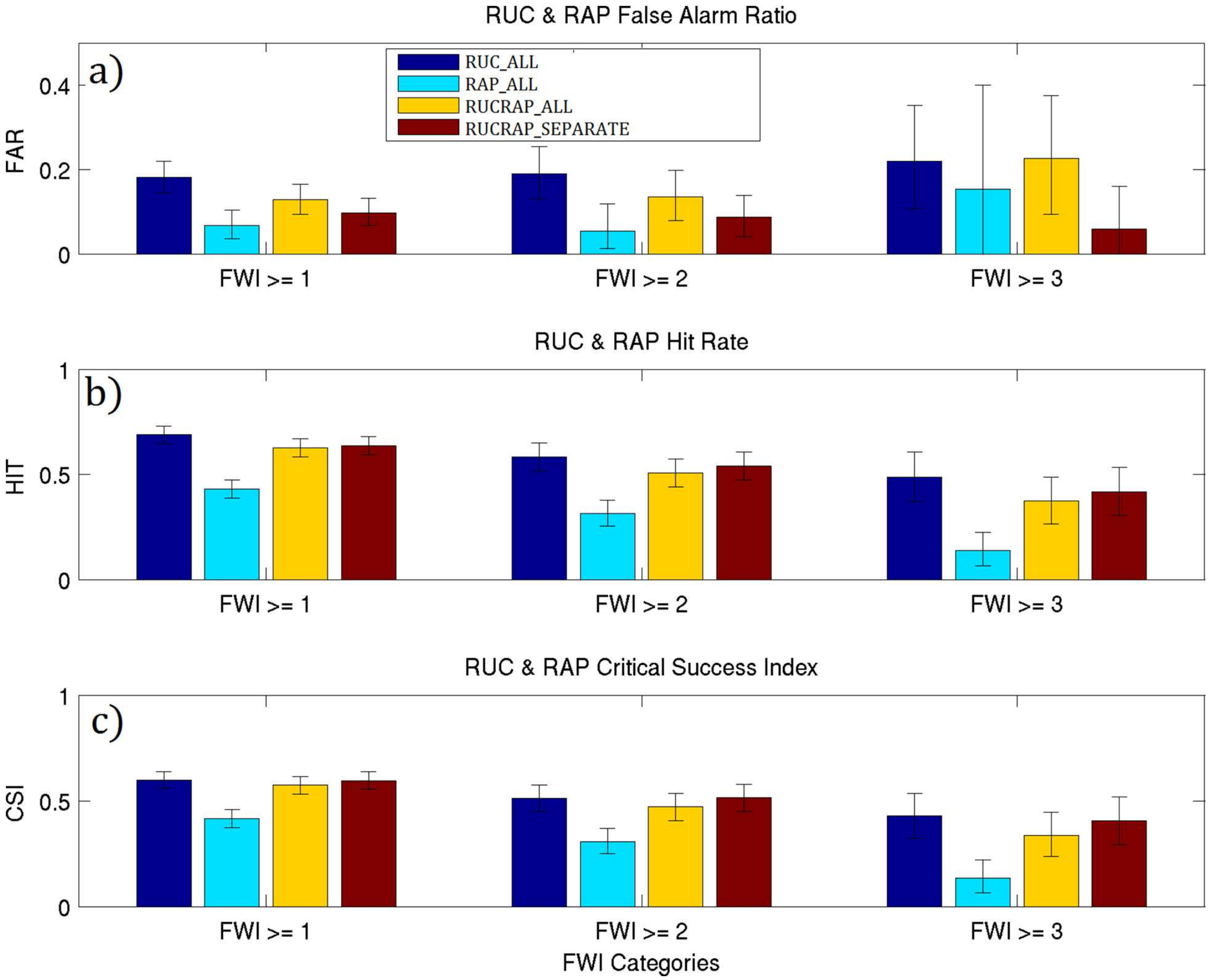

Figure 4.4: FAR (a) , HIT (b), and CSI (c) for RUC/RAP derived FWI compared to METAR derived FWI using RUC climatology (RUC_ALL; blue), RAP climatology (RAP_ALL; cyan), RUC/RAP averaged climatology (RR_ALL; yellow) and RUC climatology for the RUC analysis and RAP climatology for the RAP analysis (RR_SEPARATE; red). Error bars represent the 2.5th and 97.5th percentile of the 1000 resampled datasets.

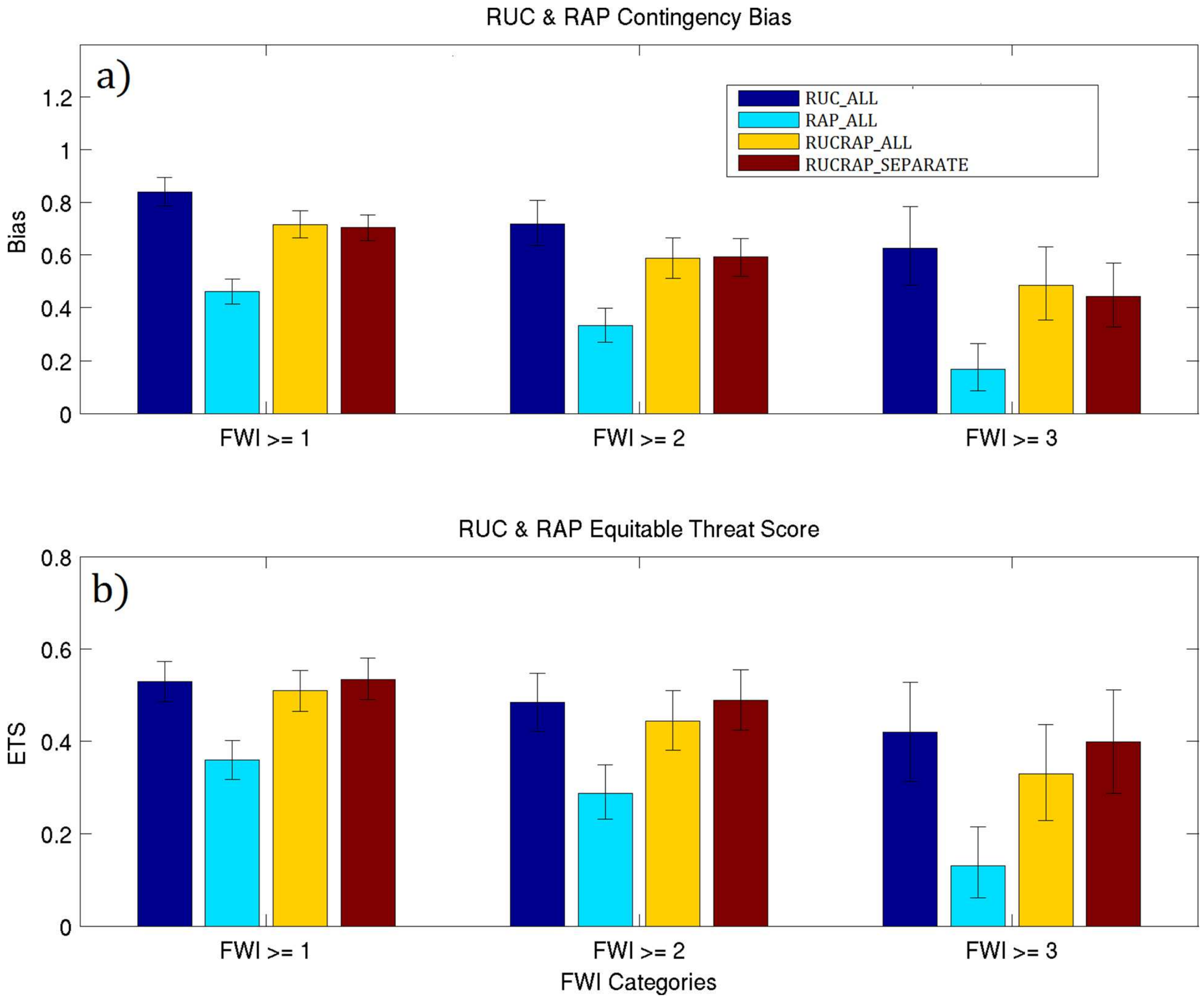

Figure 4.5: Same as figure 4.4, but for CB (a) and ETS (b).

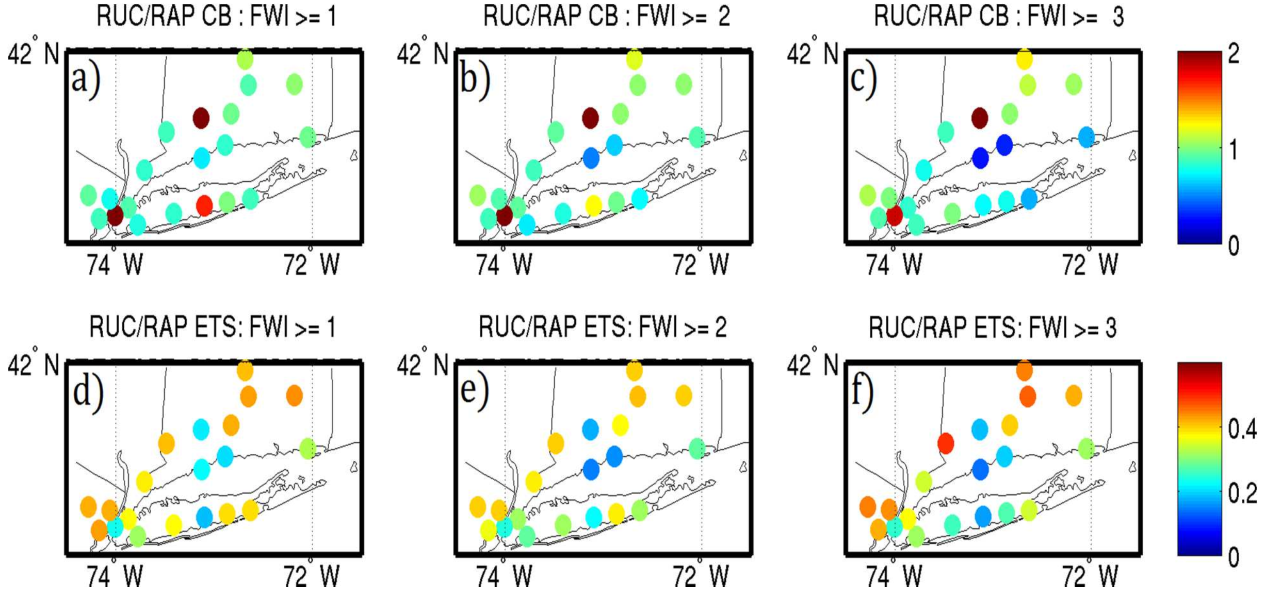

Figure 4.6: Spatial CB (a-c) and ETS (d-f) for RUC/RAP based FWI compared to METAR based FWI by threshold.

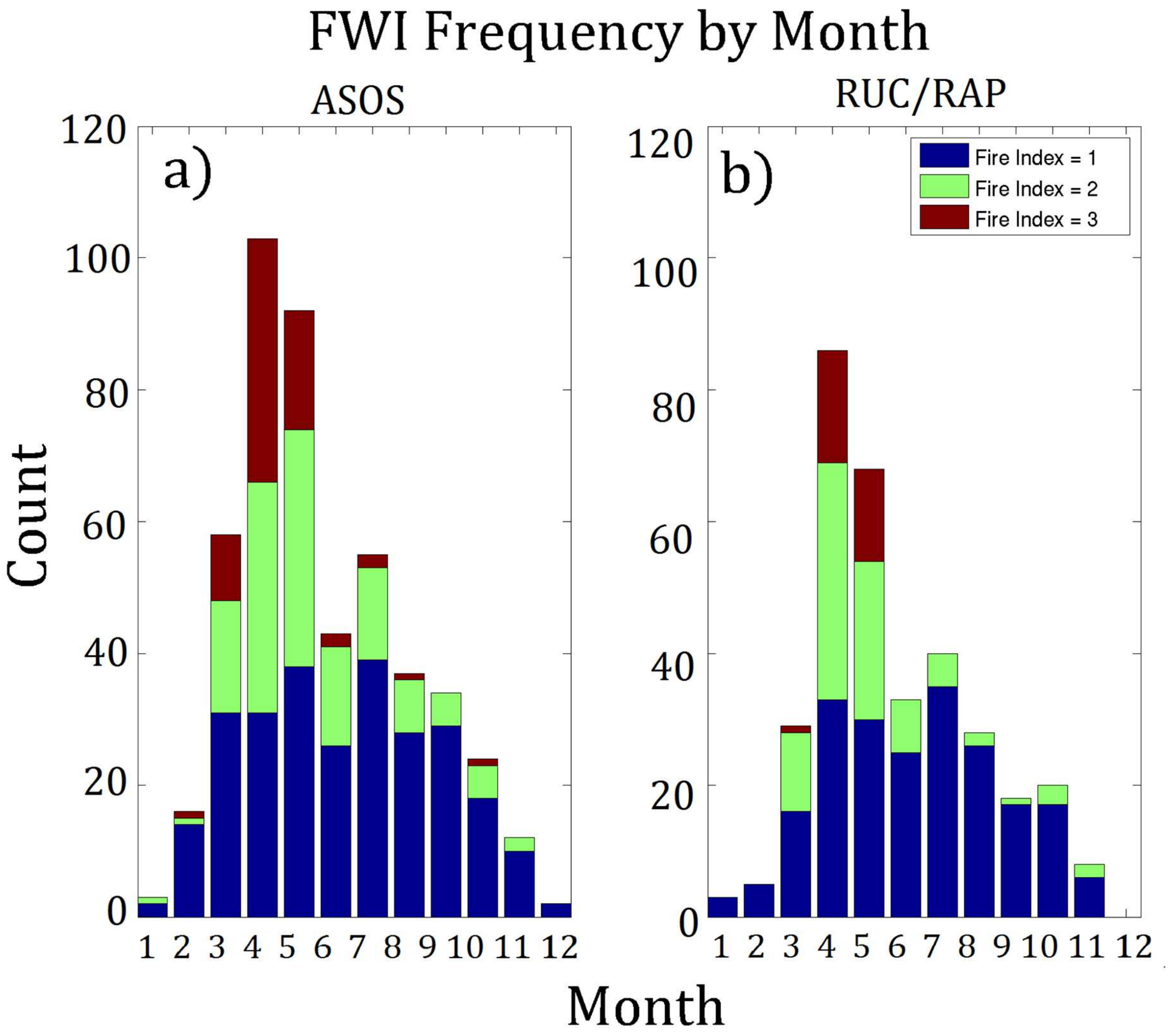

Figure 4.7: FWI Monthly climatology by index value (blue, green and red for FWI values of 1,2 and 3, respectively) for ASOS (a) and RUC/RAP (b) between 2007 and 2014.

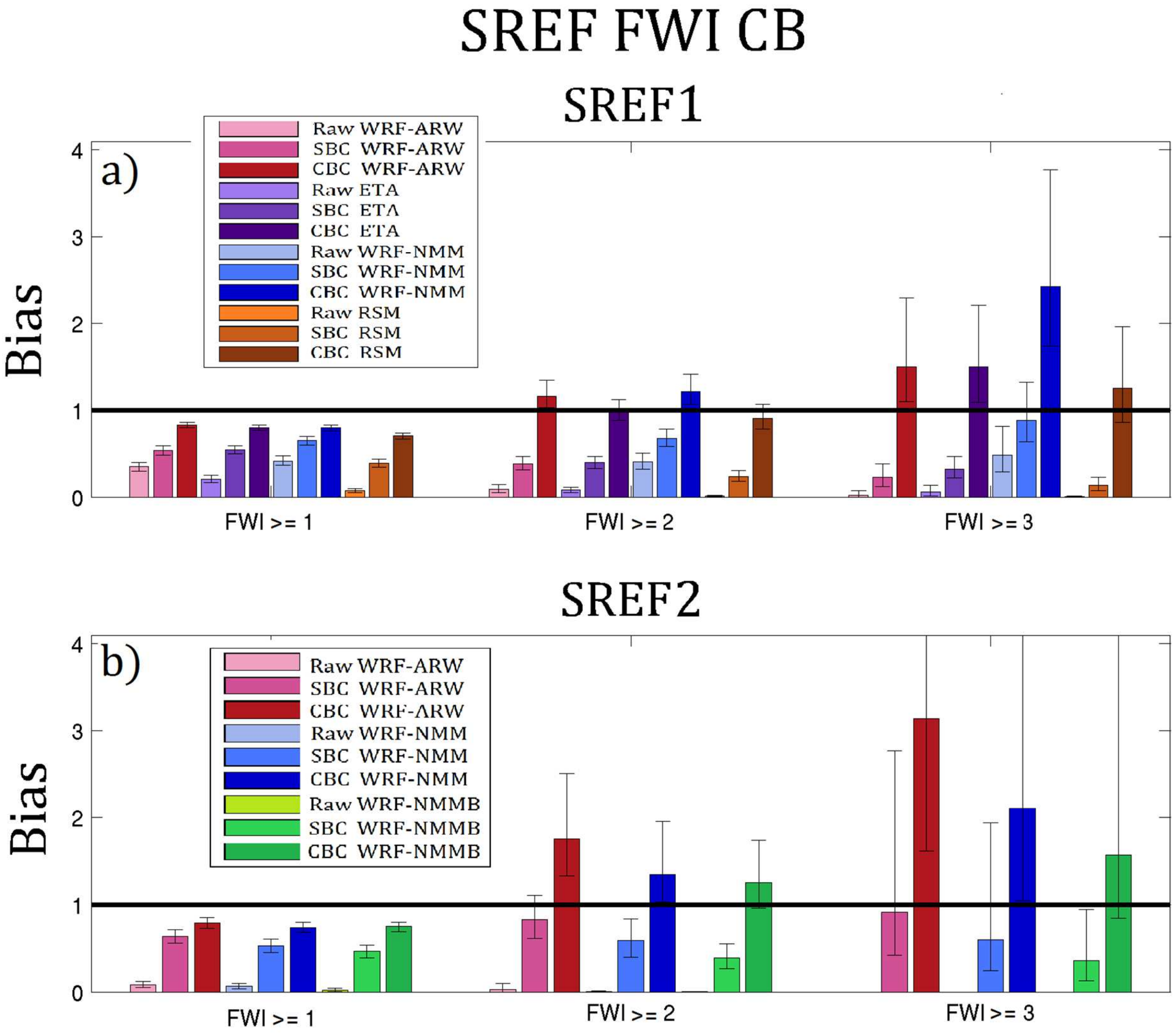

Figure 4.8: SREF based raw, SBC and CBC CB by threshold and model core (WRF-ARW – magenta, WRF-NMM – blue, WRF-NMMB – green, ETA – purple and RSM – brown) for SREF1 (a) and SREF2 (b). Error bars represent the 2.5th and 97.5th percentile of the 1000 resampled datasets.

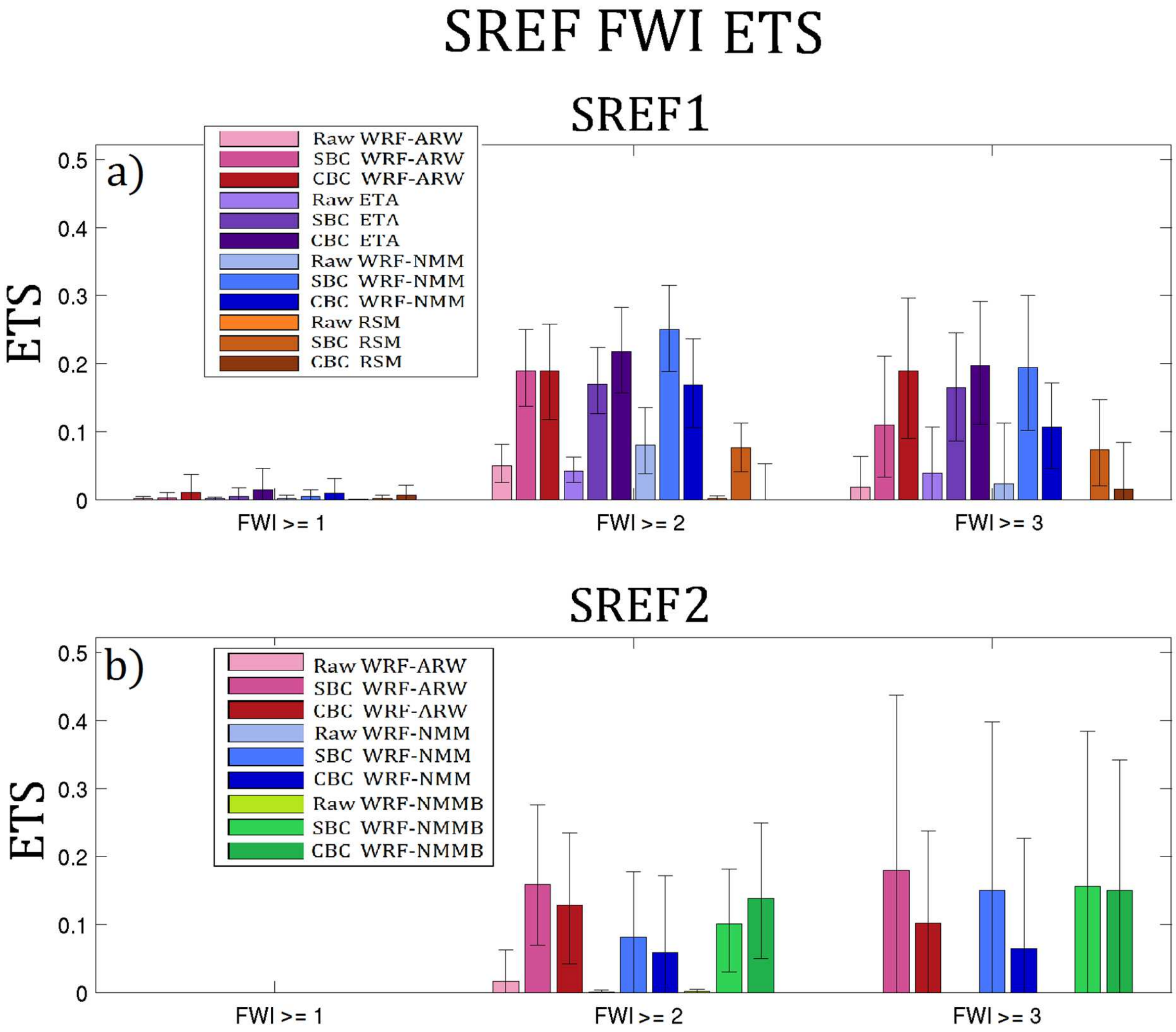

Figure 4.9: Same as figure 4.8, but for ETS rather than CB.

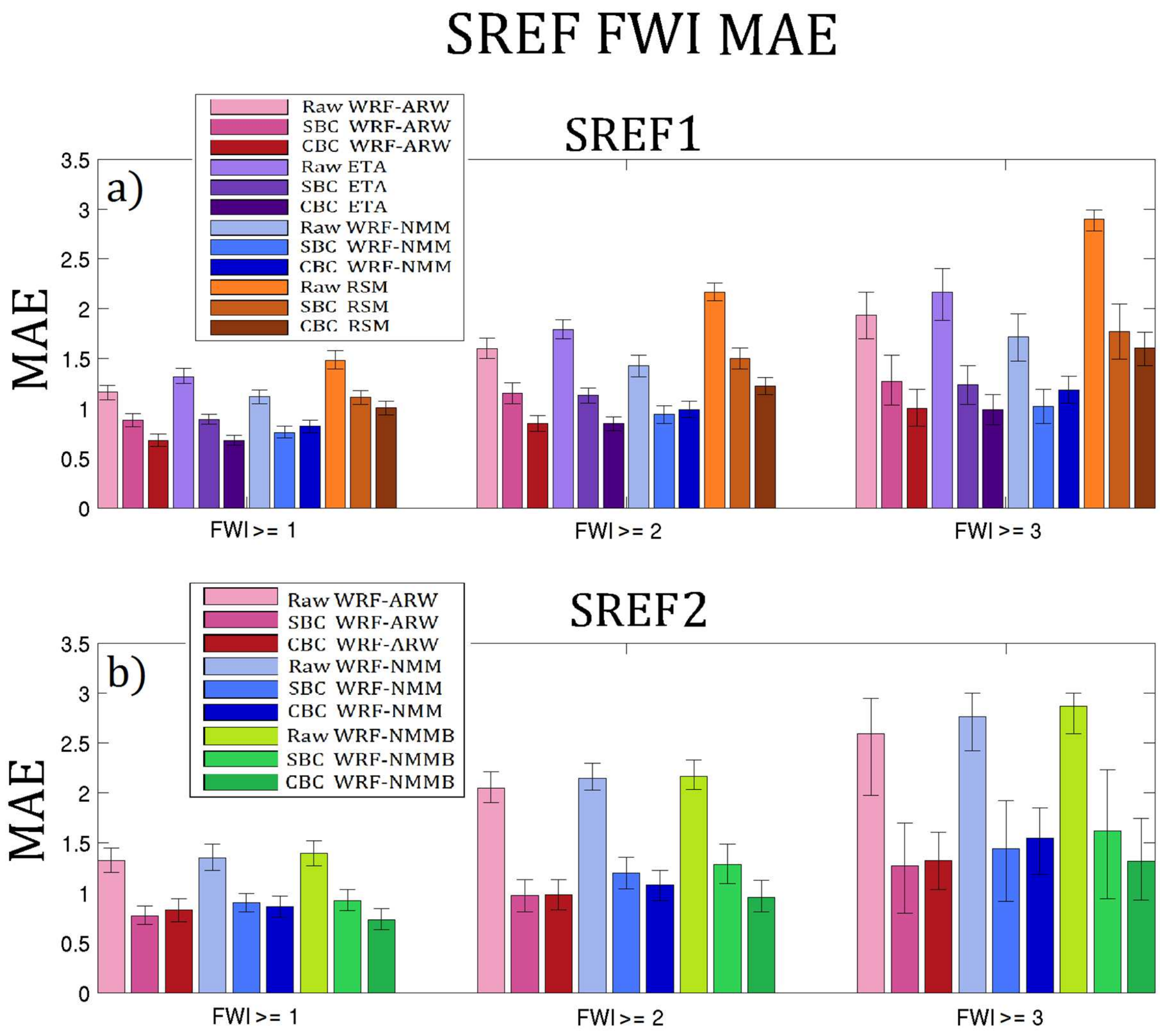

Figure 4.10: Same as figure 4.8, but for MAE rather than CB.

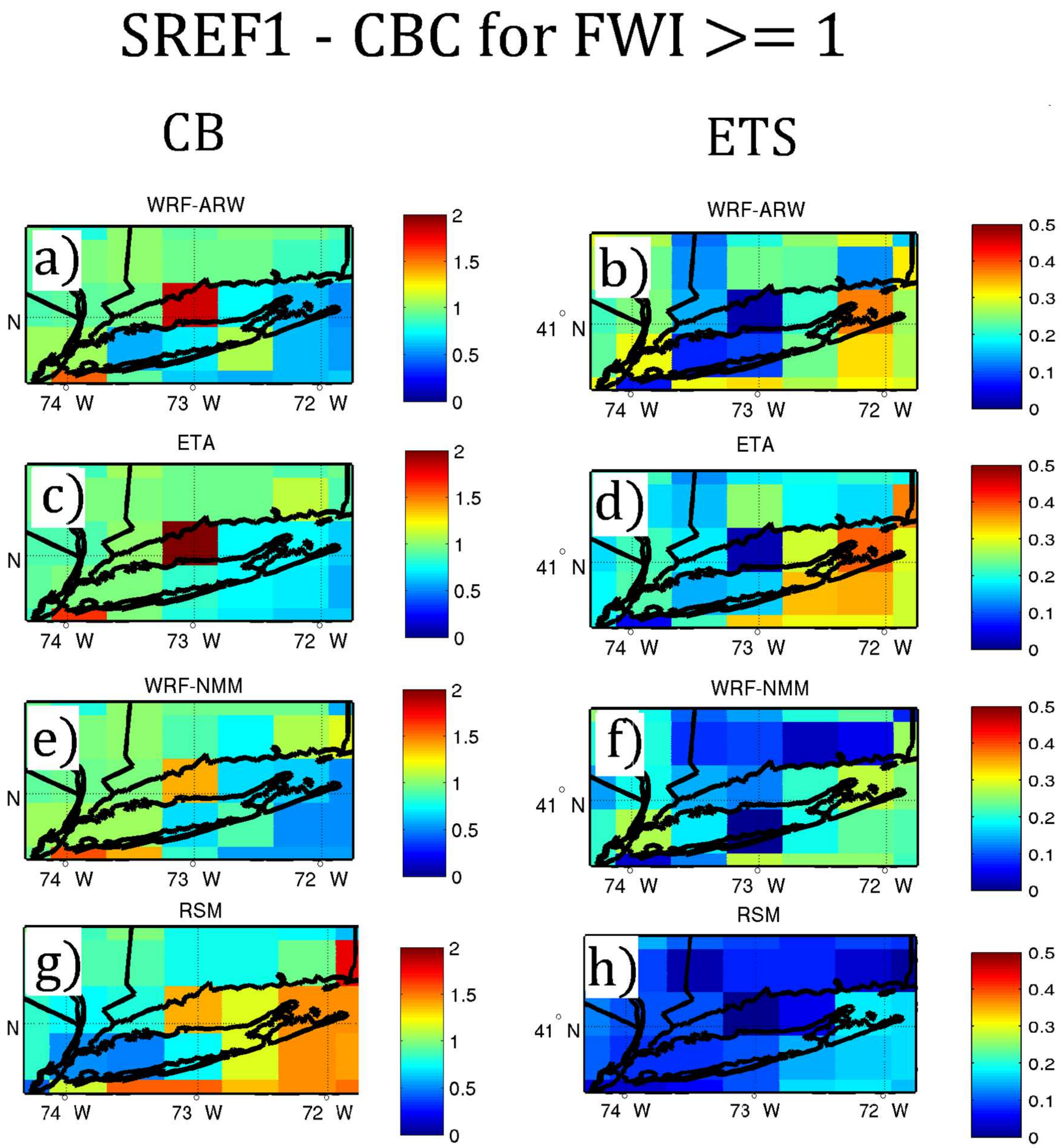

Figure 4.11: SREF1 spatial CBC at FWI >= 1 averaged over the WRF-ARW core (a,b), WRF-ETA (c,d), WRF-NMM (e,f), and RSM (g,h) for CB (a,c,e,g) and ETS (b,d,f,h).

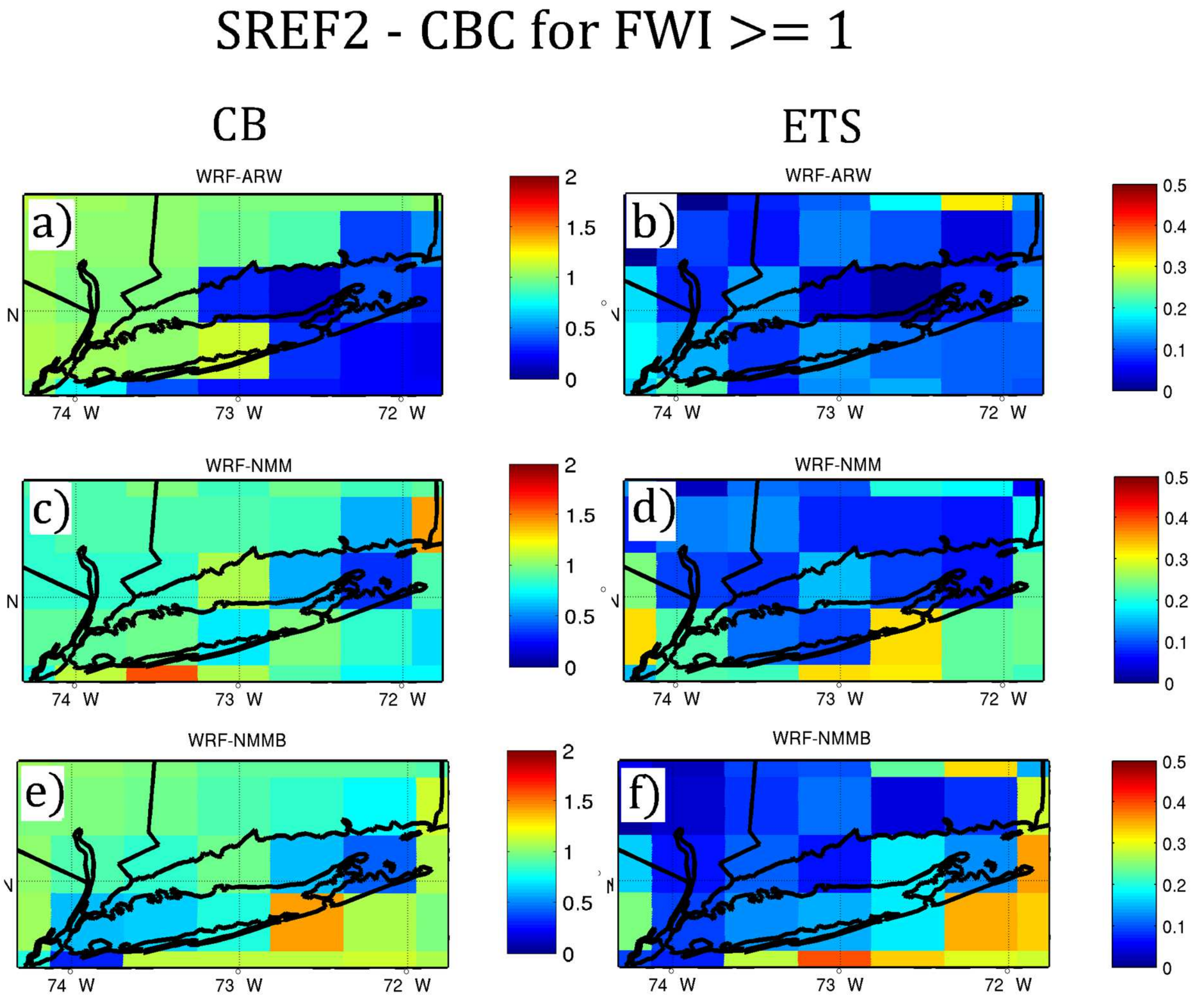

Figure 4.12: Same as Figure 4.11 but for SREF2.

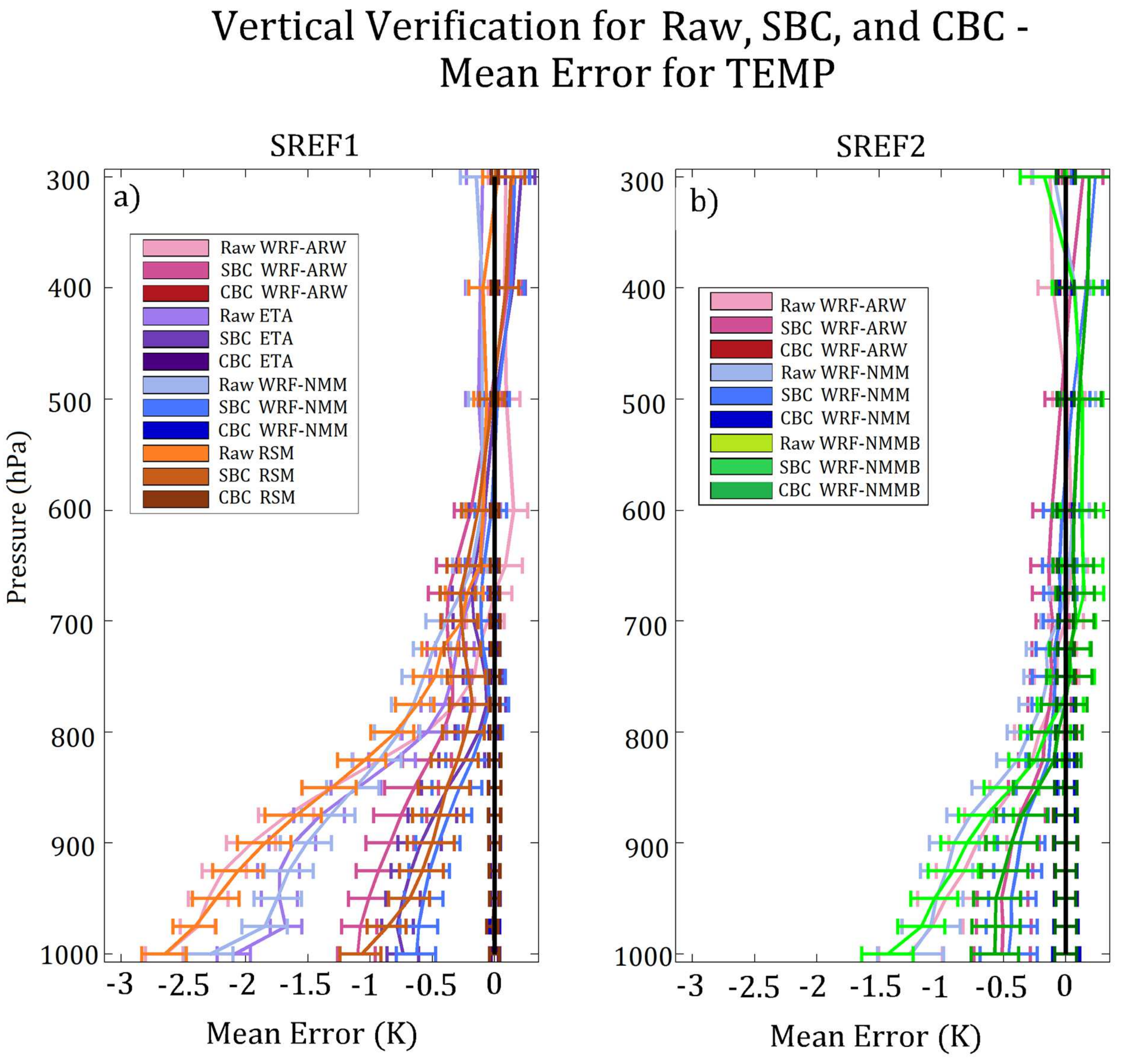

Figure 4.13: SREF based TEMP profile mean error for raw, SBC and CBC by threshold and model core (WRF-ARW – magenta, WRF-NMM – blue, WRF-NMMB – green, ETA – purple and RSM – brown) for SREF1 (a) and SREF2 (b). Error bars represent the 2.5th and 97.5th percentile of the 1000 resampled datasets.

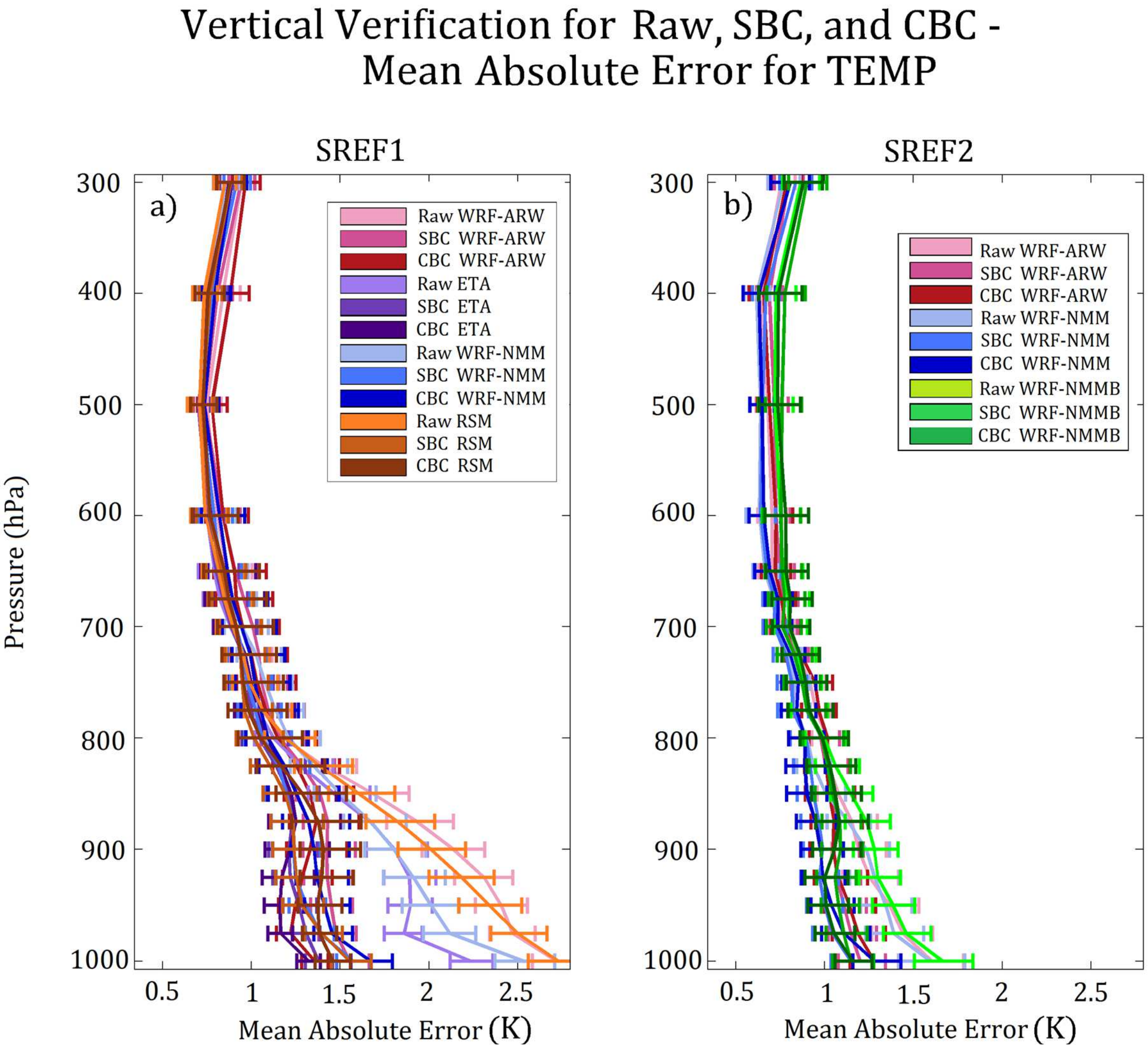

Figure 4.14: Same as figure 4.13, but for mean absolute error TEMP profiles

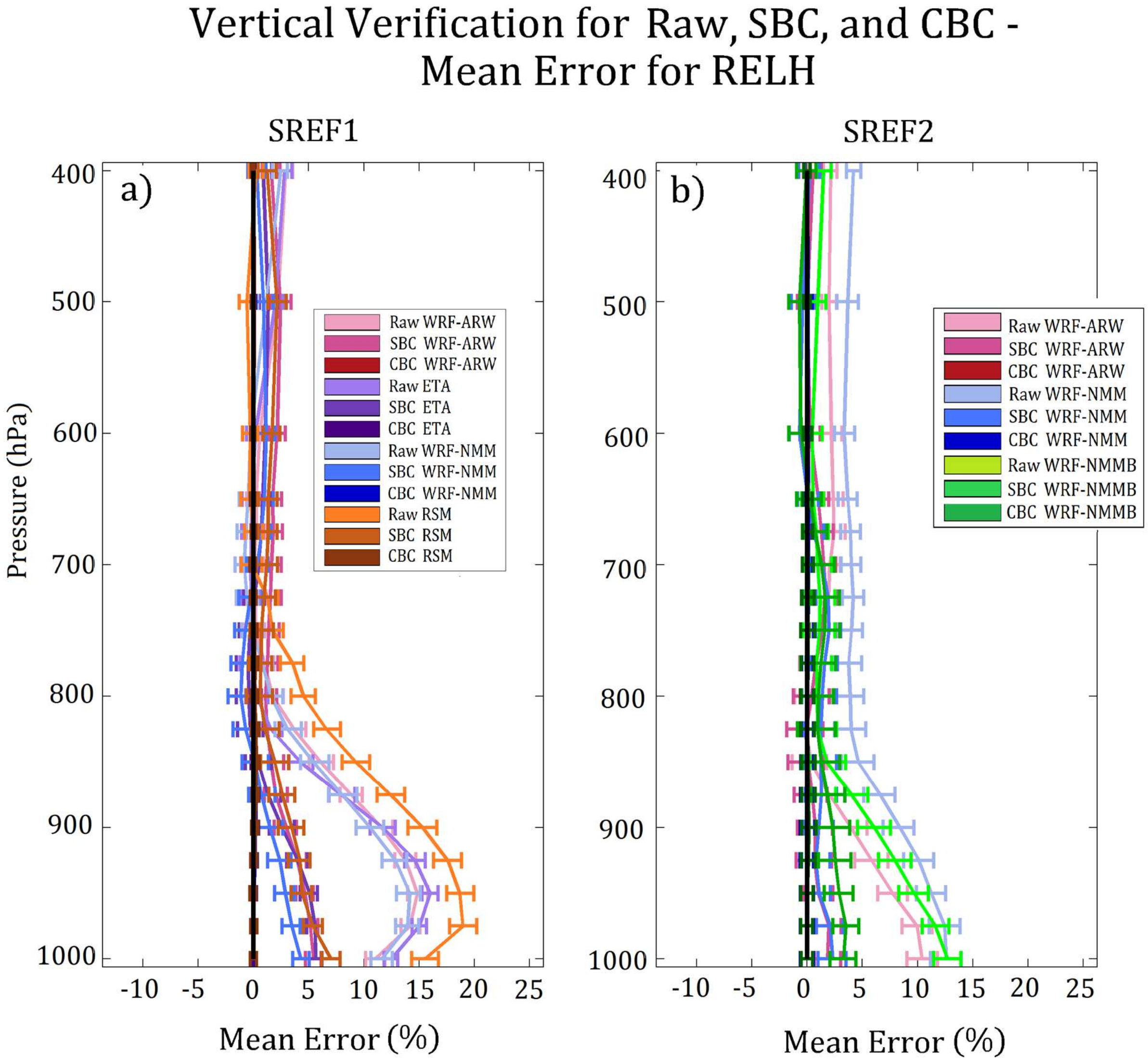

Figure 4.15: Same as figure 4.13, but for mean error RELH profiles.

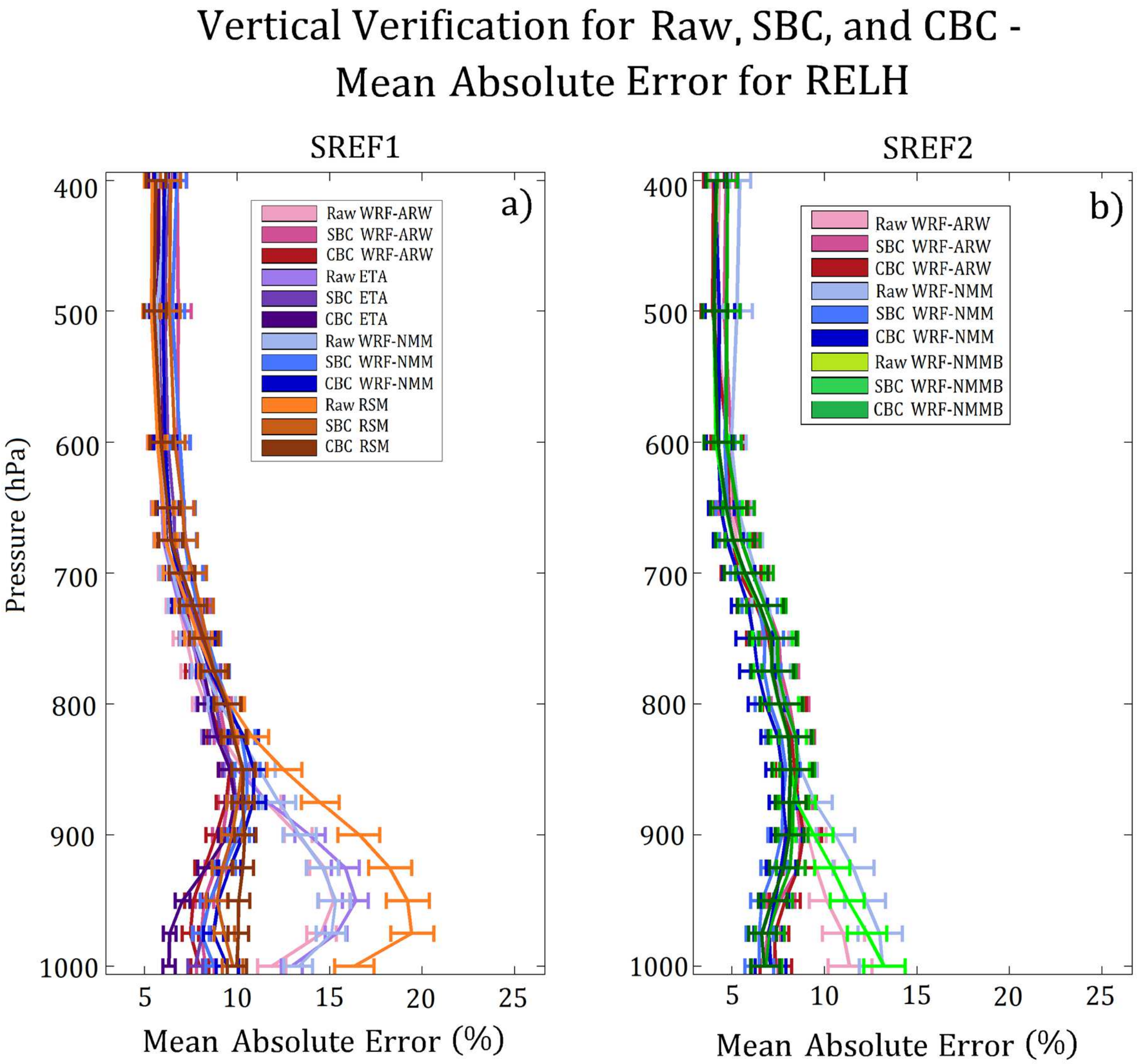

Figure 4.16: Same as figure 4.14, but for mean absolute error RELH.

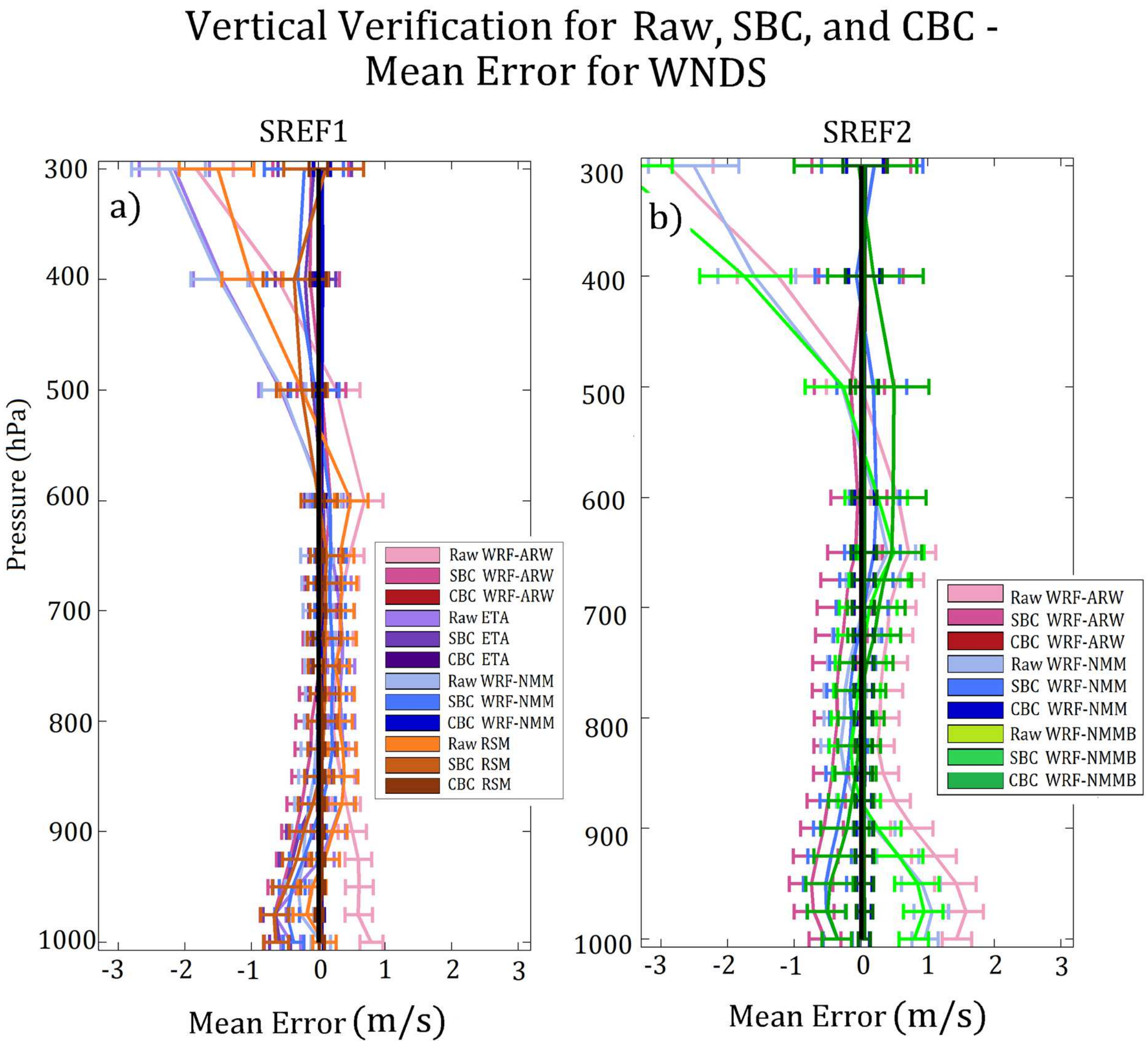

Figure 4.17: Same as figure 4.14, but for mean error WNDS profiles.

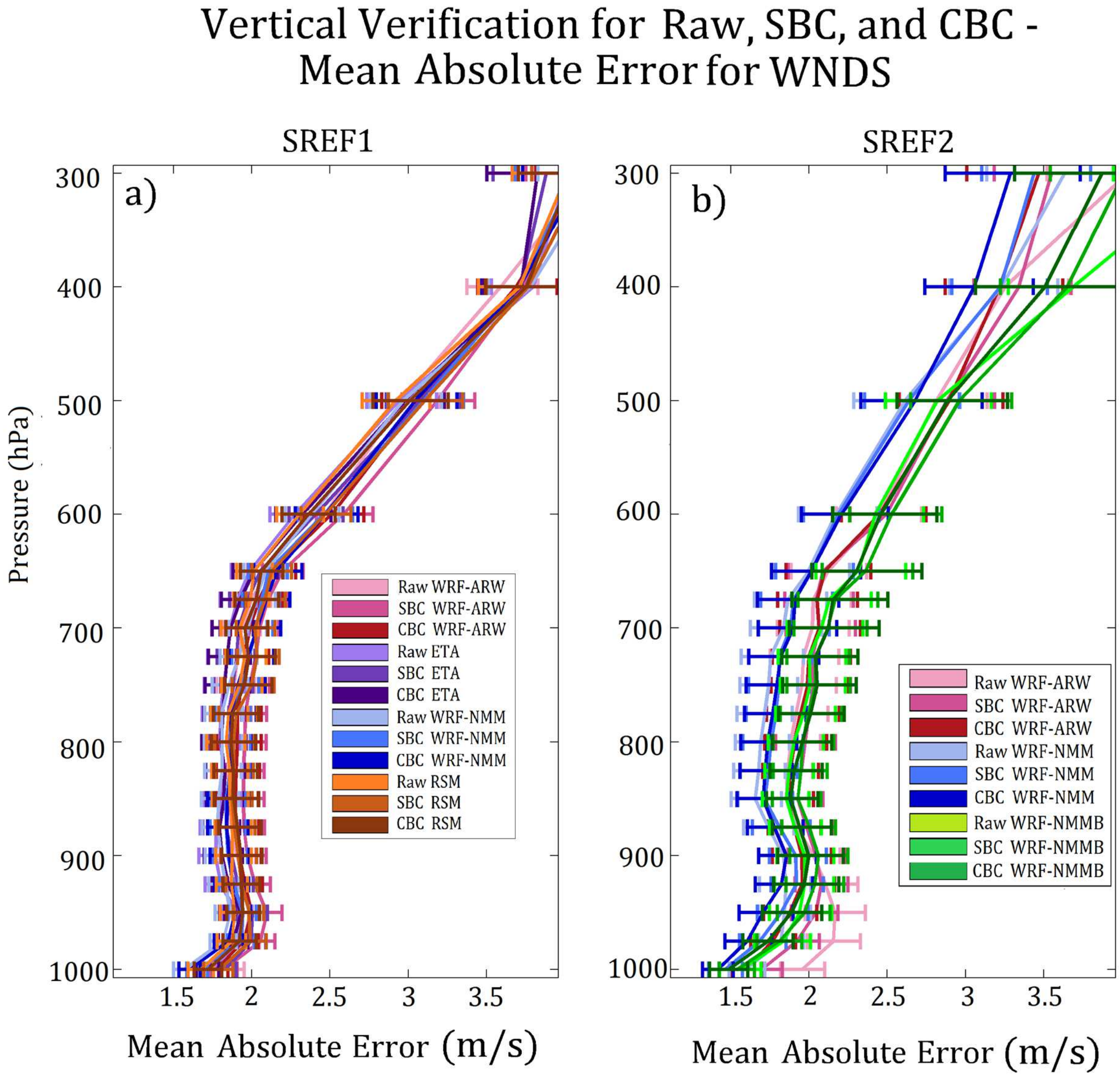

Figure 4.18: Same as figure 4.14, but for mean absolute error WNDS profiles.

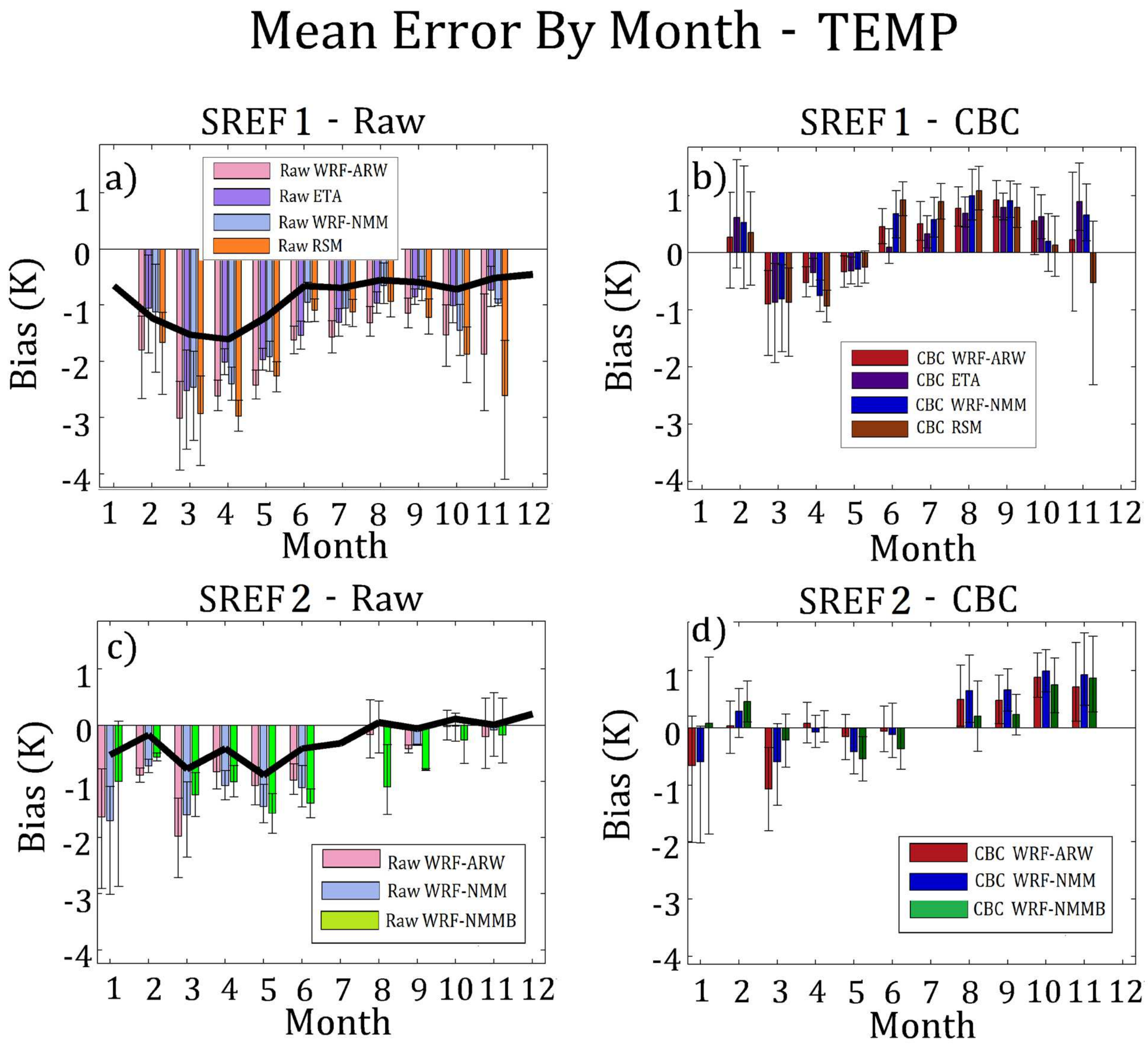

Figure 4.19: SREF1 raw (a), SREF1 CBC (b), SREF2 raw (c) and SREF2 CBC (d) based TEMP mean error by month and model core (WRF-ARW – magenta, WRF-NMM – blue, WRF-NMMB – green, ETA – purple and RSM – brown). Error bars represent the 2.5th and 97.5th percentile of the 1000 resampled datasets and the solid black line denotes the SREF ensemble averaged mean error by month (i.e. all days).

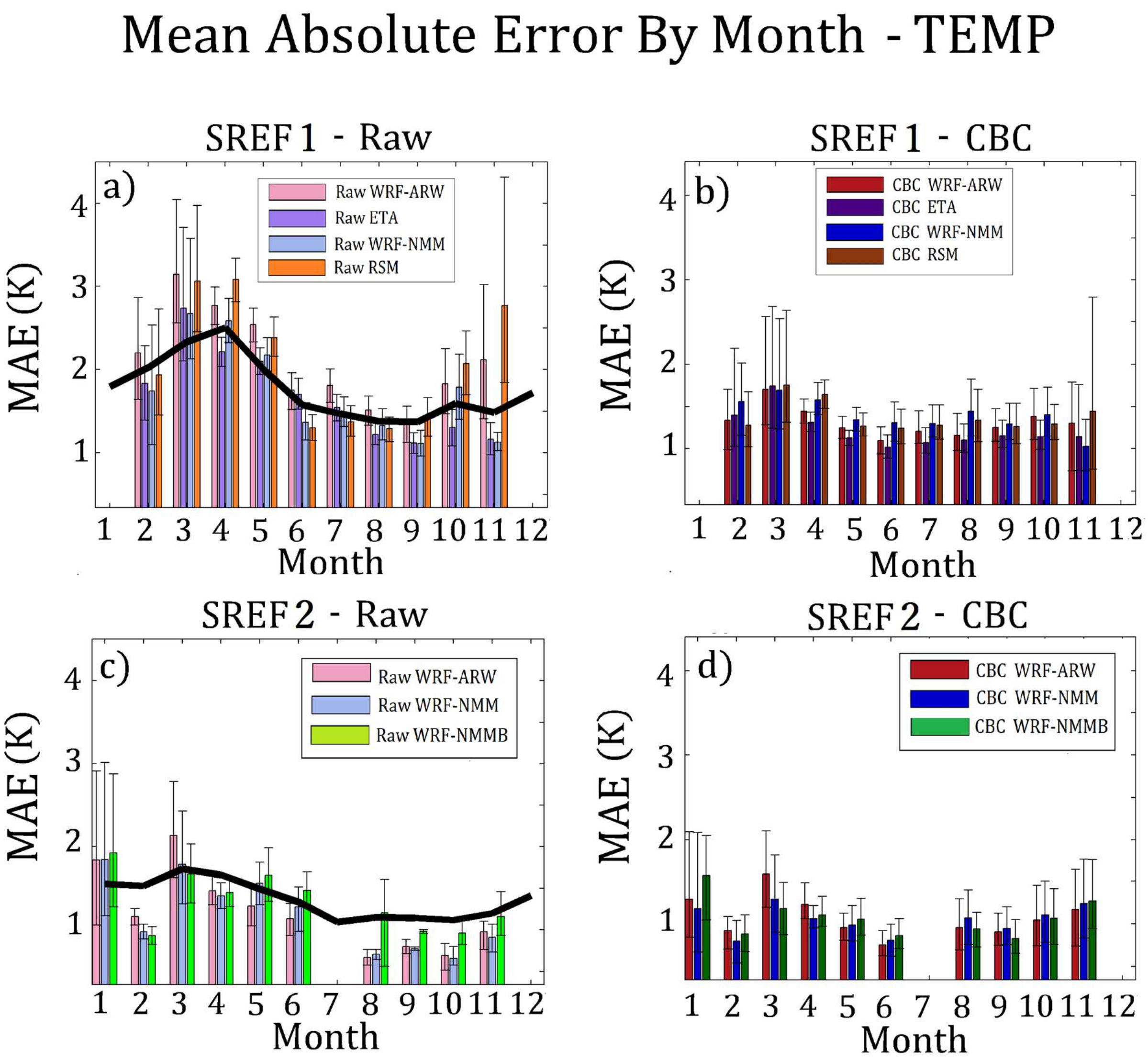

Figure 4.20: Same as Figure 4.19, but for TEMP mean absolute error by month.

Chapter 5:

Exploring Model Error with the Ensemble Kalman Filter on Fire Weather Days

5.1 Introduction

Chapters 3 and 4 quantify and correct for ensemble model bias on fire weather days (FWDs) using a variety of post-processing methods. Although operationally useful, post-processing does not directly improve the model simulation or provide insight into the structural model errors within the parameterized physics. Another approach is to explore potential sources of model error and bias specific to FWDs using data assimilation (DA). As shown in chapter 4.3c, FWDs have model bias and error maximized in the lower PBL near the surface, suggesting that the radiation scheme, land surface model, planetary boundary layer, or combinations of all three are a potential error source. In addition, chapter 3.3b shows that model biases on FWDs appear to be sensitive to the planetary boundary layer (PBL) scheme within the Weather Research and Forecast (WRF), suggesting the potential that some of the systematic bias on FWDs might be the result of the parameterized PBL.

In this chapter, the EnKF is used to explore a period of abundant FWDs in three separate month-long (April of 2012) simulations; a control run with exchangeable ensemble members, a multi-PBL ensemble run, and a simultaneous state and parameter estimation (SSPE) run to estimate relevant PBL parameters within the WRF model. This chapter addresses several novel goals with the EnKF approach on FWDs; including the quantification of how rapidly model biases develop in the model simulation, the impact and potential benefit of data assimilation, the result of a multi-PBL ensemble, and the effect of SSPE within the parameterized PBL on days that prominently feature a deep PBL.

Section 5.2 details the WRF-EnKF setup and domain, the sensitivity runs performed to adapt the EnKF for FWDs, the three month-long EnKF runs used to elucidate model error on FWDs, and the datasets used for EnKF verification. Section 5.3 presents the results for the sensitivity runs and month long runs, and details the impact of SSPE on model performance. Section 5.4 discusses the potential for structural model error in the parameterized PBL physics and some ideas for future work.

5.2 Methods

a. WRF-EnKF details

The Pennsylvania State University Ensemble Kalman Filter (PSU-EnKF; Meng and Zhang 2008 a,b) is selected as the data assimilation platform for this dissertation with forward iterations performed in six hour increments using the Weather Research and Forecasting (WRF) Advanced Research WRF (ARW) core Version 3.5 (Skamarack et al. 2008). Model physics include the WRF Single Moment 6-class (WSM6; Hong et al. 2004) microphysics, Kain Fritsch (KF; Kain and Fritsch 1990) cumulus parameterization, Rapid Radiative Transfer Model (RRTM; Mlawer et al. 1997) for short wave radiation, Dudhia long wave radiation (Dudhia et al. 1989) and the Noah Land Surface Model (LSM; Chen and Dudhia 2001). Unless otherwise mentioned, the Asymmetric Convective Model, version 2 (ACM2) PBL scheme (Pleim 2007) is also used.

The algorithm in the PSU EnKF is the square root ensemble filter (EnSRF; Whitaker and Hamill 2002) which uses the traditional Kalman gain to update the ensemble mean and a fractional Kalman gain to update the deviations from the mean. Localization of the background error covariance is applied with an element-wise multiplication of the covariance matrix using the successive covariance localization (SCL) of Gaspari and Cohn (1999). Inflation is applied to the perturbations after each analysis using the "covariance relaxation to the prior" method described in Zhang et al. (2004). For all runs, the relaxation coefficient is set to 0.6 (i.e. 60% of the updated perturbation is relaxed back to the prior). An ensemble of initial condition (ICs) and lateral boundary condition (LBCs) are created by drawing perturbations from the cv3 climatological background error covariance option (Barker et al. 2004) available with the WRF three-dimensional variational data assimilation. The WRF-EnKF system has been adapted to assimilate a variety of different observations; including data from buoys, ships, mesonet stations, profilers, rawinsonde soundings, satellite winds, Standard Aviation Observation (SAO), Aircraft Communication Addressing and Reporting System (ACARS), and Automated Surface Observing System (ASOS) stations.

b. Experimental design

April of 2012 is chosen to run the WRF-EnKF system due to the abundant number of FWDs over the NEUS region, and it coincided with the 1000+ acre Ridge-Manorville fire on 09 April 2012 in eastern Long Island, New York. FWDs in this chapter are defined using the fire weather index (FWI) developed in chapter 2. April 2012 consists of 17 FWD2 days, with 15 days reaching a category of 2, and 9 days reaching a category of 3. According to the United States Drought Monitor (Svoboda et al. 2002), a D1 to D2 drought enveloped the entire New York City region by 07 April (Svoboda et al. 2002). April 2012 was mostly dry with the exception of a significant rain event from 21 to 23 April, during which most ASOS stations in the tri-state region measured 50 mm to 75 mm of rain. There was also a minor rain event of less than 3 mm between 25-27 April.

The WRF-EnKF system consists of an outer domain (d01) at 27 km grid spacing and inner domain (d02) at 9 km grid spacing centered over the New York City Tri-State region (Fig. 5.1). The d02 domain is selected to encompass the D1 region from chapter 3 (Fig 3.1a) and chapter 4 (Fig. 4.1). All WRF-EnKF runs begin with perturbing the IC and LBC Global Forecast System (GFS) 6-hour forecasts. The 6-hour GFS forecast is selected over the zero hour analysis to be realistically timely in an operational setting, since the assimilated observations are available before the 00 hour GFS analysis. Thereafter, an initial 12 hour spin up is performed before the first data assimilation cycle to allow for physically consistent flow-dependent covariance structures to develop. Data is then assimilated sequentially every 6 hours for the remainder of the WRF-EnKF run. Only the results within d02 are presented.

c. Sensitivity studies with the PSU-EnKF over the NEUS

EnKF performance has never been tested on FWDs, and as a result some minor adjustments to the filter may be necessary to improve performance. The PSU EnKF is designed with a large number of tunable parameters, including (but not limited to) the degree of horizontal localization, vertical localization, inflation, observational error variance, observational thinning, the total number of WRF variables used in the Kalman gain, and the total number of ensemble members. Furthermore, there are several different methodologies for localization, inflation and perturbing the IC's and LBCs, although these different methodologies will not be explored in this dissertation.

Zhang et al. (2014) is used as an initial basis for the EnKF setup in this chapter, since they applied the PSU EnKF to high resolution simulations of convection over land using Meteorological Assimilation Data Ingest System (MADIS; Barth et al. 2002) and radar data. Following the localization setup of Zhang et al. (2014), the control values are set to 100 km for surface mesonet stations, 300 km for Automated Surface Observing System (ASOS), and 900 km for all observations above the surface. An observational thinning factor of 3 (0) is applied for all surface observations assimilated in d01 (d02) except for ASOS, where there is no observational thinning. The observational error for the assimilated observations comes directly from the assumed error in the MADIS dataset. However, additional tuning may be necessary to make the EnKF suitable for data assimilation on FWD2 events. Therefore, seven sensitivity runs are conducted to determine an appropriate EnKF configuration by adjusting the horizontal localization and observational error variance:

a) Control EnKF (EnKF1.0L)
b) EnKF1.0L but with the horizontal localization doubled at surface (EnKF2.0L).
c) EnKF1.0L but with the horizontal localization halved at surface (EnKF0.5L).
d) EnKF1.0L but with the observational error of surface stations reduced by half (EnKF0.5o).
e) EnKF0.5o but with the horizontal localization doubled at surface (EnKF0.5o2.0L).
f) EnKF0.5o but with the horizontal localization tripled at surface (EnKF0.5o3.0L).
g) EnKF0.5o but with the horizontal localization quadrupled at surface (EnKF0.5o4.0L).

The purpose of these sensitivity runs is not to determine the optimal EnKF configuration, but rather to identify a suitable EnKF setup for FWDs. The determination of an optimal configuration would require hundreds of separate simulations over multiple seasons in order to adequately sample the phase space for all tunable parameters. It would also require the usage of different localization, inflation, observational thinning, and filter type techniques. Such assessment is beyond the scope of this dissertation.

For these sensitivity runs, a subsample of days from April 2012 is selected between 06-11 April 2012. This time period coincides with a large number of consecutive FWDs and the occurrence of the Ridge-Manorville fire on 09 April 2012. The ensemble is initialized on at 0000 UTC 06 April 2012 and run forward for 12 hours, when the first set of observations is assimilated. Thereafter, the observations are assimilated every 6 hours until 0000 UTC 11 April. To increase sample size, the verification domain in Fig. 4.1 is expanded to encompass 39^{o}N - 42^{o}N and 78^{o}W - 70^{o}W (when available) for these trial runs. Given the small number of total days in this verification, bootstrapping is not performed to assess the statistical significance.

d. Adapting the PSU-EnKF for SSPE

There is no universal statistical formulation within the parameterized physics of the WRF that represents the optimal configuration. For this reason, multiple variations of parameterized physics packages exist to approximate what cannot be explicitly resolved. Subsequent to the assumptions within the statistical physics are static parameters, the values of which are not always known with complete certainty. Not all parameters are ideal candidates for estimation because not all parameters are identifiable (Nielsen-Gammon et al. 2010). An identifiable parameter must project strongly and monotonically onto the model phase space and be uncorrelated with the influences of other parameters (Nielsen-Gammon et al. 2010). Finding identifiable parameters requires running a large number of WRF simulations with the adjusted candidate parameter values and then analyzing their impact on the model phase space. Fortunately, Nielsen-Gammon et al. (2010) has already evaluated parameter sensitivity within the ACM2 PBL scheme for the purposes of parameter estimation and concluded that without the assimilation of profiler data, the two most identifiable parameters are PVAR (an exponent that affects the magnitude and vertical distribution of eddy diffusivity in an unstable PBL) and KVVAR (affects vertical mixing in a stable PBL). Therefore, this dissertation modifies the WRF and EnKF system to perform SSPE with PVAR and KVVAR in the ACM2 PBL.

Both the WRF and the EnKF are adapted to treat the estimated parameters as a two dimensional model state variable with SSPE. This dissertation uses the "spatial updating" method of Aksoy et al. (2006b). First, the parameter values are augmented in the EnKF state matrix as a homogenous two dimensional field. The same localization and inflation applied to the state variables are also applied to the parameters. The EnKF then updates all state variables, which creates spatial variability in the two dimensional parameter estimate field. The spatially varying parameter values are replaced with the homogenous spatial average. Thereafter, within the WRF, the updated parameter states are feed into the relevant parameterized physics modules to influence the forecast. In this manner, the WRF-EnKF SSPE attempts to estimate both the optimal parameter values and the optimal model analysis.

To create an initial ensemble of parameter estimates, values are drawn from a random uniform distribution of theoretically realistic potential values (i.e. between 1 and 3 for PVAR and between 0.0003 and 0.05 for KVVAR). The parameter ranges for PVAR and KVVAR are determined from Nielson-Gammon et al. (2010), but the lower range for KVVAR is adjusted downward due to the change in the default KVVAR from 0.1 to 0.01 within WRF ACM2. After the initial draw, the only bounds on PVAR and KVVAR are that they remain positive definite. Hu et al. (2010b) constrained the parameter values using a parameter transformation technique, but this is decided against to explore the temporal variability of the parameter values. Of particular interest are any variations in parameter values as a result of synoptic flow pattern or FWI value, which might provide insight into potential PBL structural model errors (i.e. misspecification of a terms functional form; Golaz et al. 2007). Finally the "conditional covariance inflation" technique of Aksoy et al. 2006a is applied, which preserves the variance limit of each parameter to ¼ of its original value.

e. Full EnKF runs for April of 2012

Using the ideal EnKF configuration from the sensitivity runs in section 5c, three WRF-EnKF runs are performed between 0000 UTC 01 April 2012 and 1800 UTC 30 April 2012:

a) A 45-member control ensemble (EnKF_control).
b) A 45-member multi-PBL ensemble with 15 members running the ACM2 PBL, 15 members running the Yonsei Univeriesty (YSU; Hong el al. 2006) PBL, and 15 members running the Mellor, Yamada Nakanishi and Niino (MYNN; Nakanishi and Ninno 2006) PBL (EnKF_multiPBL).

c) A 45-member control ensemble while performing SSPE for PVAR and KVVAR within the ACM2 PBL scheme (EnKF_parsest).

The ensemble is initialized at 0000 UTC 01 April 2012 and run forward for 12 hours before the first set of observations are assimilated on 1200 UTC 01 April 2012. Observations are sequentially assimilated every 6 hours until 1800 UTC 30 April 2012, and the first two days are discarded for spin up. The verification region is identical to the domain used in developing the FWI (i.e. 40.5ºN - 42ºN and 74.5ºW – 71.5ºW; see Fig 4.1) within the d02 inner nest.

f. EnKF verification

EnKF performance is complicated by a compromise between using all available observations to optimize filter performance and withholding observations for verification. In the former case, filter verification is limited by the observations not being independent from those being assimilated. This dissertation verifies EnKF performance using three different methodologies; 1) all background and updated ensemble model data is interpolated and compared to the observations directly assimilated into the filter, 2) all background and updated ensemble models are interpolated to available data in the MADIS archive and 3) the background and updated fields are compared to the Rapid Update Cycle (RUC) analysis.

EnKF assimilates many, but not all of the available MADIS datasets. Therefore, the MADIS datasets not assimilated by the EnKF could be considered independent verification. The dependent datasets include; ACARS, ACARSP (i.e. ACARS profiles or ACARS data organized into profiles over each airport location), MARINE (includes all buoys and ships), METAR (ASOS stations), MESONT (surface mesonet data), and HDW-E (satellite winds). Independent data not assimilated into the EnKF from MADIS includes CRN (Climate Reference Network), NEPP (New England Pilot Project), POES (Polar Orbital Environmental Satellites), and MULT-P (multi-agency profiler data). Results from the background and update of the EnKF represent the 45-member ensemble mean and are tri-linearly interpolated (using the log of pressure for observations above the surface) to the MADIS observations.

Several observation-space diagnostic statistics are computed to evaluate the EnKF system performance by interpolating the background and update to the location where the observations are assimilated (i.e. first method mentioned two paragraphs above). Diagnostics include the root mean squared innovation (RMSI; the root mean squared observation minus model), the ensemble spread (defined as the average ensemble deviation from the ensemble mean), and the consistency ratio (Yussouf et al. 2013; their Eq. 4). The consistency ratio compares the observational error plus the ensemble spread to the mean square innovation. A value of one indicates that the ensemble variance is a good approximation of the forecast error variance for the assumed observational error. Finally, traditional metrics such as mean error (ME) and mean absolute error (MAE) are analysed in bulk, by height, time, and spatially. ME and MAE are also used to compare EnKF performance for FWDs and non-FWDs.

5.3 Results

a. Determining an appropriate EnKF configuration

The ME results for all sensitivity runs (section 5.2c) interpolated to the MADIS observations are shown for the background (Fig. 5.2a), update (Fig. 5.2b), and RUC analysis (Fig. 5.2c). For these statistics model biases are averaged over for the last four days (07-11 April 2012) at 00 UTC, 06 UTC, 12 UTC, and 18 UTC. There is a large average cool TEMP bias (-1.21 K) in the background analysis when comparing with all surface (i.e. MARINE, METAR, CRN, MESONT, and NEPP) observations, which is considerably smaller aloft (i.e. POES, ACARS, and ACARSP; averaging -0.23 K). An exception occurs with MULT-P (-2.41 K), which has a considerably smaller sample size of 15 observations (Table 5.1) compared to other MADIS datasets. The presence of a large low-level cool TEMP bias is consistent with Fig. 3.5 and Fig. 4.13, albeit in this case the biases develop within 6 hours after initialization. The rapid growth of short-term low level cool TEMP biases suggests that even in the presence of data assimilation, short-term TEMP forecasts will still underestimate FWI values.

All EnKF updates (Fig. 5.2 b) using the default observational error (EnKF1.0L, EnKF0.5L, EnKF2.0L) have a considerably larger cool TEMP bias (averaging -0.95 K for all surface observations) compared to the RUC analysis (Fig 5.2c; -0.34 K) although all runs with the reduced observational variance (EnKF0.5o, EnKF0.5o2.0L, EnKF0.5o3.0L, EnKF0.5o4.0L) are more comparable to the RUC (averaging -0.53 K). Thus, a reduced observational error should be used on FWDs, although the localization radius appears to also influence EnKF performance. For instance, EnKF0.5o3.0L is the least biased across all observation types (while neglecting MULT-P; -0.27 K) and closest to the RUC analysis (-0.16 K).

The decrease in updated ME for the reduced observational error runs also improves updated MAE (Fig. 5.3b) for the surface observations (improvement averaging 0.23 K), while having little impact on average MAE aloft (improvement averaging 0.02 K). This is encouraging since the goal of the reduced observational error runs is to improve surface bias while not degrading assimilation results aloft. The averaged surface MAE for EnKF0.5o3.0L is best for all reduced observational error runs and averages 0.13 K less than EnKF0.5o. The EnKF0.5o3.0L run has slightly better MAE than the RUC analysis aloft (averaging 0.11 K) and at the surface (averaging 0.09 K). The comparable performance of EnKF0.5o3.0L to RUC is encouraging given that the RUC assimilates additional observations such as GPS precipitable water and radar reflectivity (Weygandt et al. 2014). Given that the PSU-EnKF assimilates fewer observations than the RUC's data assimilation, these results suggest that further improvements to the EnKF are possible by adding more observations. Furthermore, these results suggest that the filter is performing well for TEMP with a reduced observational error, particularly for the EnKF0.5o3.0L run.

The spatial TEMP ME for d02 is presented using MESONT observations averaged over a 0.2° latitude by 0.2° longitude grid for all EnKF sensitivity runs (Fig. 5.4). Grid points with greater than 80% missing data are removed to reduce noise in the statistics. Most grid points analyzed for EnKF1.0L have a negative TEMP bias with the greatest cool biases over eastern Pennsylvania, Connecticut and eastern Long Island (ME = -1.2 K; Fig. 5.4a). However, bias is more effectively removed over the more urban regions between Philadelphia and New York City (ME = -0.2 K). Altering the localization (i.e. EnKF2.0L and EnKF0.5L) does not impact the filter performance considerably when verifying with MESONT stations. However, the reduced observational variance improves the spatial ME overall (ME values between -2.00 K and 1.39 K) and is more comparable to the RUC analysis (ME values between -2.21 K and 1.87 K) across the domain.

To explore the vertical impact of the EnKF sensitivity runs, MAE is plotted by height for the POES (a,d), ACARS (b,e) and ACARSP (c,f) background, update and RUC analysis (Fig. 5.5). The background TEMP MAE for the ACARS and ACARSP is larger than the RUC MAE between 700 hPa and the surface. This is the result of a large negative bias in TEMP in the 6 hour WRF forecast discussed earlier. The average updated MAE for ACARS and ACARSP is comparable between EnKF runs and the RUC, suggesting all analyses are very similar above the surface. Therefore, the effect of surface observational error adjustment or localization does not degrade or improve the EnKF performance aloft. Although POES is not assimilated into the EnKF, the background and updated EnKF compared to POES is unchanged and lower than that of the RUC analysis. POES data should be treated with caution since verification with this dataset results in substantially different ME and MAE profiles compared to ACARS or ACARSP.

Similar to the bulk MAE for TEMP, bulk MAE for WNDS is shown in Fig. 5.6. Results are mixed when comparing the reduced and default observational error runs. However, the EnKF0.5o3.0L MAE is improved over EnKF1.0L for all observations (averaging 0.14 m s^{-1}) except for METAR and MARINE, where there is a slight degradation (averaging 0.04 m s^{-1}). These changes are very minor and the filter adjustments made to TEMP do not appear to negatively impact the results for WNDS. However, caution should be used with some near-surface WNDS observations that do not have properly mounted anemometers, since this can result in an underestimation of the observed WNDS (Benjamin et al. 2007). This appears to be a problem with the MESONT and NEPP observations, which have a collective average of 0.80 m s^{-1} lower in the background than the better regulated METAR observations. Although there are quality control (QC) checks for both the MADIS mesonet system and PSU-EnKF, the negative WNDS bias of MESONT and NEPP stations generally pass through QC. Future adjustments to the EnKF assimilation of WNDS observations should consider this likely spurious negative bias in MESONT and NEPP. Regardless, average MAE across all observations is comparable between the optimal EnKF update (EnKF0.5o3.0L; MAE = 1.82 m s^{-1}) and RUC analysis (MAE = 2.11 m s^{-1}).

WNDS vertical MAE is presented for HDW-E, ACARS and ACARSP for the background, update and RUC analysis (Fig. 5.7). In general, there is some variability between EnKF sensitivity runs in the background MAE, with EnKF0.5L having slightly higher MAE than the ensemble mean (by 0.15 m s^{-1}). As with TEMP aloft, WNDS is not significantly impacted by adjustments to the surface observational error or localization in the sensitivity runs. In addition, WNDS MAE in the update is lower compared to the RUC for all levels between 200 hPa and 1000 hPa for both ACARS (averaging 0.71 m s^{-1}) and ACARSP (averaging 0.87 m s^{-1}). Update EnKF results for MAE are mixed when verified against HDW-E (EnKF improvement averaging 0.44 m s^{-1}), with the RUC having slightly lower MAE at 400 hPa, the update having slightly lower MAE between 800 hPa and 500 hPa, and comparable results at all other levels (Fig. 5.7d). In general, the updated EnKF performs comparable to or slightly better than the RUC for most WNDS observations assimilated.

Bulk MAE averaged across observation type for SPHU is shown in figure 5.8. There is a greater MAE in the background field for surface SPHU measurements, which is the result of a positive near-surface moisture bias in the model (averaging 0.35 g kg^{-1}; not shown but similar to Fig. 3.8 and Fig. 4.16). Interestingly POES SPHU has very large MAE for the background, update and RUC analysis, suggesting it is quite different from the other data sets. The difference in POES SPHU MAE is largely the result of this dataset having higher SPHU below 700 hPa, which is maximized at 1000 hPa (SPHU = 1.2 g kg^{-1}; not shown). POES may have a strong wet bias in the lower atmosphere and caution should be used when attempting to assimilate this data (recall that POES is not assimilated by the EnKF). The updated analysis improves MAE slightly across all observation types excluding POES (averaging 0.08 g kg^{-1} for EnKF1.0L) with a small additional benefit from the reduced observational error sensitivity runs (averaging 0.17 g kg^{-1}). In addition, the reduced observational error runs have lower SPHU MAE than the RUC analysis for all observational types analyzed except for MARINE.

When considering TEMP, WNDS and SPHU results overall, the greatest impact on the updated fields is related to halving the observational error when assimilating surface observations. This results in improved TEMP MAE and a reduction in surface bias. Furthermore, WNDS and SPHU also exhibit slight improvements in surface MAE. Although many of the observations used in this verification are not independent, comparisons to the RUC analysis provide perspective since many of the same observational datasets are used to create both analyses. Furthermore, the halved surface observational runs reduce near-surface ME and MAE when verified against independent observations from NEPP and CRN. Therefore, the observational error is reduced by half when running the full EnKF throughout April 2012. The determination of the ideal localization is more difficult since the results are generally less sensitive to adjustments in the radius of influence. Given that the average optimal MAE for TEMP is achieved with EnKF0.5o3.0L while not degrading the results for WNDS or SPHU, EnKF0.5o3.0L is chosen as the ideal configuration for the month long simulations of EnKF_control, EnKF_multiPBL and EnKF_parsest

b. Observation-space diagnostics for the April 2012 EnKF_control run

Before analyzing EnKF performance on FWDs versus non-FWDs, a few common EnKF observation-space metrics are analyzed to ensure the filter is behaving properly for EnKF_control. The verification in this section uses observations directly assimilated into the EnKF with model data interpolated to the assimilated observation location. This approach differs from that in section 5.3a, where available observations from the MADIS archive are used. The consistency ratio time series of EnKF_control is shown in Fig. 5.9 for the U component of the wind (UWND), V component of the wind (VWND), theta (THTA), temperature (TEMP), specific humidity (SPHU), and dew point (TMPD). Fujita et al. (2007) found that assimilating THTA and TMPD may be more beneficial in the presence of a well-mixed PBL. Therefore, if pressure is available, all assimilated moisture observations are converted to TMPD and all assimilated TEMP observations are converted to THTA. As a result, the majority of THTA observations are sampled above the surface and the majority of TEMP observations are taken near the surface.

Average consistency ratio values for UWND and VWND are greater than one (1.91 and 1.62 respectively), suggesting that the observational error variance is too large or the ensemble is overspread. It might be possible to decrease the consistency ratio by reducing the relaxation coefficient or increasing the localization of the WNDS observations. However, the WNDS consistency ratios appear to be stable in time and a comprehensive optimization of the EnKF is beyond the scope of this dissertation. THTA consistency ratio values initialize high (2.58) but generally average around 1.15 for most of the simulation. The high SPHU consistency ratios (averaging 3.45) should be viewed with caution due to a lack of assimilated observations (consisting of 6530 out of 613546 total simulations within d02 for the 30 day simulation). TEMP and TMPD have consistency ratios below one (averaging 0.42 and 0.39), which is probably the result of a lack of ensemble spread from a heavy reliance on the parameterized model physics near the surface.

Consistency ratio averaged by height is shown in Figure 5.10 for all variables but SPHU, which is removed due to its limited sample size. The high consistency ratios of UWND and VWND are continuous throughout the depth of the troposphere and vary between 1.3 and 2.7. THTA exhibits lower consistency ratios with a peak value of 1.7 at 500 hPa. The near-surface consistency ratios are considerably smaller and vary between 0.39 and 0.66 for TMPD, UWND, VWND and TEMP. These results are generally consistent with the findings from Fig 5.9.

Figure 5.11 shows the time averaged vertical mean error (ME) by model state variable for the background (dashed) and update (solid). A background cool THTA bias exists from the surface (-2.94 K) to 900 hPa (-0.59 K), which is somewhat corrected in the update at the surface (-1.19 K) and at 900 hPa (-0.20 K). There is also a model dry bias for TMPD (-2.81 K) that is largely removed after DA (-0.40 K). Other notable biases are found for background surface TEMP (-1.83 K) and a low level northerly wind bias (VWND = -2.82 m s^{-1}), which are largely removed in the update (TEMP = -0.15 K and VWND = 0.37 m s^{-1}). The low level cool bias is consistent with the warm season average cool bias found in Fig. 3.5, but the low level dry bias differs slightly from the warm season average wet bias of Fig. 3.8. There are significant variations of ME with height for UWND and VWND, which is likely the result of errors in the placement and orientation of various jet stream positions throughout the month as the average WNDS value increases with height. In practically all cases, the updated field reduces ME compared to the background for all variables shown. As a result, most updated model state variables exhibit lower MAE compared to the background (Fig. 5.12). The greatest improvement in MAE is found near the surface for THTA (2.15 K), TEMP (1.40 K), UWND (0.13 m s^{-1}) and VWND (0.73 m s^{-1}), with UWND and VWND exhibiting respectable improvements in MAE near the jet stream (300 hPa averaging 1.48 m s^{-1} and 2.44 m s^{-1} for UWND and VWND). From these results, the EnKF appears to be working to reduce the short-term development of model bias and error for all model state variables and heights analyzed.

Figure 5.13 shows the time series of total ensemble spread (ensemble spread plus observational error variance; dashed) and RMSI (solid) for UWND and VWND (a), THTA and TEMP (b) and SPHU and TMPD (c). X's in the top of Figure 5.13a denote FWI values greater than zero; with a small, medium, and large X indicating a FWI of 1, 2, and 3, respectively. UWND and VWND have similar total ensemble spread and RMSI, although they are not well correlated in time. Between 01-20 April 2012, the total ensemble spread is slightly greater than RMSI, resulting in the larger consistency ratios noted in Figure 5.9 and Figure 5.10. After 20 April 2012, total ensemble spread and RMSI are more comparable for both UWND and VWND. Although subjective, this coincides with a period of stormy weather between 22-26 April 2012 and the presence of increased upper level winds through the remainder of the month. Therefore, consistency ratios may vary depending on the synoptic flow pattern. RMSI and total ensemble spread are roughly comparable for THTA on the average (Fig. 5.13b), although the spread exhibits very little variability in time. On average, TEMP has significantly greater RMSI than total ensemble spread, which is likely the result of near-surface parameterizations reducing the ensemble portion of the spread. A similar affect occurs with TMPD (Fig. 5.13c), which is likewise primarily observed near the surface. SPHU RMSI (Fig. 5.13c) averages less than the total ensemble spread, which is supported by the high consistency ratio time-series in figure 5.9.

Overall, the state-space diagnostic plots (total ensemble spread, RMSI, and consistency ratio) are within the ranges found for previous EnKF applications (Yussouf et al. 2013, Tanamachi et al. 2013), suggesting that the current implementation of the filter is appropriate for studying FWDs. Furthermore, the EnKF reduces the short-term development of model bias and non-systematic error in the 6-hour WRF forecast after data assimilation for most model state variables and heights analyzed. However, the filter is not necessarily optimized and non-trivial future work could focus on improving the effectiveness of the EnKF updated field while simultaneously decreasing (increasing) the consistency ratio values for WNDS aloft (all surface observations).

c. Comparing the EnKF_control, EnKF_multiPBL, and EnKF_parsest runs

The background TEMP ME time series of EnKF_control, EnKF_multiPBL, EnKF_parsest, and RUC analysis are presented for each observation type (Fig. 5.14). X's in the top of each figure denote FWI values greater than zero; with a small, medium, and large X indicating a FWI of 1, 2, and 3, respectively. Note that all EnKF simulations are from a 6-hour WRF forecast (i.e. the background), while the RUC analysis have already assimilated all available observations. Therefore, the RUC analysis is expected to have lower bias and error than the background fields from all EnKF simulations. In general, the RUC and all EnKF simulations are well correlated with each other for all observations, with the EnKF simulations exhibiting greater variability from a mean zero bias. However, there are critical differences between the EnKF analyses, which are well demonstrated by discussing the MESONT ME time series (Fig. 5.14b). Firstly, the beginning of the WRF-EnKF simulation (01-10 April 2012) features almost continuous FWDs. During this time, there is an apparent diurnal cycle in model bias using MESONT with a negative peak in the afternoon for all forecasts similar to Fig. 3.7. Interestingly, the RUC analysis has greater negative biases in the morning when verified against some surface datasets, suggesting the RUC data assimilation is less effective at removing the negative bias in the morning compared to the afternoon. EnKF_parsest has the smallest ME of all EnKF runs (by 0.46 K over EnKF_control), suggesting that parameter estimation may be improving the simulation. However, EnKF_parsest appears to degrade model bias (either positively or negatively) on non-FWDs using MESONT, such as on 15 April 2012, 18 April 2012, and from 22-28 April 2012. Interestingly, 22 April 2012 features a general change in average model bias sign from negative (EnKF_control ME = -0.93 K) to positive (EnKF_control ME = 0.31 K) for both MESONT and METAR datasets. This coincides with a change in the weather pattern from frequent FWDs to occasional rain and increased general cloudiness. The impact of FWDs and the sign flip in model bias around 22 April 2012 is not apparent aloft (ACARS, ACARSP and POES) or over the ocean (MARINE).

Two surface assimilation cycles are examined that both feature large increments (update minus background) and large differences between EnKF_parsest and EnKF_control as noted in the black arrows of Fig. 5.14; 1800 UTC 09 April 2012 (Fig. 5.15) and 0600 UTC 26 April 2012 (Fig. 5.16). 1800 UTC 09 April 2012 is noteworthy for the 1000 + acre Ridge-Manorville fire that developed during the afternoon on eastern Long Island in New York. Using North American Regional Reanalysis (NARR) data, a cold frontal passage occurred on Long Island in New York around 0000 UTC 09 April 2012, advecting a dry modified continental polar air mass into the region (not shown). During the afternoon of 09 April 2012, ideal fire weather conditions occurred at Brookhaven Airport (KHWV; approximately 5 km from the fire location) with a recorded maximum WNDS of 19.2 m s^{-1}, a minimum RELH of 15%, and a maximum TMPD of 17.8^{o}C under partly to mostly sunny skies.

The 1800 UTC 09 April 2012 assimilation cycle is examined using 2-m TEMP, 10-m WNDS, and SLP for the background (a-c), update (d-f) and increment (g-i) fields. The EnKF_control background field (Fig. 5.15a) has a strong negative bias compared to the update (Fig. 5.15d), which exceeds 5^{o}C over the Catskills of New York as shown in the increment (Fig. 5.15g). The EnKF_multiPBL run reduces this bias in the increment by up to 2^{o}C over the Catskills but is considerably less effective over Long Island and the southern half of New Jersey (Fig. 5.15h). The reduction of the increment from EnKF_multiPBL could be caused by a cancellation of errors resulting from taking the mean over three unique PBL scheme simulations. However, the individual impacts from each PBL scheme require further examination in the future. The greatest reduction of mean increment comes from EnKF_parsest, which practically eliminates the cool bias in the background from New Jersey to Massachusetts (Fig. 5.15i).

According to the NARR, the 26 April 2012 (Fig. 5.16) case features a weak low pressure to the northwest of the New York City region, broad easterly to southeasterly flow, and light precipitation across the region. For instance, KHWV received 2 mm of precipitation, a minimum daily RELH of 62%, a maximum daily TEMP of 15.6^{o}C, and a maximum WNDS of 8.9 m s^{-1}. The background EnKF_control (Fig. 5.16 a) shows the broad southeast flow over the region and the presence of an urban heat island over New York City and Philadelphia. The update is considerably cooler (Fig. 5.16 d), particularly from eastern Pennsylvania to New Jersey, which might be the result of stronger northerly winds in the increment (Fig. 5.16 g). The EnKF_multiPBL (Fig. 5.16 b,e,h) run is largely similar to the EnKF_control run. Unlike the 0600 UTC 09 April 2012 case, the EnKF_pareest case degrades the background (Fig. 5.16 c) forecast compared to the control by overwarming a larger area (notably Southern New England) and veering the low level winds toward a more southerly direction than EnKF_control.

Disturbingly, many observations are rejected in the case of EnKF_parsest, resulting in a sub-optimal analysis for this update cycle. The most striking deficiencies are with temperature over New Jersey (biases up to -10 K) and wind direction over Long Island and southern New England (south-southeast rather than east-northeast). There is a careful balance in the EnKF between observational error and ensemble spread, and the 0600 UTC 26 April 2012 treads dangerously close to filter divergence due to a large number of rejected observations (about 10%). There are similar but less severe cases of surface observations being rejected on 23 April 2012 and 27 April 2012 as shown in the large ME plots of Fig. 5.14 a,b. This problem could potentially be alleviated by assimilating observations more frequently or forcing the EnKF to assimilate all observations. The latter option is not ideal, suggesting that EnKF_parsest WRF simulations should be treated with caution depending on the synoptic flow pattern. The rejection of observations in data assimilation schemes would likely result from a combination of ensemble underdispersion and rapid model error growth. Therefore, most modern reanalysis products would not commonly suffer from such deficiencies but they are more likely to occur on FWDs than non-FWDs. The 09 April 2012 and 26 April 2012 cases will be reexamined while taking into consideration the parameter variability of SSPE in section 5.3d.

From the 09 April 2012 and 26 April 2012 cases, caution must be exercised with EnKF_parsest, since it does not always result in an improved analysis over EnKF_control. The performance of EnKF_parsest may depend on the synoptic variability. Given the large number of days where the FWI is greater than zero in April of 2012, an FWD for this chapter is defined as having an FWI equal to three. Furthermore, since TEMP model bias is maximized during the afternoon, only the 18 UTC and 00 UTC data assimilation cycles are presented hereafter.

TEMP bulk ME by observation type and EnKF run is presented for FWDs and non-FWDs for the background (Fig. 5.17a), update (Fig. 5.17b), and RUC (Fig. 5.17c). Background model biases are more negative in the EnKF_control on FWDs when verified against the METAR and MESONT datasets (averaging -0.89 K) and observations aloft (averaging -0.37 K). EnKF_multiPBL has a slightly greater background negative bias on FWDs (by -0.19 K) than EnKF_control. Most importantly, EnKF_parsest significantly reduces the background surface land cool bias on FWDs compared to EnKF_control (by 0.58 K). EnKF_parsest also helps to reduce the land based cool bias on non-FWDs (by 0.35 K). The EnKF update effectively removes most of the TEMP bias (Fig. 5.17b) from the background regardless of the simulation, with lingering biases similar to that of the RUC analysis (Fig. 5.17c).

Bulk MAE by observation type and EnKF run is also presented for the background, update, and the RUC analysis (Fig. 5.18). In general, the EnKF simulations with higher ME from Figure 5.18 exhibit higher MAE in Fig. 5.19. The surface MAE over land on FWDs for EnKF_control is about 0.38 K higher than on non-FWDs. Likewise, MAE for both FWDs and non-FWDs are very similar between the EnKF_control and EnKF_multiPBL, suggesting that there is little benefit on the average with running a multi-PBL ensemble. Nonetheless, the biases and error for all members using a unique PBL scheme should be examined. It is possible that a multi-PBL EnKF could be optimized by choosing different PBL parameterizations. EnKF_paresest reduces surface MAE over land on FWDs (by 0.32 K), but increases MAE on non-FWDs (by 0.35 K). This further suggests that the performance of EnKF_parsest may vary with the synoptic variability. These results are consistent with the two case studies in Fig. 5.15 and Fig. 5.16, although 09 April 2012 and 26 April 2012 represent extreme examples from the EnKF_parsest run. Consistent with Figure 5.17, all EnKF simulations effectively reduce MAE in the update (Fig. 5.18b), which is comparable to the RUC analysis (Fig. 5.18c). For some of the datasets with less observations (MARINE and NEPP), there is a slightly higher MAE on FWDs in the update compared to non-FWDs (difference averaging 0.34 K).

Spatial ME of the EnKF runs are compared to MESONT TEMP and averaged on a 0.2^{o} latitude by 0.2^{o} longitude grid in Figure 5.19. In order to analyze the spatial variability over a larger area, the domain size is increased for this figure only to encompass 39^{o}N to 42^{o}N and 78^{o}W to 70^{o}W. The background spatial ME on non-FWDs are generally similar in all EnKF simulations (mostly ranging between -2.5 K and 1.2 K), although there are small isolated regions of up to 1.4 K positive bias. On FWDs, the EnKF_control run has a large region of negative bias (ranging between -4.4 K and 0.3 K) that is somewhat improved in EnKF_parsest (ranging between -4.3 K and 0.8 K). The largest improvements to the negative bias on FWDs for EnKF_parsest are over land and in the corridor between Philadelphia and New York City. On the other hand, most regions with a cool bias exceeding -4 K in the EnKF_parsest run are just offshore of the New Jersey coast (some stations just offshore of the coast are lumped in with the MADIS dataset).

The vertical structure of TEMP ME and MAE for the EnKF runs is analyzed vertically using the POES, ACARS, and ACARSP datasets (Fig. 5.20). The POES behaves differently than the ACARS observations (Fig. 5.20a,d) and should be used with caution. For instance, there is a large negative bias when comparing POES to the RUC analysis below 700 hPa averaging about 1.2 K. ACARS and ACARSP background ME (Fig. 5.20b) is considerably lower on FWDs for the EnKF_control and EnKF_multiPBL below 700 hPa compared to non-FWDs, which is consistent with Fig. 4.13. The EnKF_parsest reduces this bias below 800 hPa (to -0.22 K at 1000 hPa) on FWDs, but does not change the considerably bias above 800 hPa. As a result, the EnKF_parsest reduces ACARS and ACARSP MAE compared to EnKF_control below 750 hPa (by 0.22 K at 1000 hPa), such that it is slightly better than the MAE on non-FWDs. The EnKF_multiPBL increases the negative cool bias and MAE below 800 hPa compared to the EnKF_control.

Similar to TEMP, bulk ME for SPHU is analyzed by observation type for the background, update, and RUC analysis (Fig. 5.21). The total number of SPHU observations available throughout the April of 2012 simulation are more limited than for TEMP and WNDS and total 3492, 396, 399, 261, 3042, 351, and 19953 for POES, ACARS, ACARSP, MARINE, METAR, CRN, and MESONT, respectively. EnKF_control exhibits a greater positive SPHU bias on FWDs compared to non-FWDs (by 0.3 g kg^{-1}) when using METAR observations. The EnKF_multiPBL simulation is very similar to the EnKF_control run. However, the EnKF_parsest has a reduced positive moisture bias with METAR compared to EnKF_control on FWDs (by 0.19 g kg^{-1}) and an increased positive moisture bias on non-FWDs (0.08 g kg^{-1}). METAR observations are likely the most trustworthy of the observation types since the station quality is regulated with an abundant amount of observations. Verification with other observation types for SPHU is not consistent, with MESONET and MARINE observations showing a general positive (negative) bias on FWDs (non-FWDs). Verification against ACARS and ACARSP observations show a consistent but slight positive bias on FWDs and non-FWDS, while using POES observations results in a strong negative bias that should be treated with caution. One consistent result with SPHU is that biases are generally reduced but not removed in the updated field, with the updated field less biased than the RUC analysis.

The vertical impact of the EnKF runs is analyzed for WNDS in Figure 5.22. In general, there is little difference in terms of ME and MAE between the EnKF_control and EnKF_multiPBL on FWDs and non-FWDs. One minor difference is that FWDs exhibit a negative WNDS bias at 700 hPa compared to non-FWDs (difference averaging 1.78 m s^{-1} in EnKF_control). This might be the result of model simulations underestimating the PBL depth on FWDs. Similarly, EnKF_parsest on FWDs have very similar ME and MAE compared to EnKF_control and EnKF_multiPBL. However, EnKF_parsest has a large positive WNDS bias on non-FWDs throughout the entire troposphere (averaging 2.40 m s^{-1}), which increases the MAE (by 1.78 m s^{-1}) compared to EnKF_control. This is clearly not a desirable impact from SSPE, although it is not surprising that substantial improvements to some aspects of the forecast can also have a negative impact on other variables at different times.

d. Effect of SSPE on the EnKF

As mentioned earlier, the motivation for running EnKF_parsest is to look for structural model error in the ACM2 PBL on FWDs. An additional secondary benefit would come from SSPE improving the overall WRF simulation. From the results in section 5.3c, the effect of SSPE on the WRF simulations is mixed.

Determining the cause and effect of SSPE within the EnKF_parsest run is difficult given the complexities of the WRF model and the ACM2 PBL scheme within WRF. However, some insight can be gained by analyzing the parameter variability for EnKF_parsest throughout April of 2012. The time series for PVAR and KVVAR for the month long run is shown in Figure 5.23. PVAR starts near its default value of 2 before quickly dropping on 0000 UTC 02 April 2012 and remaining between 0.5 and 1.5 for the remainder of the simulation. Nielson-Gammon et al (2010) determined PVAR to be the most sensitive daytime parameter, which exhibits negative correlations with PBL WNDS and TEMP. This can be understood visually in Figure 5.24, which shows how PVAR partially controls the shape and magnitude of eddy diffusivity in the PBL. With the PVAR default value of 2, eddy diffusivity is maximized in the middle of the PBL. A decrease of PVAR results in an increase of eddy diffusivity that is more focused toward the top of the PBL. Note that PVAR mostly influences the daytime eddy diffusivity in the ACM2 PBL (Nielson-Gammon et al. 2010). The second parameter, KVVAR, has a default value of 0.01 in the ACM2 PBL scheme and remains near that default value in SSPE until about 12 April 2012 (Fig. 5.23b). Thereafter, KVVAR steadily increases to about 0.04 and oscillates between 0.03 and 0.04 from 21 April 2012 to 1800 UTC 30 April. KVVAR is proportional to the minimum eddy diffusivity as a function of layer thickness. Nielson-Gammon found KVVAR had a positive (negative) correlation with WNDS during the day (night).

Overall, it is likely that the smaller PVAR and larger KVVAR values lead towards a net effect of increasing eddy diffusivity in the PBL particularly towards the top, which allows for a deeper, drier, warmer and well mixed PBL on FWDs. Additional research is needed to confirm this hypothesis, but the reduction in TEMP (Fig. 5.20) and SPHU (Fig. 5.21) biases on FWDs support this idea. Due to a change in the synoptic weather pattern from FWDs to occasional storminess, it is possible that the parameter values are not optimal between 22-28 April 2012. Although the parameter error variance is always inflated (section 5.2d) after each DA cycle, it is possible that the parameter values could not adjust rapidly enough starting on 22 April 2012.

The potential impact of SSPE on the two case studies from 1800 UTC 09 April 2012 (Fig. 5.15) and 0600 UTC 26 April 2012 (Fig. 5.16) are reexamined. For 1800 UTC 09 April 2012, EnKF_parsest had a very beneficial impact by reducing the TEMP and SPHU (not shown) biases that developed on an extreme FWD. At this time, PVAR (KVVAR) has a value of 0.55 (0.01). Since KVVAR is estimated at its default value, PVAR is the only parameter that could change the WRF simulation. Since the PVAR value is significantly reduced from 2 to 0.5 (cyan line in Fig. 5.24), the magnitude of eddy diffusivity increased at a location higher in the PBL.

On 0600 UTC 26 April 2012, PVAR (KVVAR) had a value of 1.07 (0.03). However, Nielson-Gammon et al. (2010) discuss that PVAR has a reduced impact on WRF simulations at night. Therefore, the greater value of KVVAR likely degrades the EnKF_parsest forecast compared to EnKF_control. In this case, it is possible that spuriously high eddy diffusivity contributed to the poor forecast. During this DA cycle, there is a weak surface easterly wind with stronger southerly winds aloft. Overmixing could explain the spurious stronger southerly winds being mixed down from aloft, which might overcome the marine layer and overwarm Long Island, New York in EnKF_parsest compared to EnKF_control.

If structural model errors exist in the ACM2 PBL, parameter values may adjust rapidly to compensate for these errors during certain synoptic flow patterns. As a result, PVAR and KVVAR may vary with FWI index or model biases. Therefore, the relationship between FWI value and parameter estimate (error bars denote 2.5th and 97.5th percentile from the total dataset) is shown in Figure 5.25. There are no statistically significant variations in PVAR or KVVAR with FWI index, however there appears to be a slight sensitivity between lower KVVAR and higher FWI value. This makes sense given that greater eddy diffusivity would occur with a higher FWI value.

Spatially averaged TEMP model bias verified against ACARS and MARINE data from the background EnKF_control is regressed on the EnKF_parseset parameter values for PVAR and KVVAR (Fig. 5.26). Given the potential diurnal sensitivity of WRF simulations to PVAR and KVVAR and the diurnal variability in model bias (Fig. 3.5), regressions are performed for each time of day separately. The relationship between parameter value and model bias is statistically significant for KVVAR (KVVAR) using ACARS (MARINE) data at 00 UTC (00 UTC). With KVVAR, larger parameter values result when the control run TEMP biases are more negative in the MARINE dataset. Therefore, negative biases occur in EnKF_control when KVVAR favors increased eddy diffusivity in EnKF_parsest. However, this conflicts with the positive correlation between KVVAR and ME using ACARS.

5.4. Discussion

An Ensemble Kalman Filter (EnKF) coupled with the Weather Research and Forecast (WRF) model is used to explore the impact of continuous cycling data assimilation (DA) on fire weather days (FWDs) throughout April of 2012. The EnKF assimilates a variety of Meteorological Assimilation Data Ingest System (MADIS) datasets from the surface [buoys ships, mesonet stations, Automated Surface Observing System (ASOS), Standard Aviation Observation (SAO)] and aloft [profilers, rawinsonde soundings, satellite winds, aircraft communication addressing and reporting system (ACARS)].

To insure the EnKF is performing well on FWDs, sensitivity runs are conducted between 06-11 April 2012 while adjusting the localization and observational error variance. The EnKF performs well regardless of sensitivity run aloft for temperature (TEMP), specific humidity (SPHU), and wind speed (WNDS). However, EnKF performance near the surface benefits from an increased radius of influence and a halving of the observational error determined from the MADIS dataset. Thereafter, three month long EnKF runs are conducted; a control run (EnKF_control), a multi-physics run using the Mellor-Yamada Nakanishi and Niino (MYNN) planetary boundary layer (PBL), the Yonesei University (YSU) PBL, and the ACM2 PBL (EnKF_multiPBL), and a third run performing simultaneous state and parameter estimation (SSPE) of PVAR and KVVAR with the ACM2 PBL scheme (EnKF_parsest).

In general, cool and wet PBL model biases (similar to results from chapters 3 and 4) develop within 6 hours of the WRF simulation on FWDs in the lower troposphere. The EnKF is effective at removing this bias for TEMP, SPHU and WNDS. The background field of EnKF_multiPBL exhibits a slightly greater cool and wet bias than EnKF_control, suggesting that the additional PBL schemes degrade the quality of the forecast. EnKF_parsest removes about 60% (90%) of the low-level cool TEMP (SPHU) bias from EnKF_control on FWDs. However, EnKF_parsest sometimes produces background fields with greater MAE than EnKF_control on non-FWDs. This is particularly true for WNDS, where a large positive bias (1.53 m s^{-1}) is present on non-FWDs within EnKF_parsest. In general, EnKF_parsest improves short-term WRF simulations on FWDs but degrades WRF performance on non-FWDs. Therefore, caution is emphasized when using parameter estimation to improve WRF simulations, since model improvement is likely conditional on the synoptic flow pattern. This is caused by structural model error within the PBL scheme or parameter estimation attempting to correct for deficiencies elsewhere in the parameterized physics of the WRF.

The purpose of running EnKF_parsest is not to create an optimal WRF simulation for FWDs, but rather explore the structural model errors within the ACM2 PBL. In reality, PVAR and KVVAR should not vary substantially from one day to the next. It is likely that during an FWD, the parameter values change to spuriously overmix the PBL, resulting in a reduction of

SPHU and TEMP biases. When the weather pattern changes, the PBL is still being overmixed, resulting in greater model error even in a 6 hour WRF forecast. Therefore, the parameter values provide clues for the sources of model error within the WRF model.

The 1800 UTC 09 April 2012 and 0600 UTC 26 April 2012 data assimilation cycles provide a potentially insightful contrast into EnKF_parsest performance. EnKF_parsest should be run over a much shorter interval with a focus on both of these cases to explore differences in parameter estimation and model performance. It is possible that EnKF_parsest is sensitive to a training period determined in the beginning of the EnKF-WRF simulation. More comprehensively, the sensitivity studies of Nielson-Gammon et al. (2010) could be repeated for each case separately to check for consistency in identifiable parameters. For instance, Nielson-Gammon et al. (2010) determined their identifiable parameters used in this dissertation from a high pollution event similar to the FWD of 09 April 2012. They emphasize that identifiable parameters may vary with synoptic flow pattern, and the results from this chapter potentially reinforce this idea.

It is possible that the main error source may originate from outside of the PBL scheme entirely. For instance, if the soil moisture is initialized too high in the model, this could lead to the development of an overly shallow, cool and wet PBL. Similarly, issues with the land surface model, deviations in the timing of green-up as represented in WRF or radiation scheme could lead to similar effects. Future work is needed to explore potential error sources outside of the PBL and to carefully analyze the parameter values and their relationship with the synoptic flow pattern. Furthermore, EnKF_parsest should be rerun to estimate one parameter at a time so that their individual impacts could be analyzed separately.

Table 5.1: Total MADIS observations used to verify the EnKF trial runs between 1200 UTC 06 April 2012 and 0000 UTC 11 April 2012within 39ºN - 42ºN and 78ºW - 70ºW.

Column1	**TEMP (K)**	**WNDS (m s^{-1})**	**SPHU (g kg-1)**
ACARS	4548	4486	168
ACARSP	3477	3454	150
MARINE	73	71	41
METAR	1093	969	1084
MESONT	10929	10026	8150
HDW-E	0	318	0
CRN	38	0	38
NEPP	133	76	0
POES	3193	0	2792
MULTI-P	15	0	0

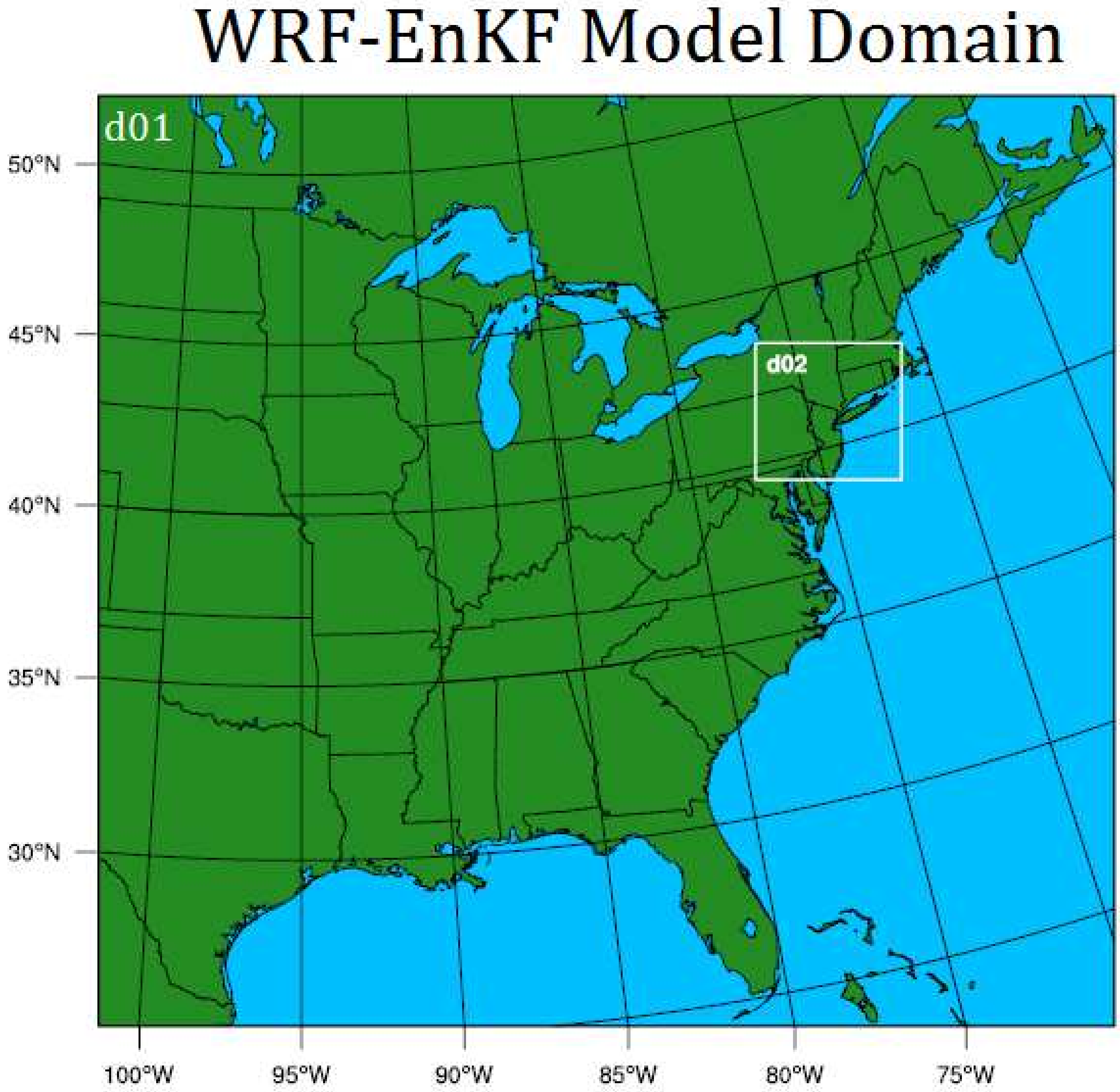

Figure 5.1: Map of WRF-EnKF model outer domain (d01) and inner domain (d02).

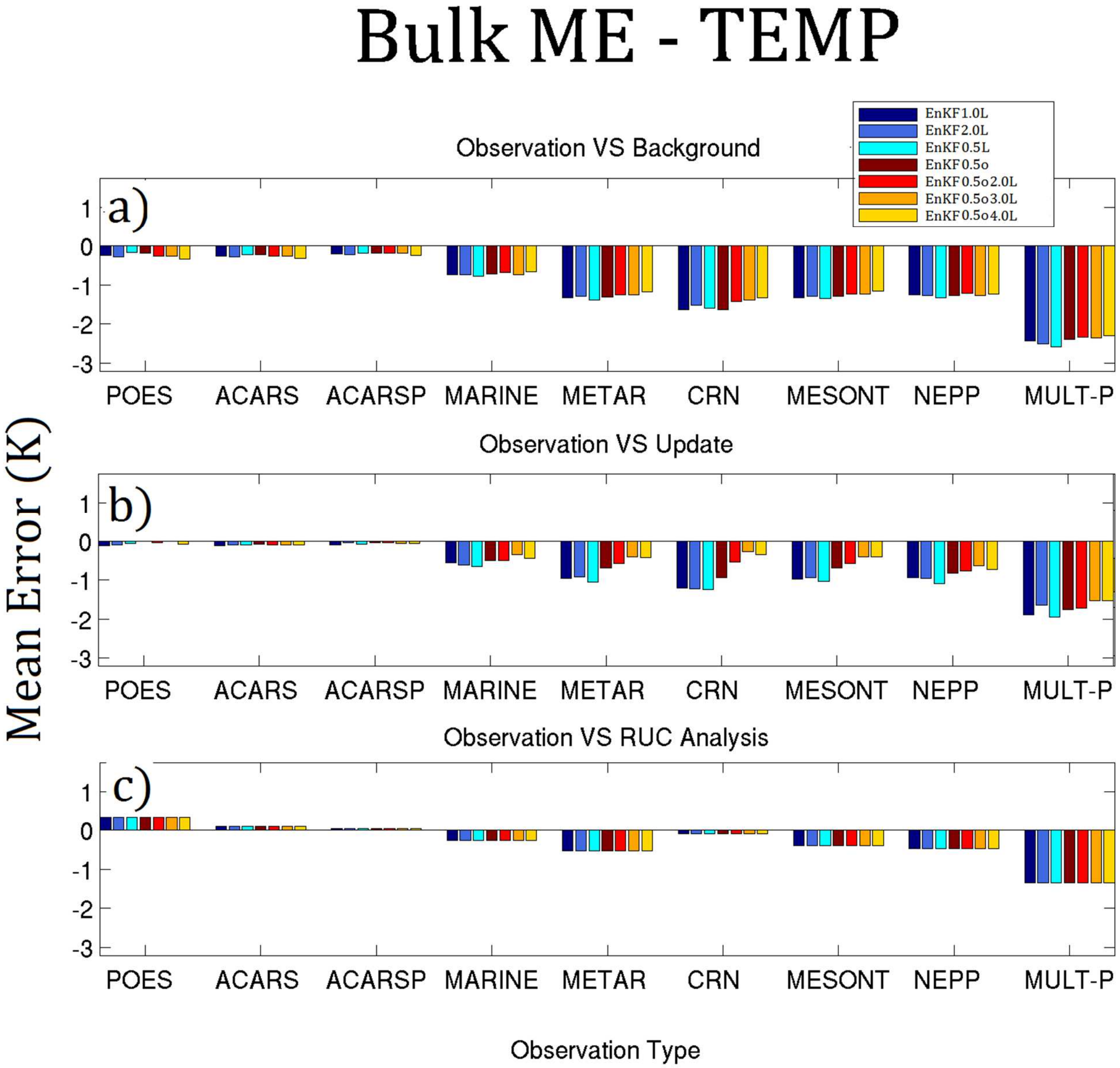

Figure 5.2: Bulk temperature (TEMP) mean error with different WRF-EnKF sensitivity runs (EnKF1.0L, EnKF0.5L, EnKF2.0L, EnKF0.5o, EnKF0.5o2.0L, EnKF0.5o3.0L, EnKF0.5o4.0L; see text) verified by MADIS observation type for EnKF background (a), EnKF update, (b) and RUC analysis (c).

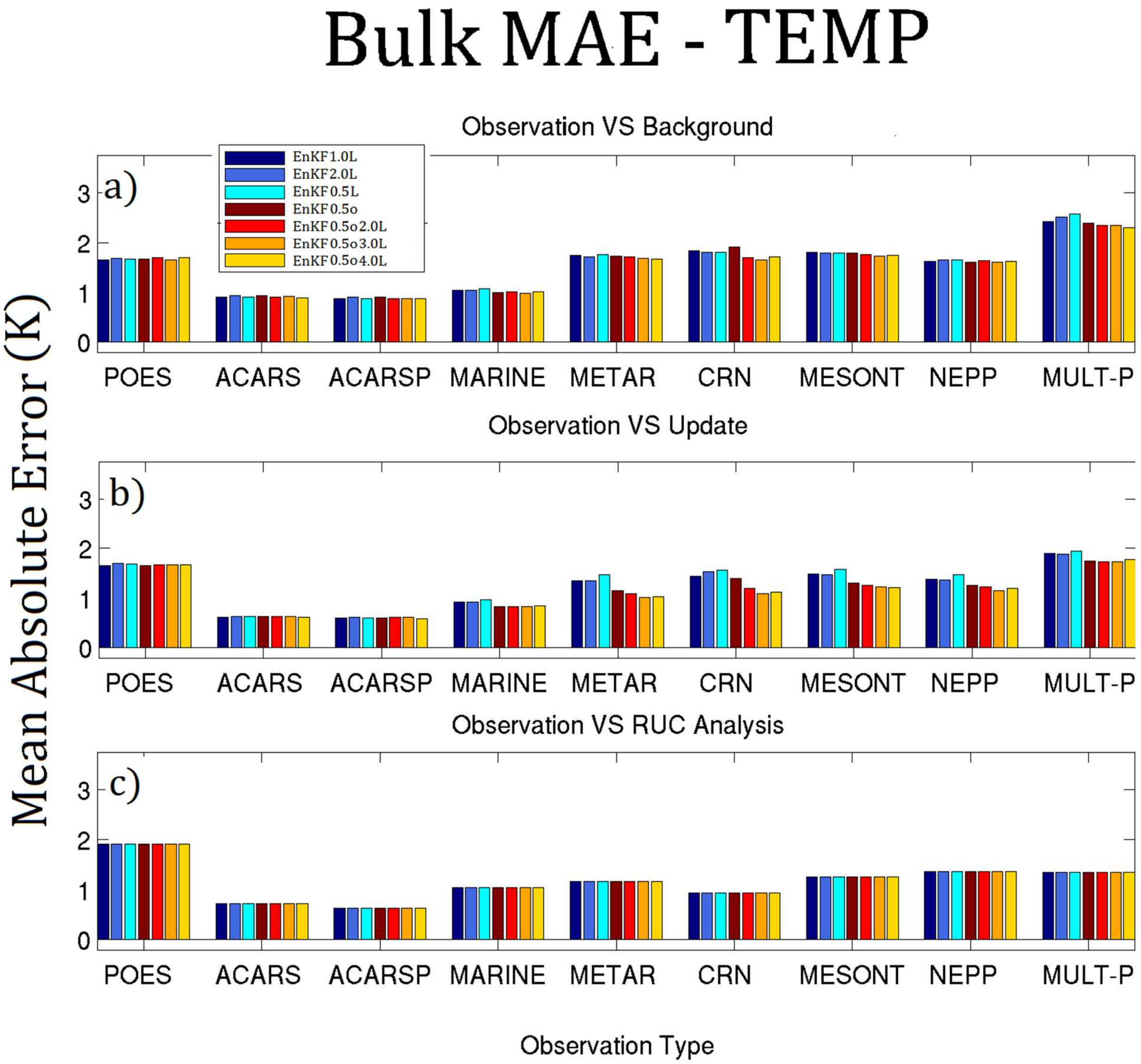

Figure 5.3: Same as figure 5.2, but for bulk MAE.

Spatial TEMP ME - MESONT Update

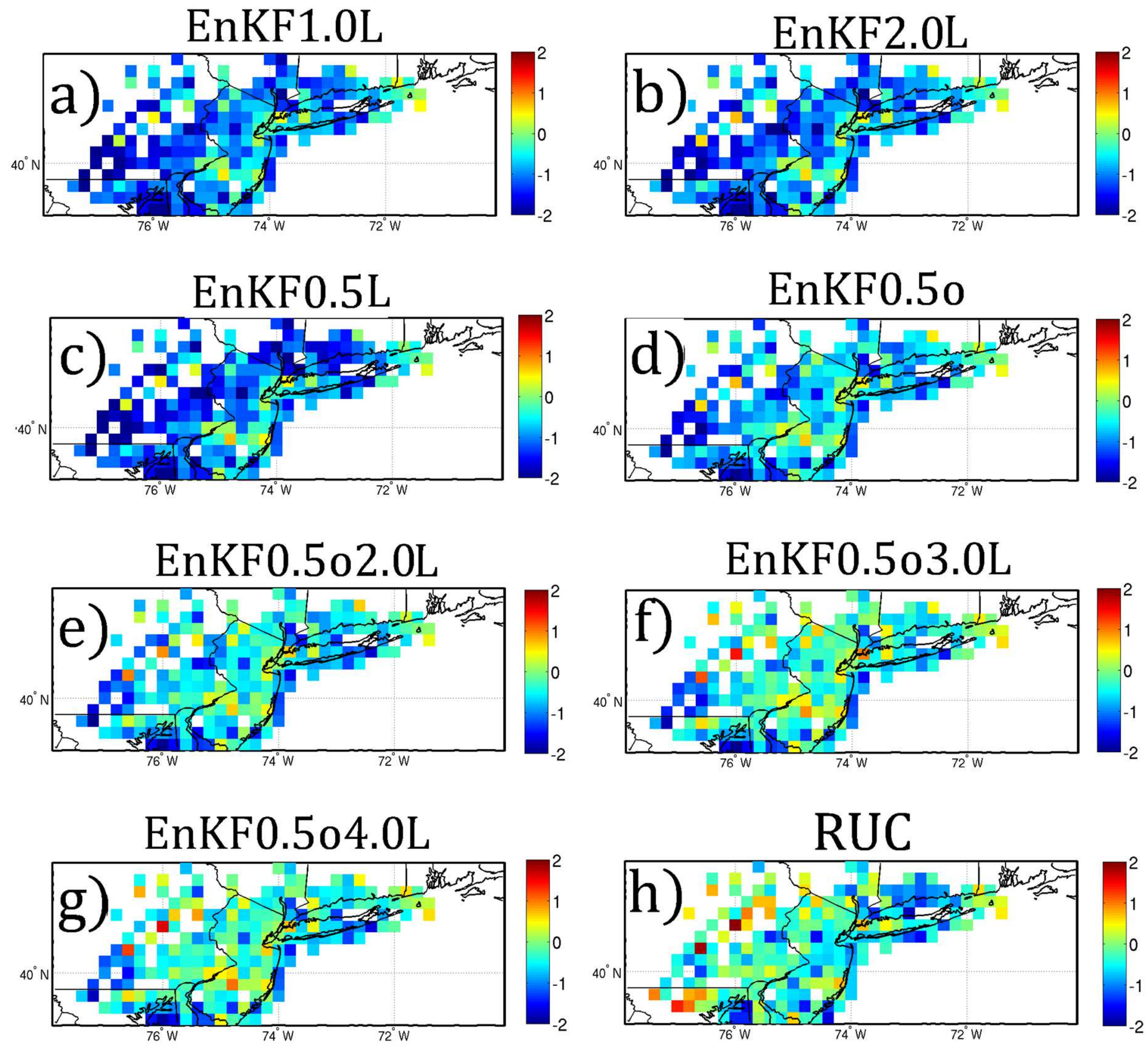

Figure 5.4: Update spatial TEMP MAE verified against MESONT observations on a 0.2° by 0.2° latitude/longitude grid for EnKF1.0L (a), EnKF2.0L (b), EnKF0.5L (c), EnKF0.5o (d), EnKF0.5o2.0L (e), EnKF0.5o3.0L (f), EnKF0.5o4.0L (g), and RUC analysis (h).

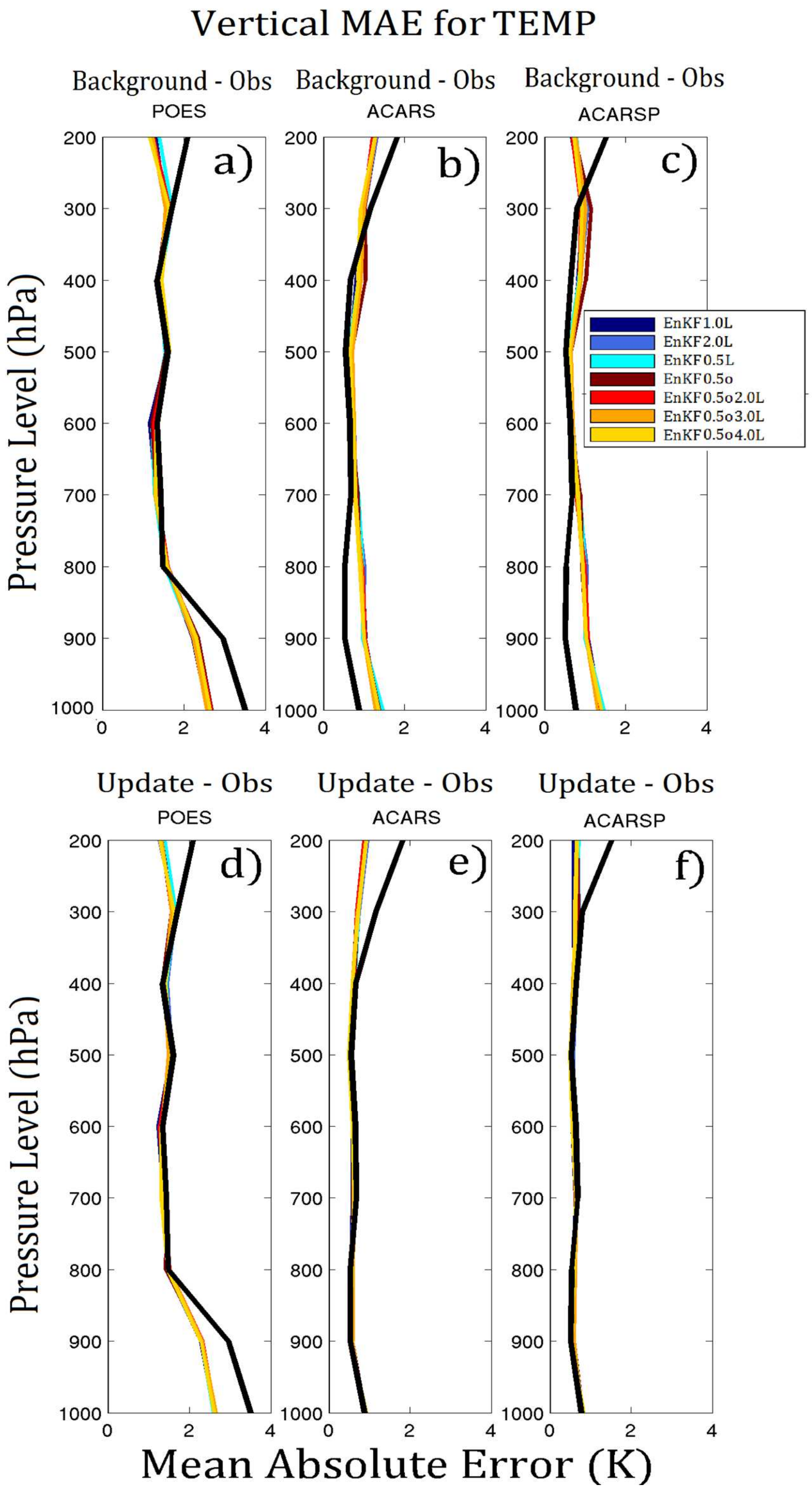

Figure 5.5: Vertical MAE verified against POES (a,d) ACARS (b,e) and ACARSP (c,f) observations over all WRF-EnKF sensitivity runs (see text) for the background state (a,b,c) and update (d,e,f). The RUC MAE verified against the observations is shown in solid black.

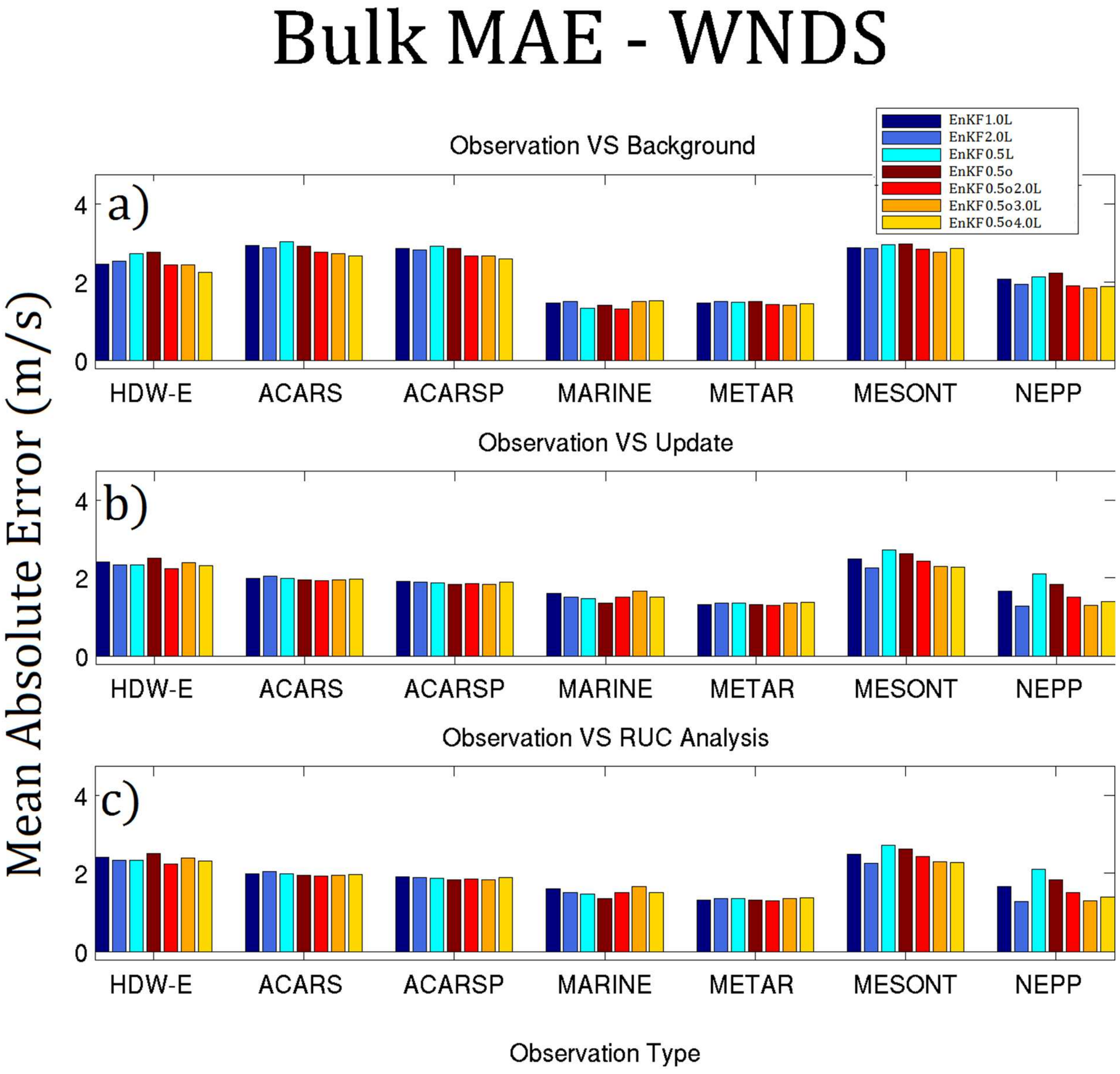

Figure 5.6: Same as figure 5.2, but for WNDS bulk MAE.

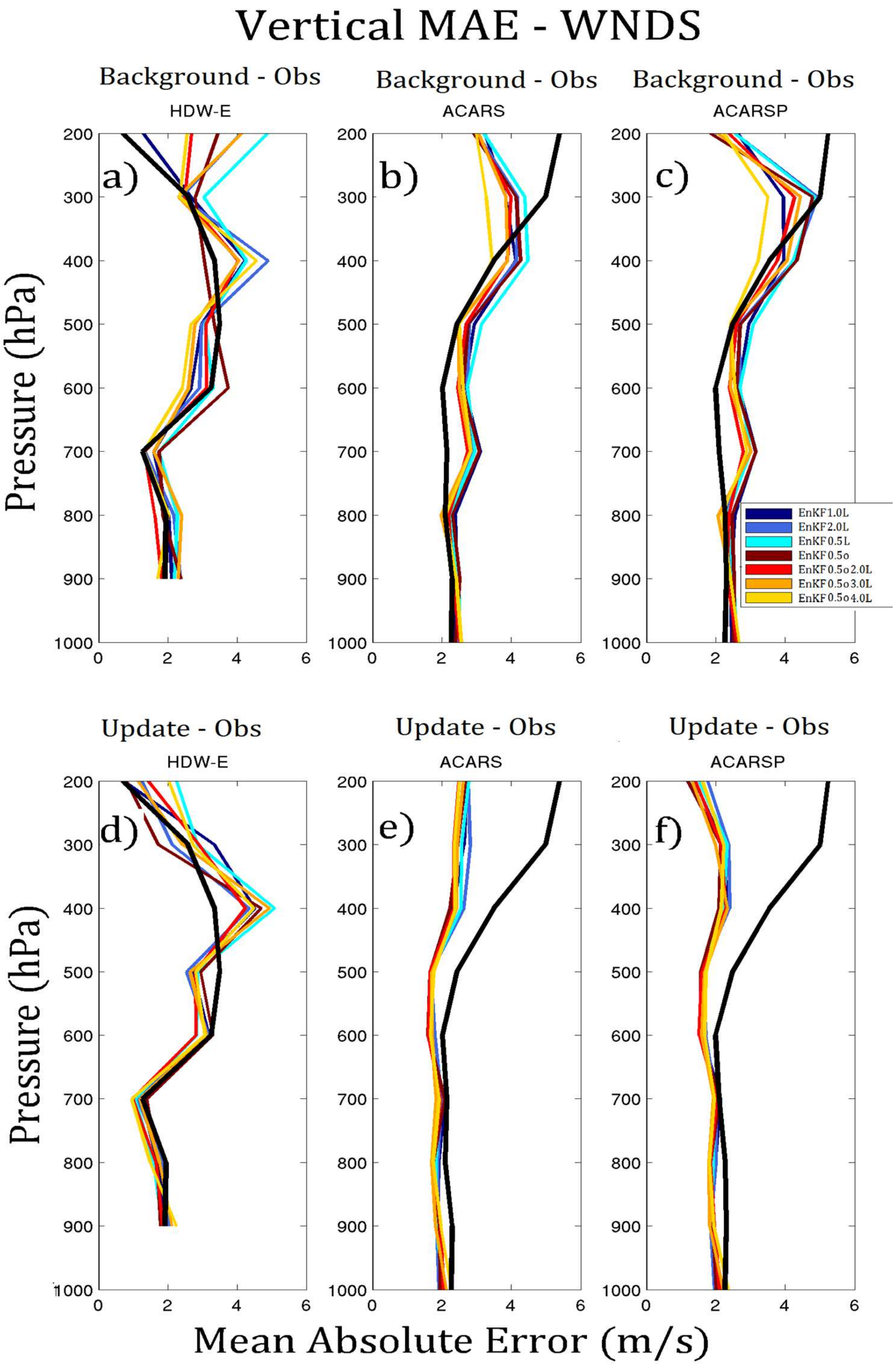

Figure 5.7: Same as figure 5.5 but for WNDS verified against MADIS HDW-E (a,d), ACARS (b,e) and ACARSP (c,f).

Bulk MAE - SPHU

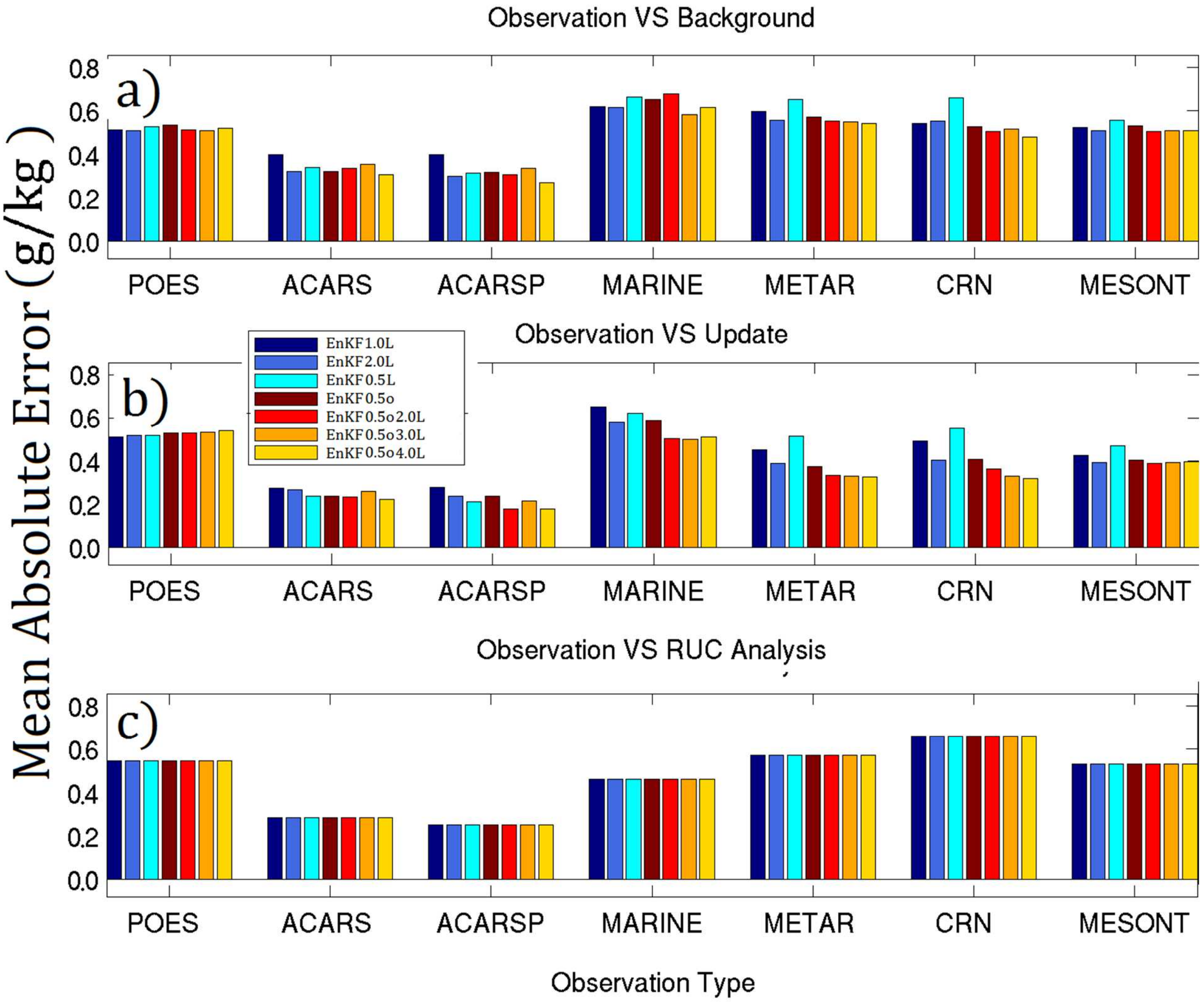

Figure 5.8: Same as figure 5.2, but for SPHU bulk MAE.

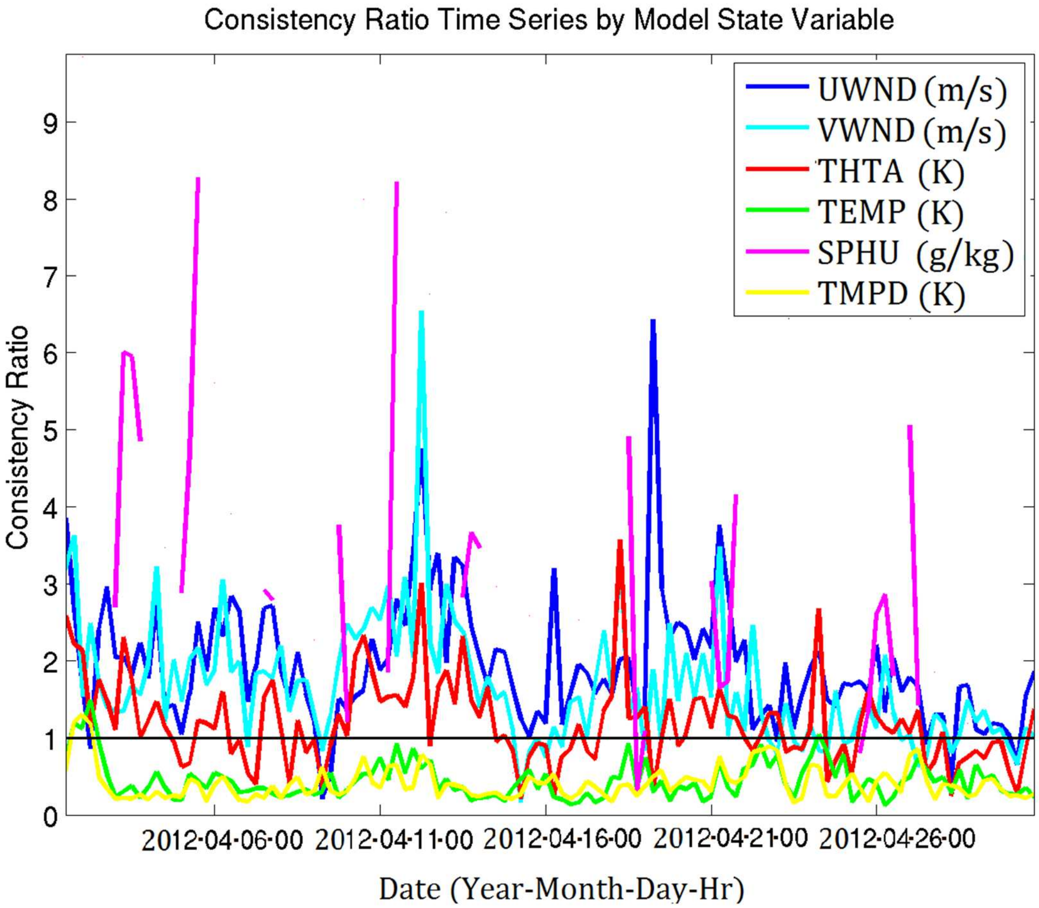

Figure 5.9: EnKF_control averaged time series of consistency ratio throughout April of 2012 for U (blue), V (cyan), T (red), K(green), Q (magenta) and D (yellow).

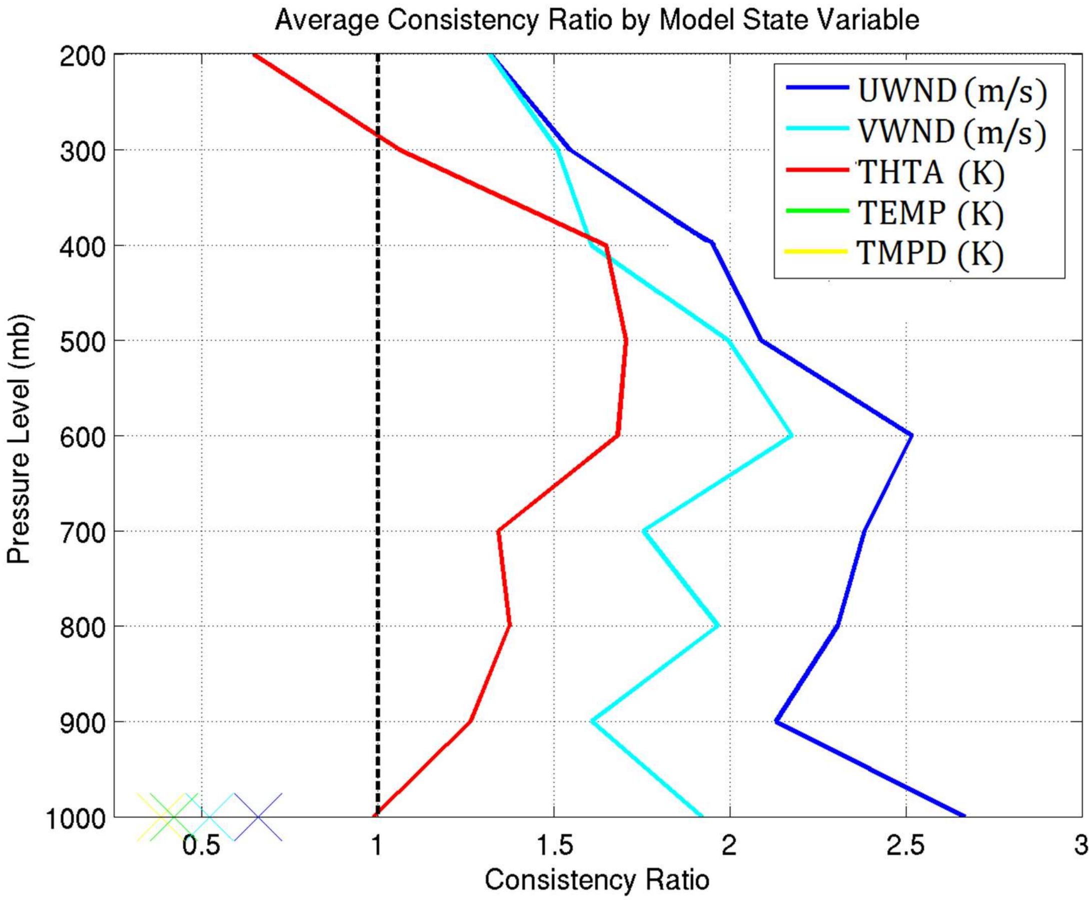

Figure 5.10: EnKF_control consistency ratio averaged by height throughout April of 2012 for UWND (blue), VWND (cyan), THTA (red), TEMP (green), and TMPD (yellow). When available, Xs denote surface values for each variable. Dashed black line denotes a consistency ratio of one.

.

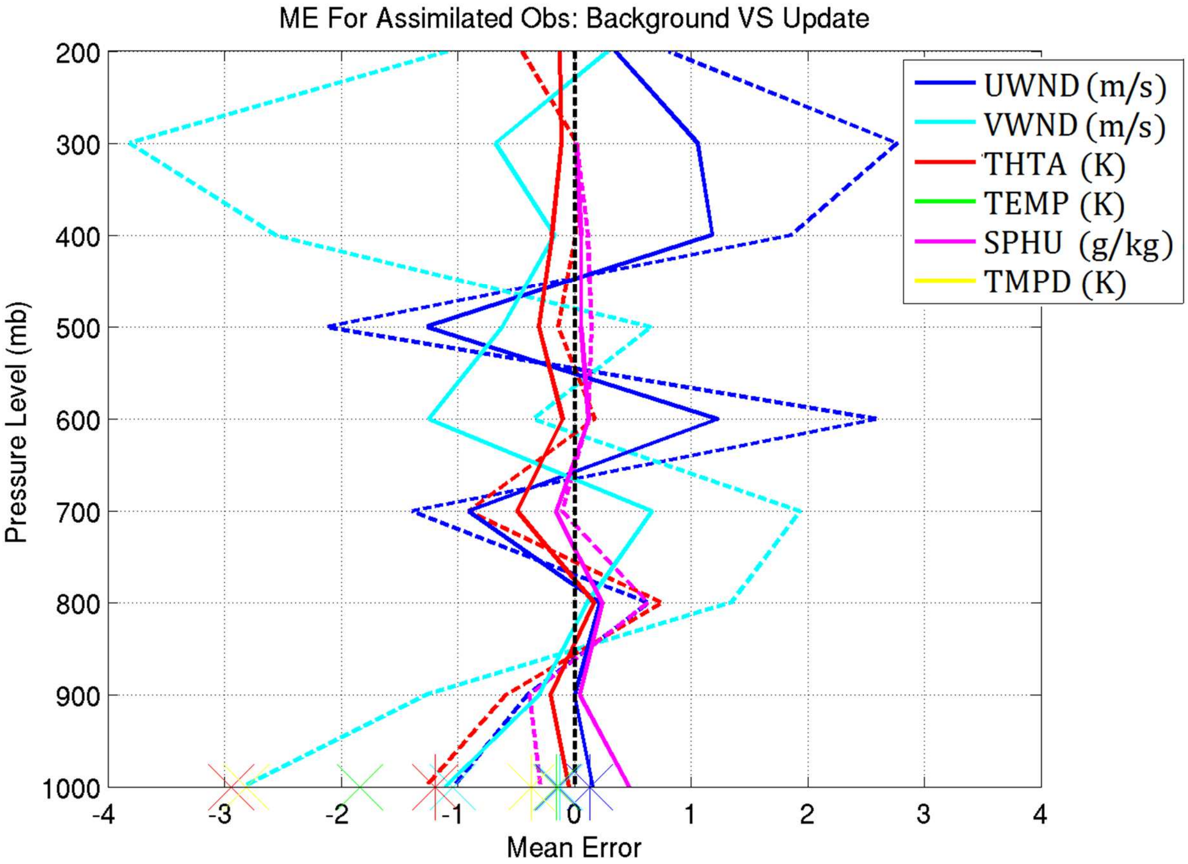

Figure 5.11: EnKF_control averaged mean error by height throughout April of 2012 for background (dashed) and update (solid) UWND (blue), VWND (cyan), THTA (red), TEMP (green), SPHU (magenta) and TMPD (yellow). When available, X's denote background surface values and +'s denote updated surface values.

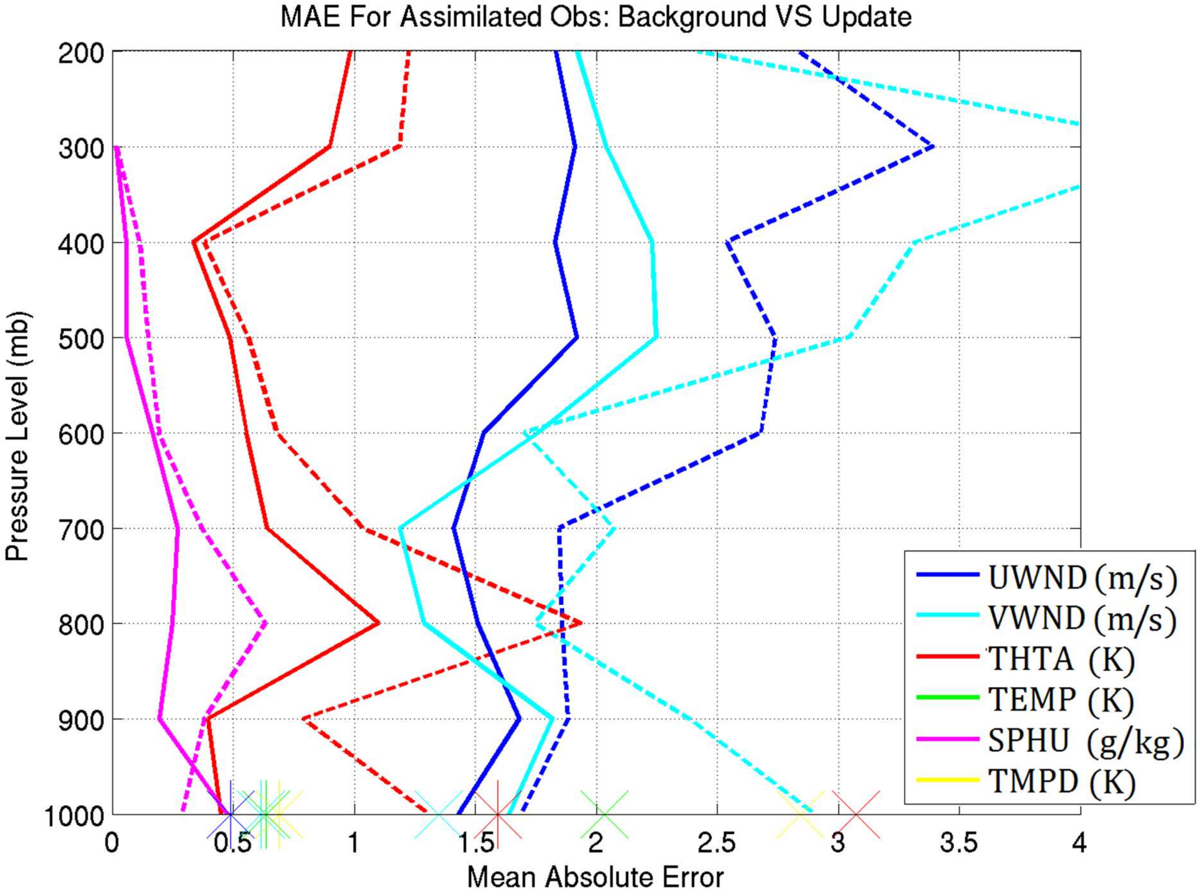

Figure 5.12: Same as figure 5.11, except averaged MAE by height.

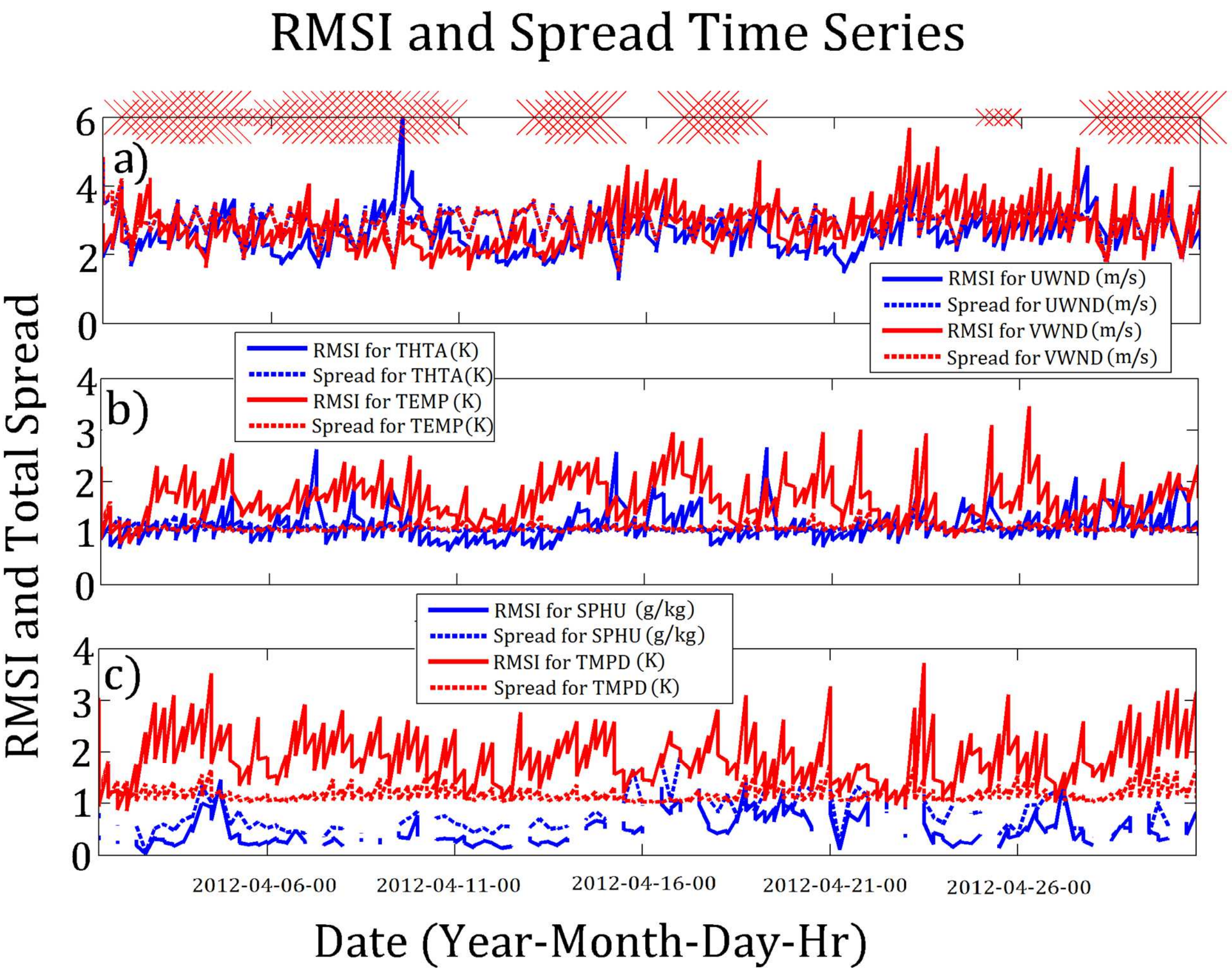

Figure 5.13: EnKF_control RMSI (solid) and spread (dashed) time series for UWND/ VWND (a), THTA/TEMP (b), and SPHU/TMPD (c). X's on top of (a) denote FWDs, with a small, medium, and large X indicating an FWI of 1, 2, and 3, respectively.

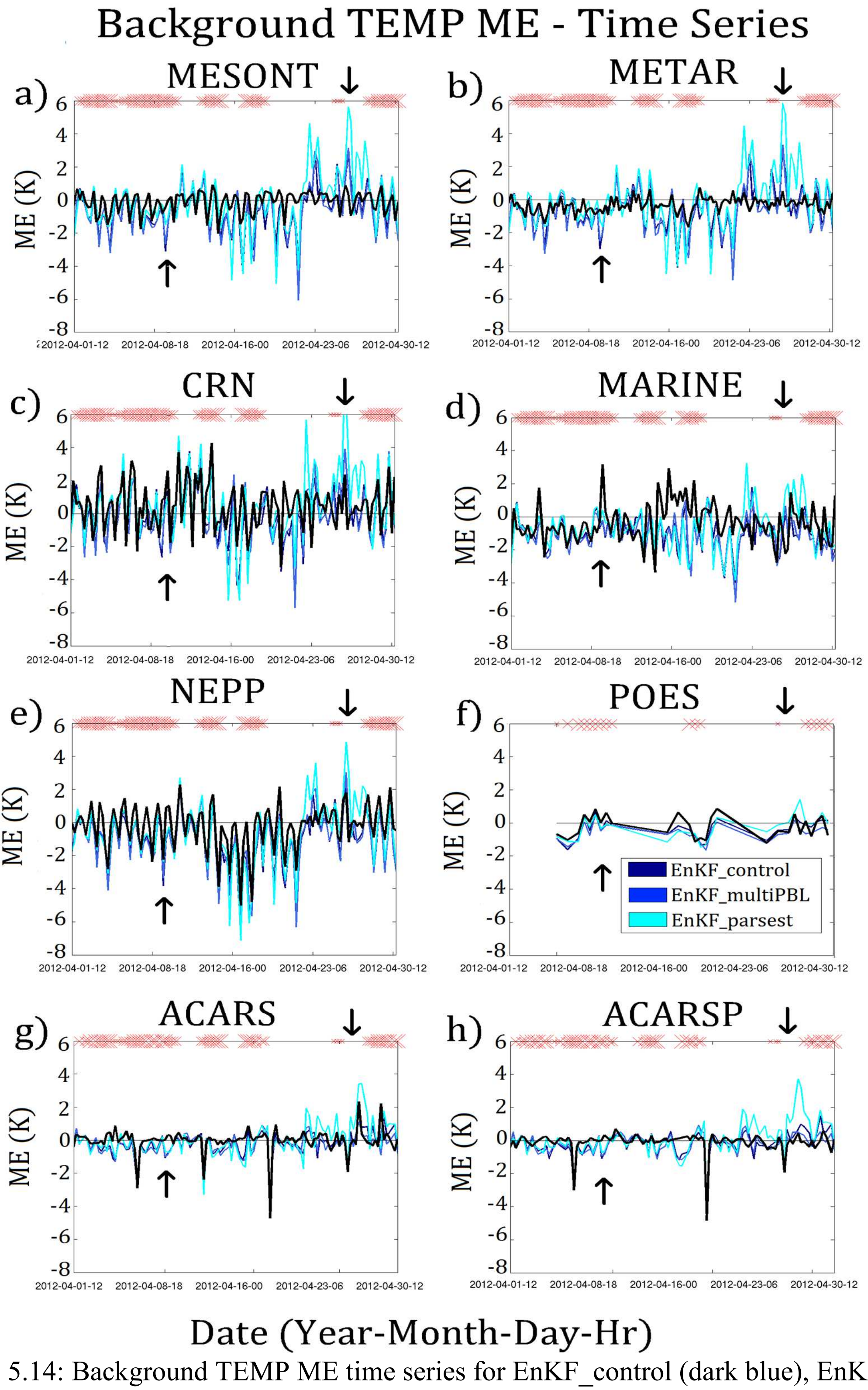

Figure 5.14: Background TEMP ME time series for EnKF_control (dark blue), EnKF_multiPBL (blue), EnKF_parsest (cyan) and RUC analysis (black) by observation type. X's denote FWDs, with a small, medium, and large X indicating a FWI of 1, 2, and 3, respectively. Arrows indicate the location of the 1800 UTC 09 April 2012 and 0600 UTC 26 April 2012 assimilations.

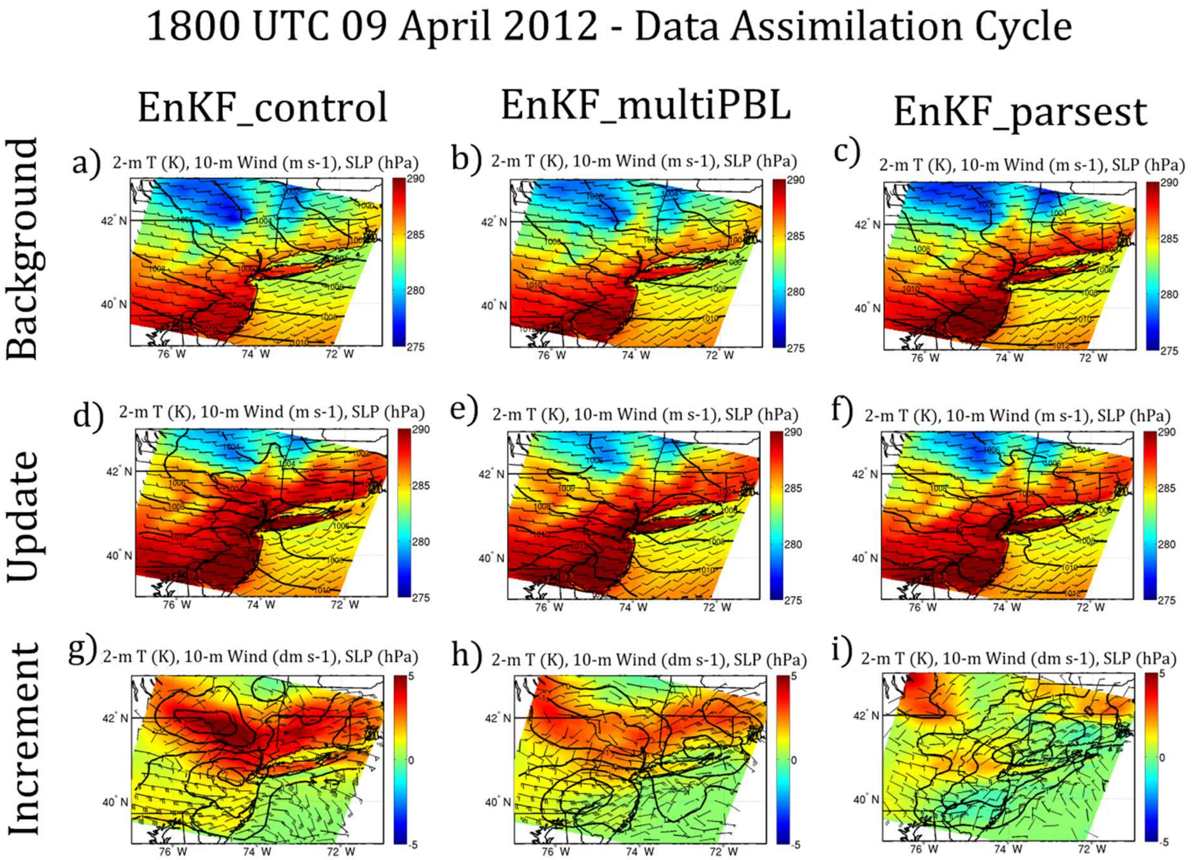
1800 UTC 09 April 2012 - Data Assimilation Cycle
EnKF_control
EnKF_multiPBL
EnKF_parsest
Background
Update
Increment
a) 2-m T (K), 10-m Wind (m s-1), SLP (hPa)
b) 2-m T (K), 10-m Wind (m s-1), SLP (hPa)
c) 2-m T (K), 10-m Wind (m s-1), SLP (hPa)
d) 2-m T (K), 10-m Wind (m s-1), SLP (hPa)
e) 2-m T (K), 10-m Wind (m s-1), SLP (hPa)
f) 2-m T (K), 10-m Wind (m s-1), SLP (hPa)
g) 2-m T (K), 10-m Wind (dm s-1), SLP (hPa)
h) 2-m T (K), 10-m Wind (dm s-1), SLP (hPa)
i) 2-m T (K), 10-m Wind (dm s-1), SLP (hPa)
42° N
40° N
76° W
74° W
72° W
290
285
280
275
5
0
-5

Figure 5.15: 2-m TEMP (K), 10-m WNDS, and SLP (hPa) background (a-c), update (d-f) and increment (g-i) fields on 1800 UTC 09 April 2012 at 18 UTC for the EnKF_control (a,d,g), EnKF_multiPBL (b,e,h) and EnKF_parsest (c,f,i). Note WNDS are in m s^{-1} for the background and update and in dm s^{-1} for the increment.

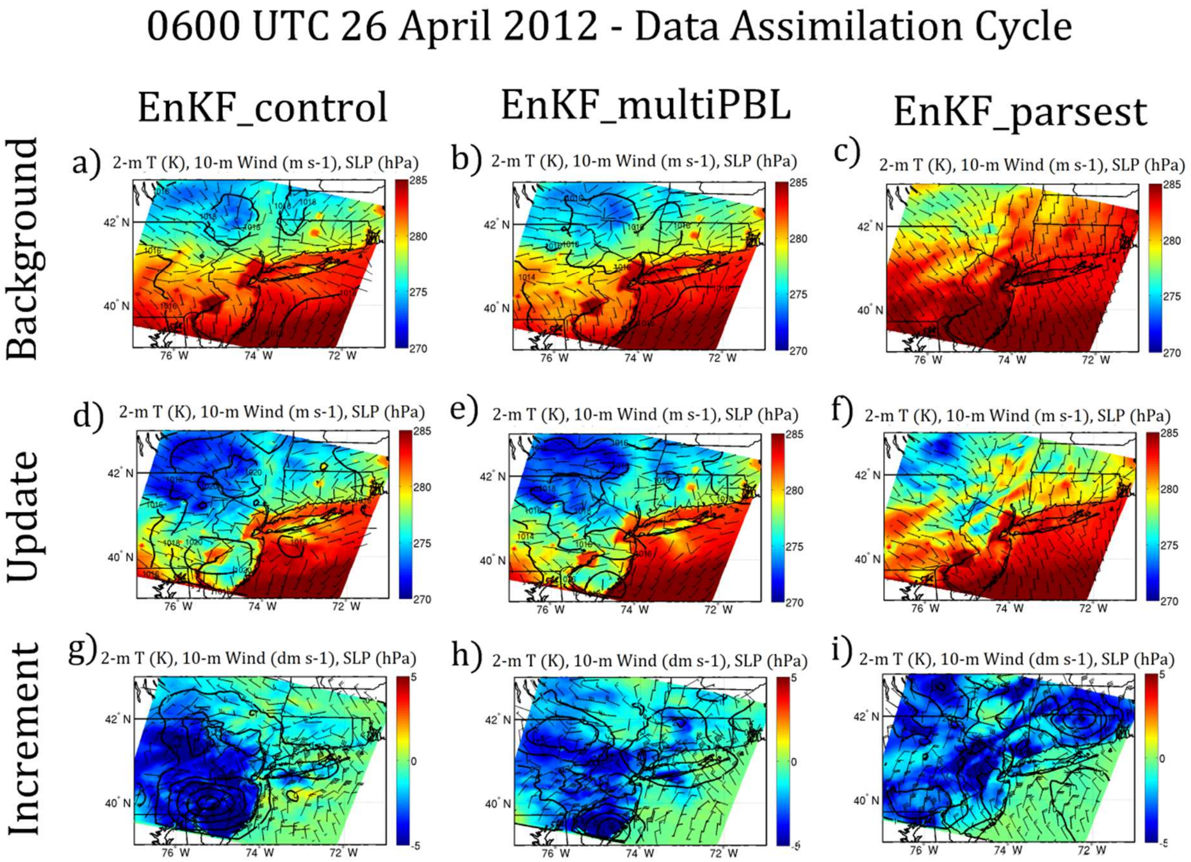
0600 UTC 26 April 2012 - Data Assimilation Cycle
EnKF_control
EnKF_multiPBL
EnKF_parsest
Background
Update
Increment
a) 2-m T (K), 10-m Wind (m s-1), SLP (hPa)
b) 2-m T (K), 10-m Wind (m s-1), SLP (hPa)
c) 2-m T (K), 10-m Wind (m s-1), SLP (hPa)
d) 2-m T (K), 10-m Wind (m s-1), SLP (hPa)
e) 2-m T (K), 10-m Wind (m s-1), SLP (hPa)
f) 2-m T (K), 10-m Wind (m s-1), SLP (hPa)
g) 2-m T (K), 10-m Wind (dm s-1), SLP (hPa)
h) 2-m T (K), 10-m Wind (dm s-1), SLP (hPa)
i) 2-m T (K), 10-m Wind (dm s-1), SLP (hPa)
42° N
40° N
76° W
74° W
72° W
285
280
275
270
5
0
-5

Figure 5.16: Same as Fig. 5.15 but for 0600 UTC 26 April 2012.

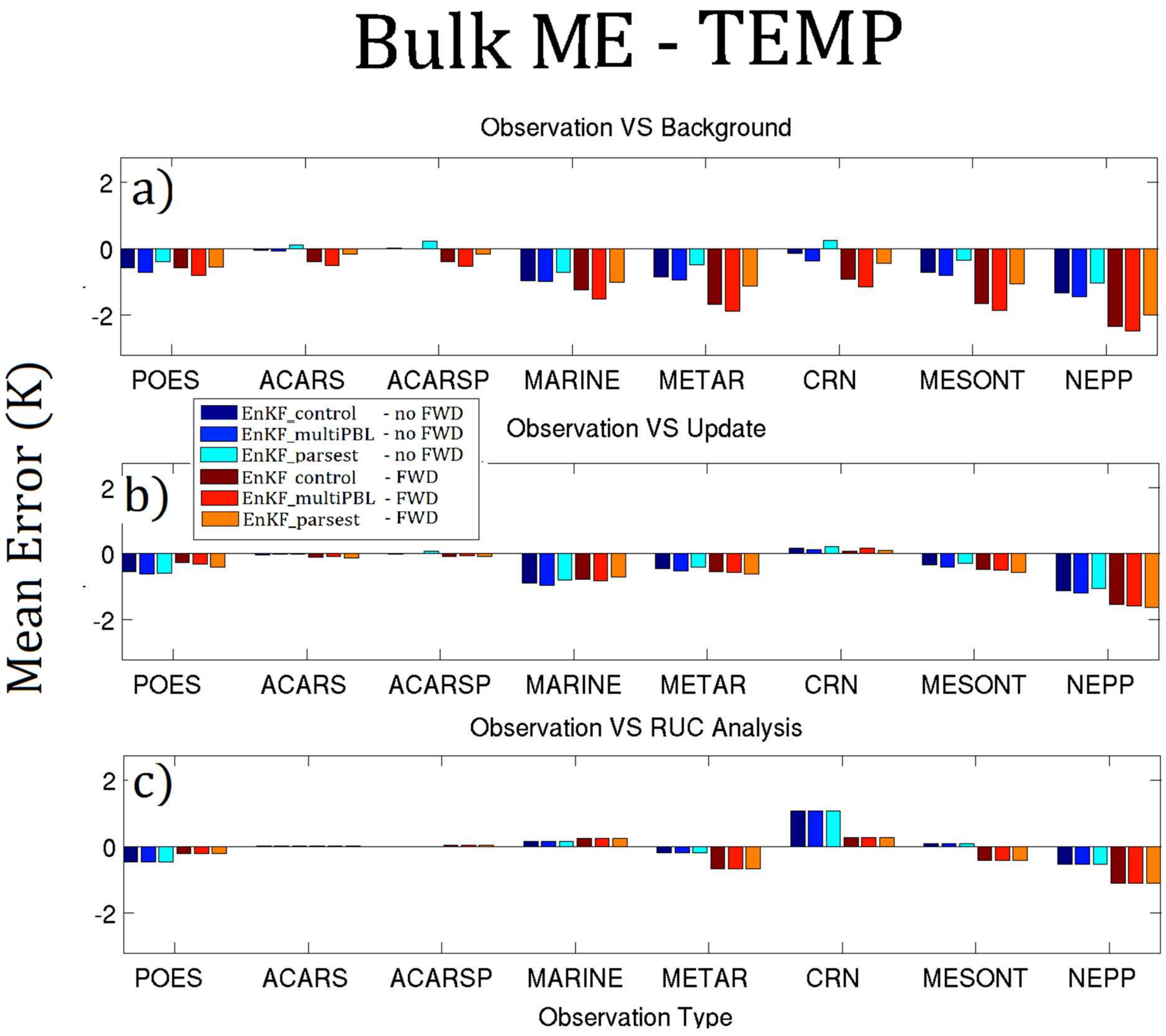

Figure 5.17: Background (a), update (b), and RUC (c) bulk TEMP ME for EnKF_control, EnKF_multiPBL, and EnKF_parsest on FWDs (FWI = 3; warm colors) and non-FWDs (cool colors).

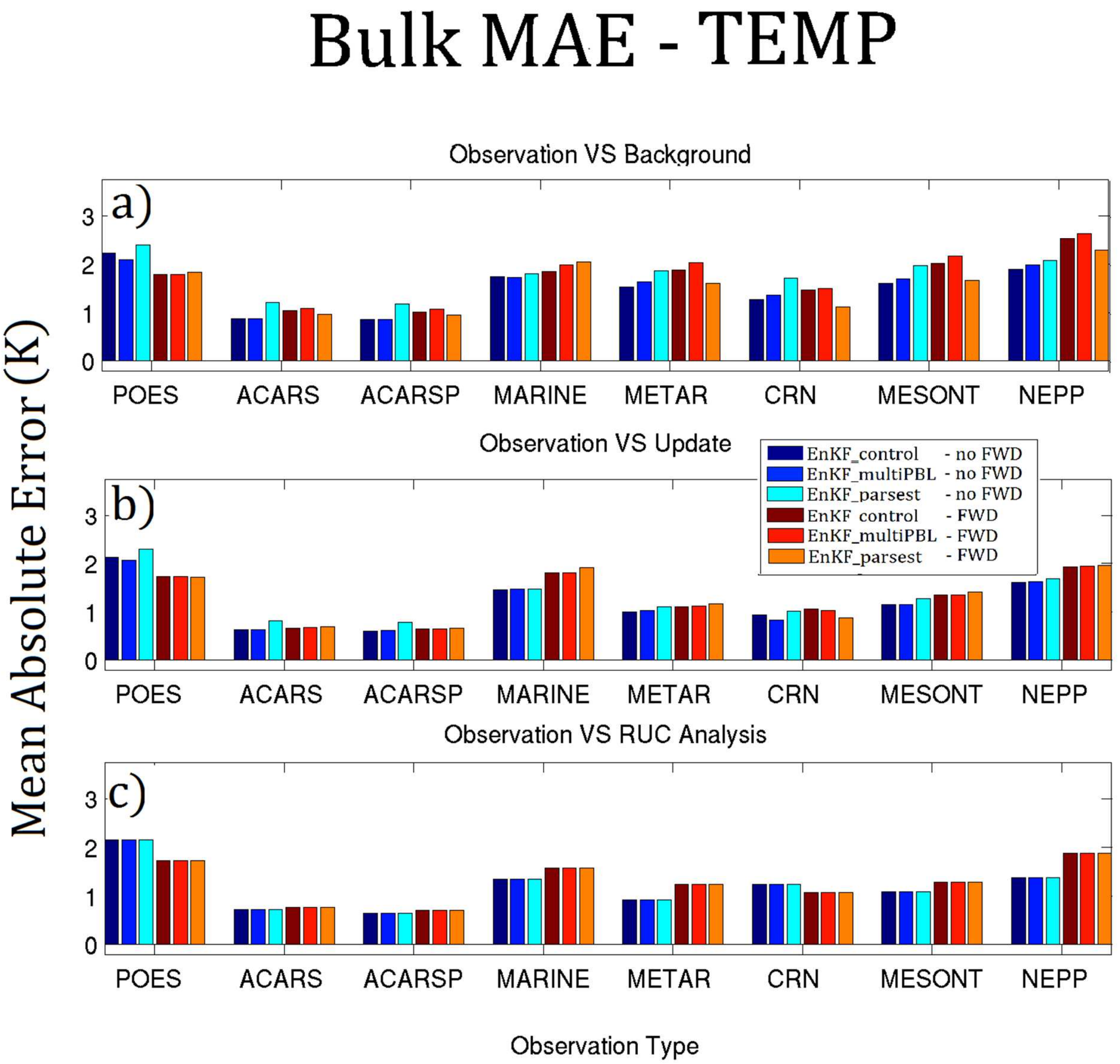

Figure 5.18: Same as Fig. 5.17, but for bulk TEMP MAE.

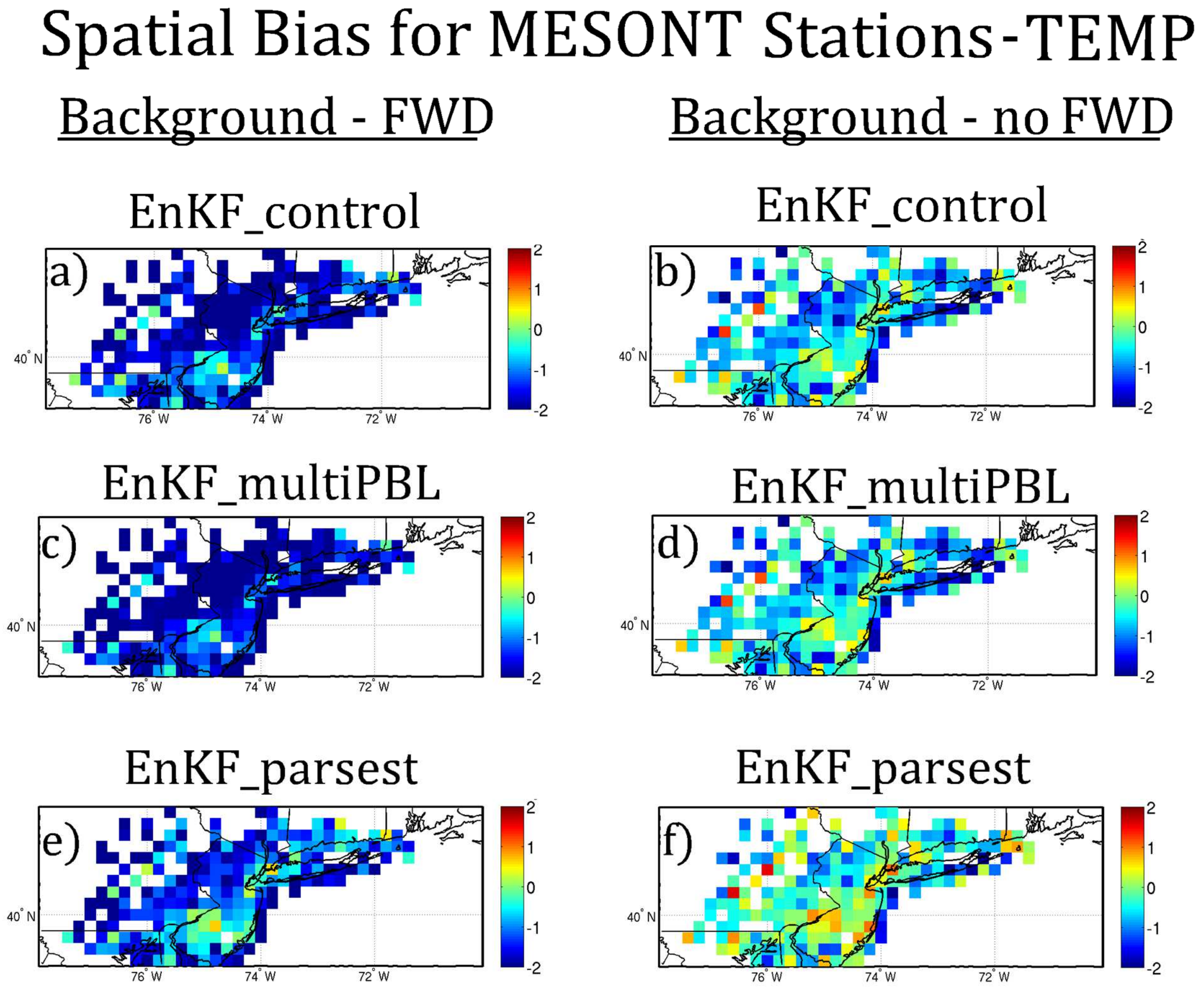

Figure 5.19: Background spatial TEMP MAE verified against MESONT observations on a 0.2° by 0.2° latitude/longitude grid subset by FWDs (b,d,f) and non-FWDs (a,c,e) for EnKF_control (a,b), EnKF_multiPBL (c,d), and EnKF_parsest (e,f).

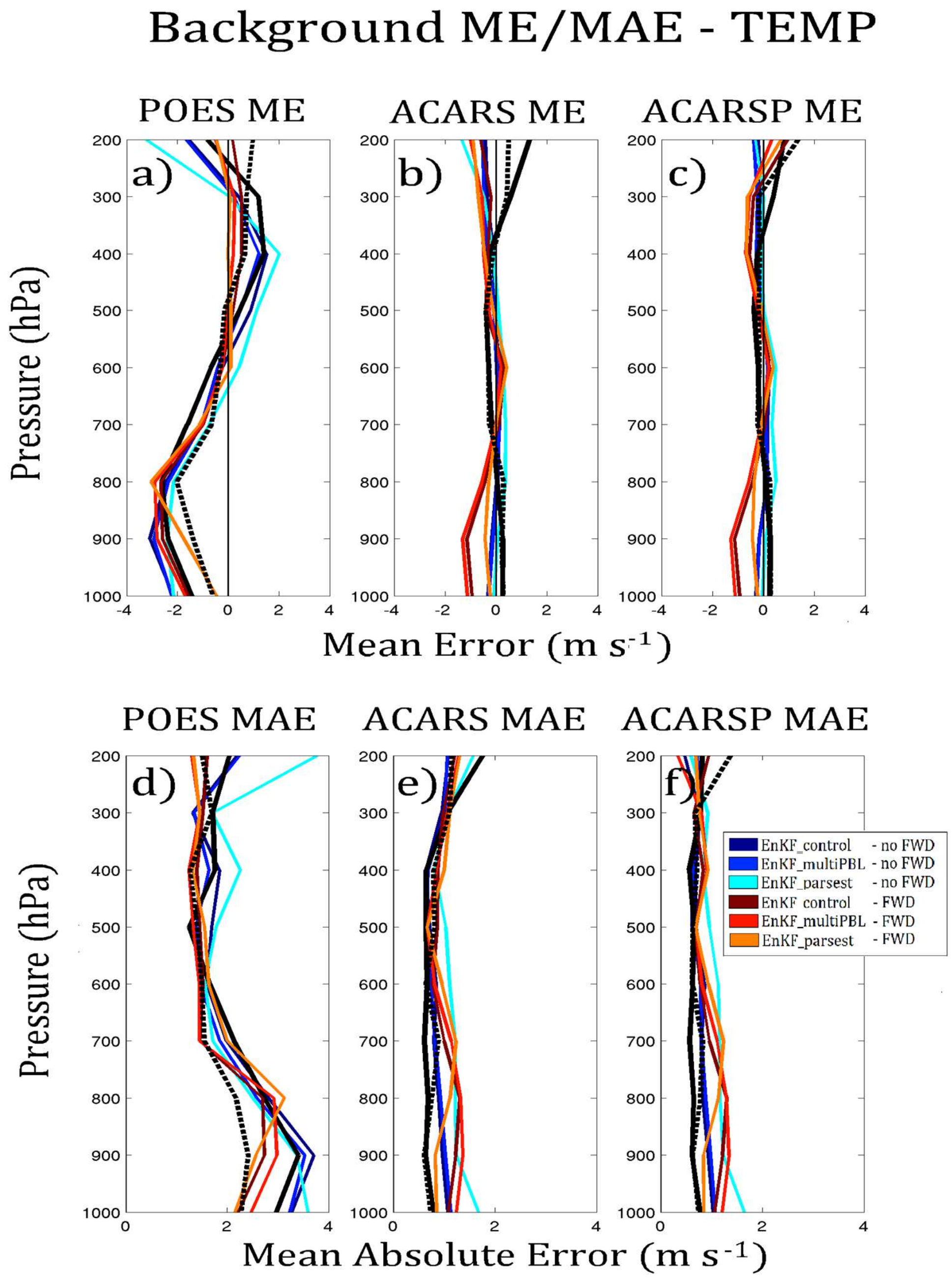

Figure 5.20: TEMP vertical ME (a,b,c) and MAE (d,e,f) separating FWDs and non-FWDs for POES (a,d), ACARS (b,e), and ACARSP (c,f). Solid (dashed) black line shows the RUC analysis for non-FWDs (FWDs).

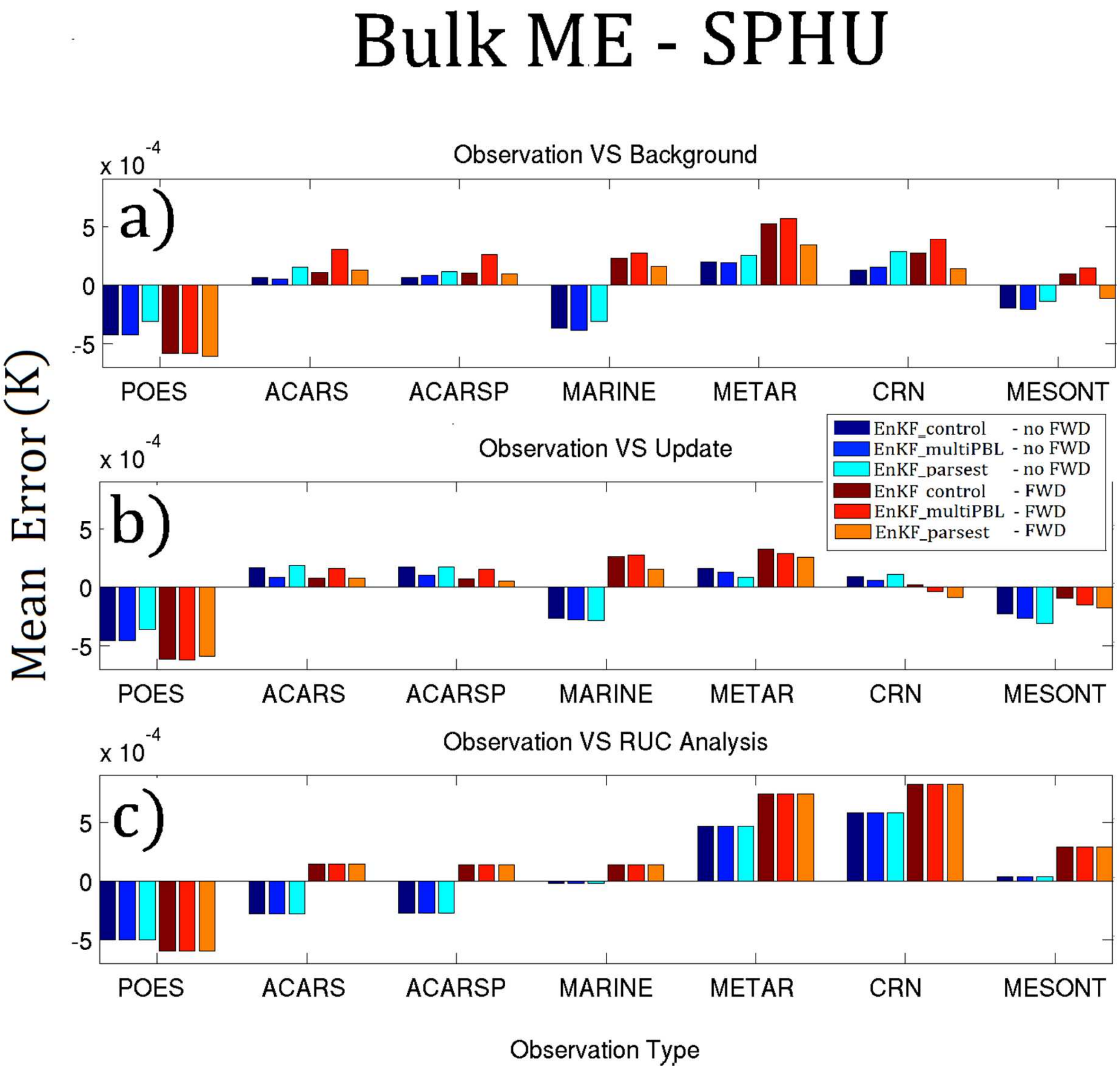

Figure 5.21: Same as Figure 5.17, but for SPHU.

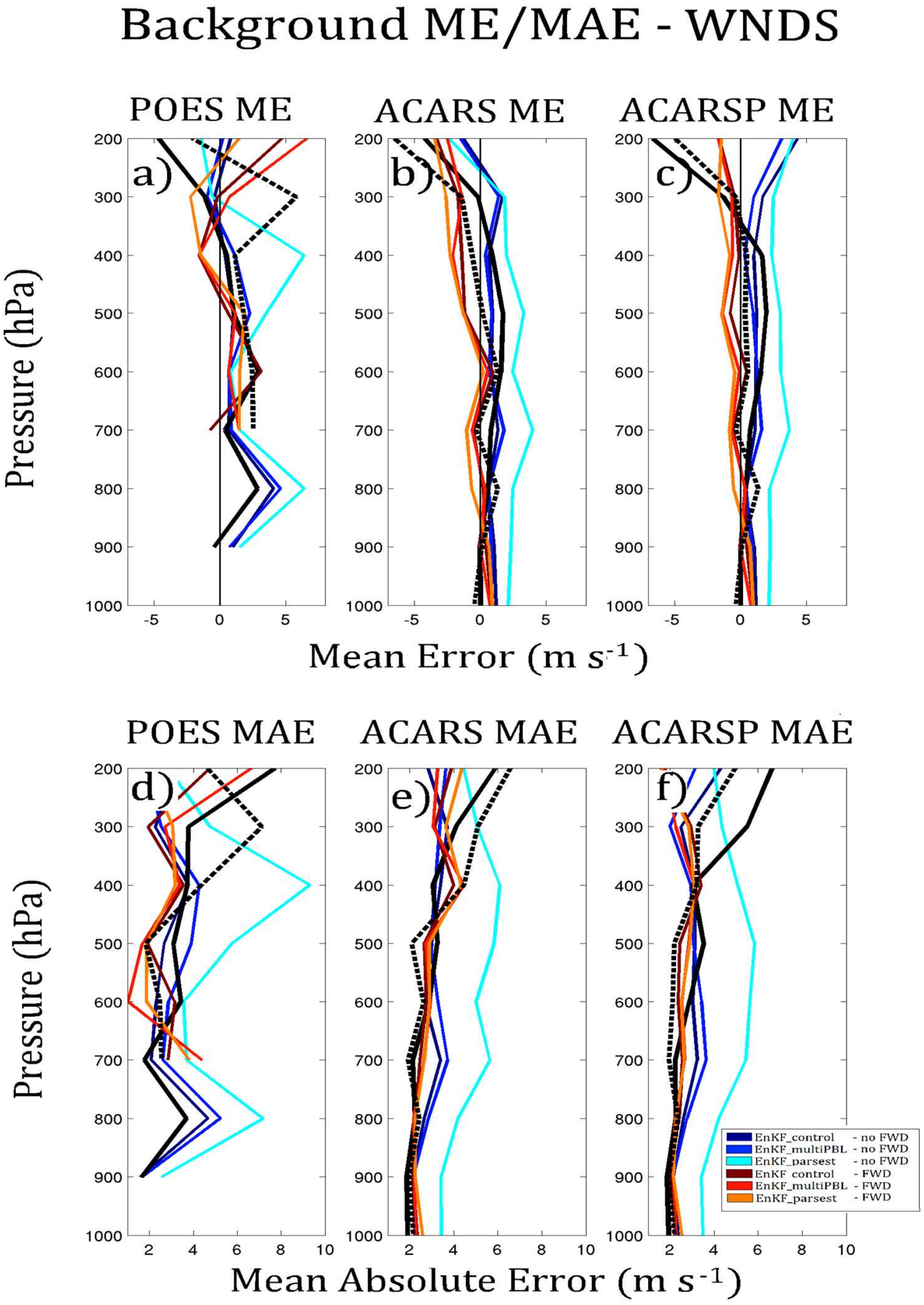

Figure 5.22: Same as Figure 5.20, but for WNDS.

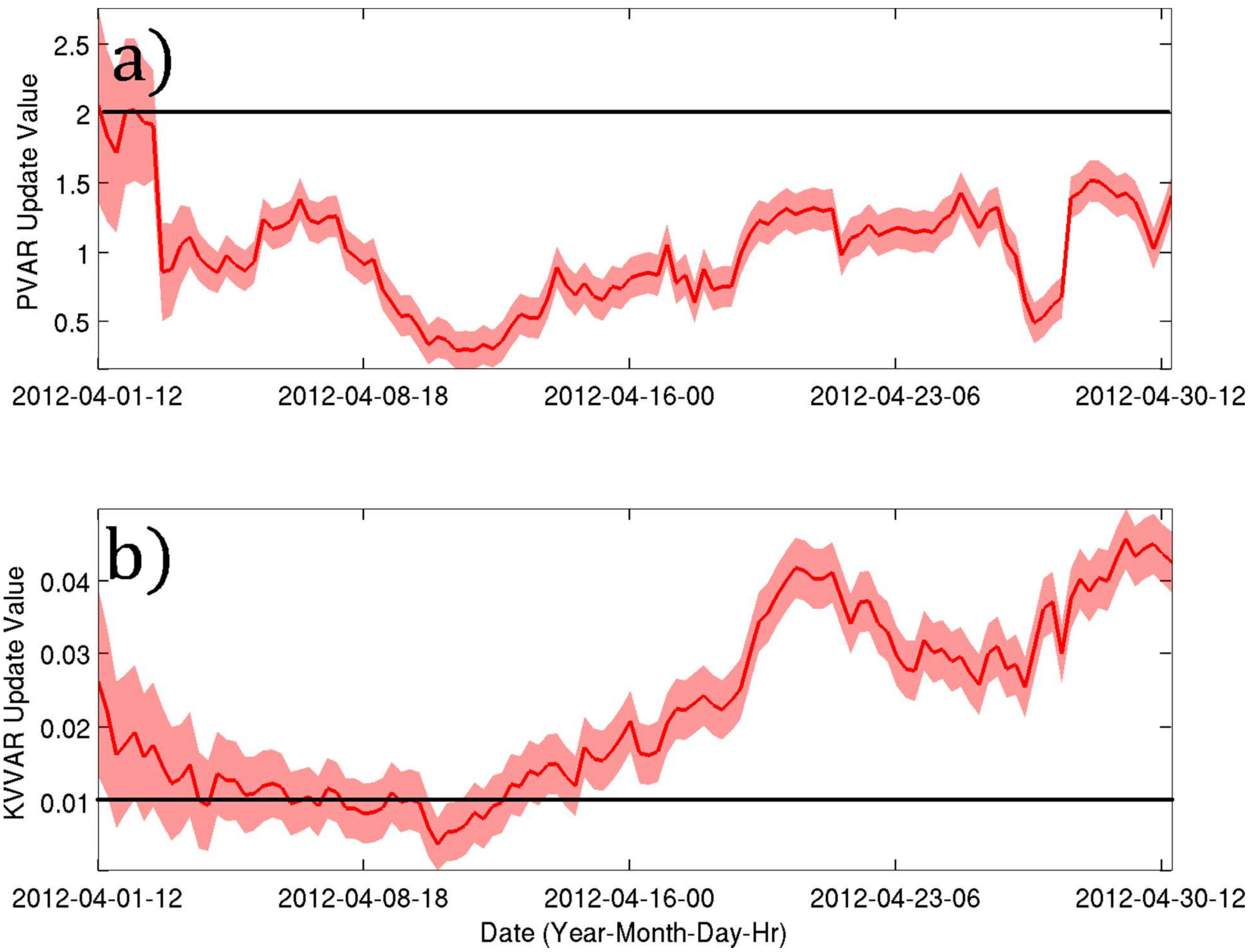

Figure 5.23: April of 2012 time series plots for PVAR (a) and KVVAR (b) with one standard deviation shading of the ensemble member variability.

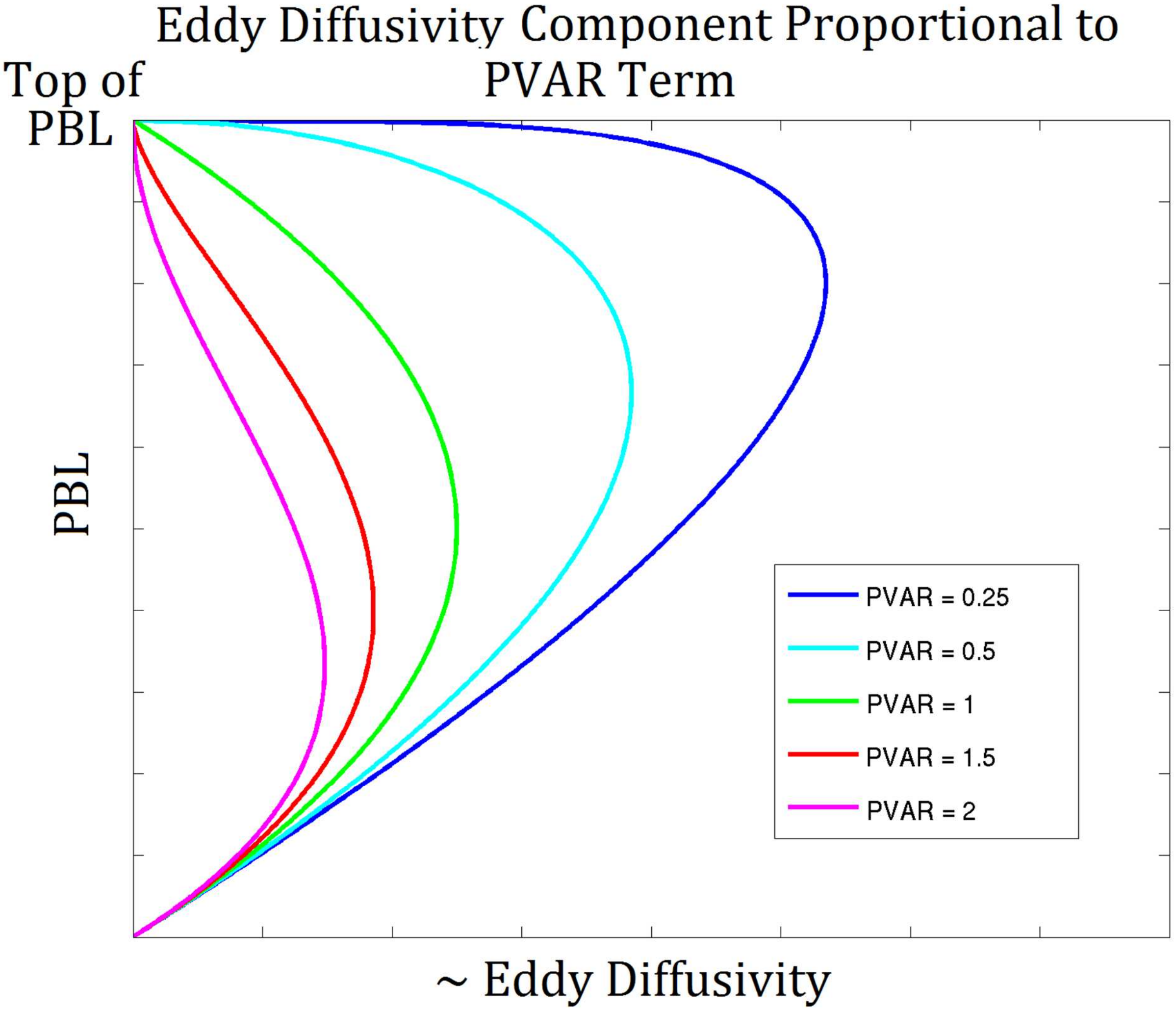

Figure 5.24: The contributing influence of the PVAR parameter to the shape and magnitude of vertical eddy diffusivity within the ACM2 PBL scheme.

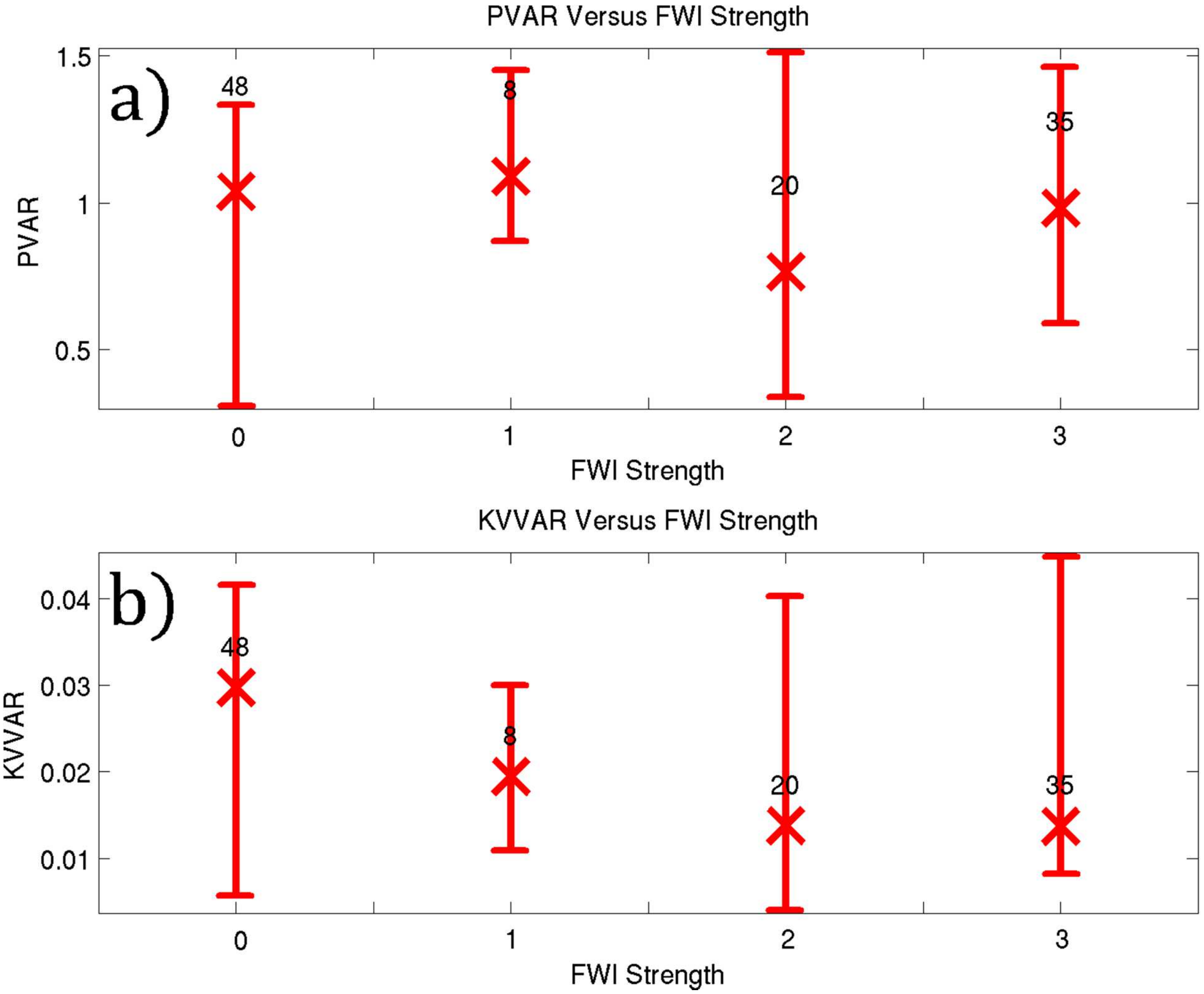

Figure 5.25: Median PVAR (a) and KVVAR (b) value by FWI strength with total sample size numbered and bars denoting the 2.5th and 97.5th percentile.

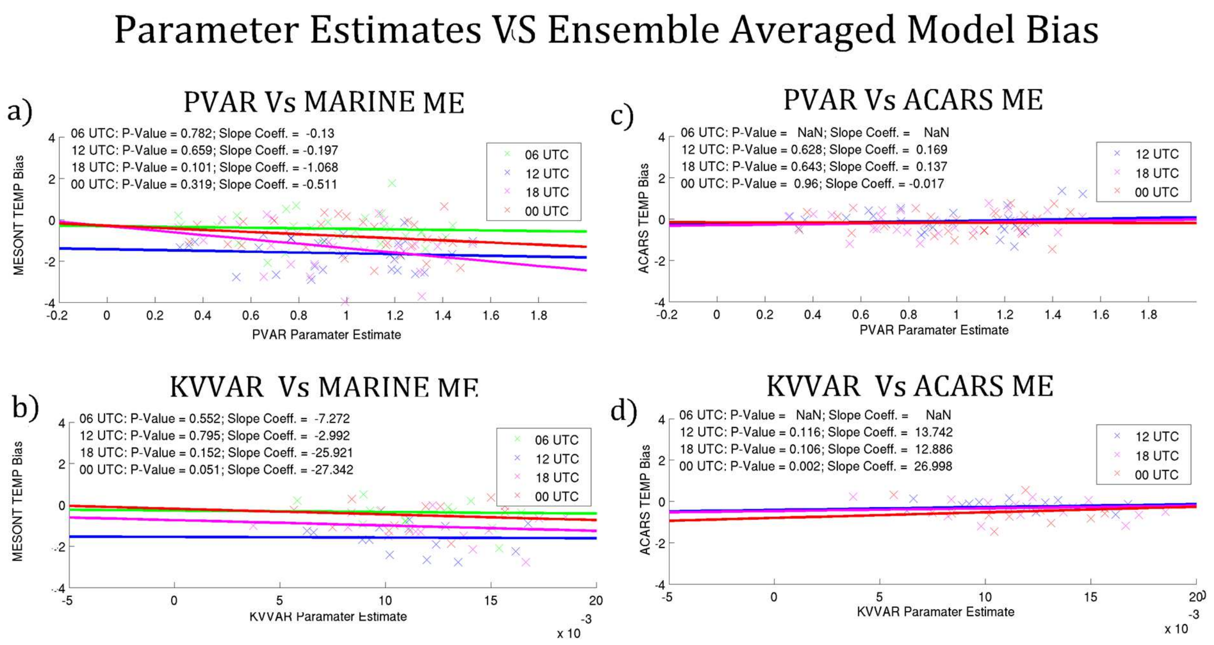

Figure 5.26: MESONT (a,b) and ACARS (c,d) domain averaged TEMP bias regressed on PVAR (a,b) and KVVAR (c,d) parameter estimates by time of day (i.e. 00 UTC, 06 UTC, 12 UTC, and 18 UTC). Slope coefficients and p-values for the slope coefficients shown in the top left portion of each subplot.

Chapter 6:

Conclusions and Future Work

6.1 Summary

This dissertation is devoted to exploring and improving model performance on fire weather days (FWDs) through the use of ensemble modeling and data assimilation over the Northeast United States (NEUS). Since the majority of fire weather studies typically focus over the western United States, a new statistical fire weather index (FWI) is developed based on near-surface weather variables and fire occurrence data within the NEUS. Thereafter, the FWI and more traditional fire risk indices, such as the National Fire Danger Rating System (NFDRS), are used to subset ensemble model data for verification and post-processing using a variety of different statistical methods. Finally, an Ensemble Kalman Filter (EnKF) is used to both simultaneously correct and explore potential sources of model error stemming from the parameterized planetary boundary layer (PBL) during a period of extensive FWDs.

6.2 Conclusions

a. Determination of fire weather days

There is no formal definition for the atmospheric conditions that represent a FWD. Furthermore, very few studies have focused on fire weather over the NEUS. As a result, this dissertation employs two different definitions. The first definition, referred to as FWD1, uses the Fire Potential Index (FPI) and when unavailable the NFDRS between 2007 and 2009. A FWD1 must have at least 10% of the NEUS domain achieving a "high" category for the FPI or NFDRS while the remainder of the domain is at least in the "moderate" category. Therefore, FWD1 represents a relatively simple method using preexisting indices to subset days of enhanced fire risk.

The FPI and NFDRS are largely based on fire occurrence data over the western United States and may not be appropriate for use over the NEUS. Therefore, a new statistical FWI is developed based on fire occurrence data within the NEUS to represent FWD2. The fire occurrence data within the NEUS is collected from Pollina et al. (2013) and regressed on different combinations of variables from the Automated Surface Observing System (ASOS) observations in a binomial logistic regression model. The optimal binomial logistic regression formulation using independent verification consists of two predictors; 2-m relative humidity (RELH) and 2-m temperature (TEMP). This model shows good average reliability with probability of fire occurrence and produces brier skill scores better than climatology for two separate domains centered over the New Jersey and the New York City (NYC) region.

The statistical FWI categories are based on the probability of fire occurrence output from the binomial logistic regression model. A FWI of zero has a statistical probability of fire occurrence below 30%, a FWI of one has probabilities between 30% and 40%, a FWI of two between 40% and 50%, and a FWI of 3 indicates probabilities greater than 50%. Furthermore, a FWD2 event occurs with a FWI value of greater than zero. Unlike FWD1, FWD2 has a direct relationship to fire occurrence within the NEUS domain. The FWI exhibits statistically significantly improved (> 95% confidence) critical success index (CSI), false alarm ratio (FAR), and hit rate (HIT) compared to climatology. In addition, FWD2 behaves similarly to the observed climatology of wildfires between 1999 and 2008.

b. Verification and post-processing of fire weather days

An extensive ensemble verification and post-processing of FWDs has never been performed. Quantifying and correcting model biases specific to FWDs is critical given that ensembles can be used to predict the atmospheric component in fire weather forecasts. Site-specific verification and post-processing of FWD1 events are analyzed and compared to the warm season average for the NCEP National Centers for Environmental Prediction (NCEP) Short Range Ensemble Forecast (SREF) and Stony Brook University (SBU) ensembles between 2007 and 2009. The optimal bias correction methodology depends on the model state variable analyzed; with an additive bias correction working optimally for 2-m temperature (TEMP) and a cumulative distribution function (CDF) bias correction method working best for 10-m wind speed (WNDS). After bias correction, Bayesian Model Averaging (BMA) is applied to TEMP and WNDS to calibrate the ensemble output. The two post-processing methods (i.e. bias correction and BMA) are trained using two different types of training windows; one with the most recent 14 days (sequential training) and another with the most recent 14 FWDs (conditional training). To test for statistical significance, the data is bootstrapped 10000 times with the training and verification windows kept separate to preserve independence.

FWD1 events exhibit a greater cool TEMP bias than the warm season average (by 1.34 K), which is reflected in the lingering bias after sequential bias correction (SBC). Conditional bias correction (CBC) removes all of the lingering average bias and reduces the mean absolute error (MAE; by 0.21 K). Most of the cool TEMP bias occurs in the afternoon hours and averages about 2.0 K colder than the warm season average during this time. In addition, FWD1 exhibits a positive 2-m specific humidity (SPHU; by 0.74 g kg^{-1}) bias and a weaker positive or greater negative WNDS bias (by -0.55 m s^{-1}) compared to the warm season average. In all cases CBC is more effective at removing bias than SBC, with varying improvements to MAE.

In addition to 1st moment biases, the SREF and SBU ensembles also exhibit underdispersion in their solutions compared to reality. As a result, BMA is applied to a subset of the SBU and SREF ensembles to inflate the ensemble variance. BMA is shown to effectively calibrate ensemble probabilities for TEMP and WNDS when combined with CBC (BMA-CT), but cannot correct for the lingering bias if SBC is used before BMA (BMA-CT) for TEMP. As a result, conditional training is necessary for reliable probabilities to be generated with BMA. Furthermore, the performance of BMA is sensitive to the smaller ensemble selected from the total members within SBU and SREF. In general, the selection of the control SREF members and the best performing SBU members within each planetary boundary layer (PBL) subset results in a better BMA calibration than selecting random members. This suggests that BMA performs better when calibrated with unique and optimal ensemble members (i.e. novel quality over quantity).

The verification and post-processing results of FWD1 events provide a good framework for biases specific to FWDs and the potential effectiveness of operational post-processing on these days. As mentioned earlier, there is no quantitative known relationship between FWD1 events and fire occurrence over the NEUS. Therefore, the verification and bias correction of

FWD2 events for the NCEP SREF are explored using the gridded Rapid Update Cycle (RUC) and Rapid Refresh (RAP) analyses between 2007 and 2014. In order to subset and verify FWD2 events, the FWI methodology is adapted to a gridded framework.

As with FWD1 events, SBC on FWD2 events result in lingering model biases that are too cool (TEMP = -0.73 K), moist (RELH = 3.74 %) and slightly less windy (WNDS = -0.44 m s^{-1}) at 1000 hPa. These model biases generally extend upward through the PBL to about 800 hPa, suggesting the potential for the model error source to originate at the land-surface interface or in the PBL scheme. The newer version of the SREF implemented in 2012 generally reduces but does not eliminate biases for TEMP (by 0.37 K) and RELH (by 0.37 %) at 1000 hPa. Regardless, biases in 2-m TEMP and 2-m RELH result in a severe underprediction of the FWI in the SREF (ME = 0.43 at FWI >=1). Similar to the FWD1 subset verification, CBC performs optimally by largely removing the average model bias and generally reducing MAE depending on the model state variable analyzed. If only SBC can be used operationally to bias correct ensemble model output, it should be applied to TEMP and RELH but not WNDS. Otherwise, CBC should be used on all model variables during FWDs.

TEMP model biases vary by month, with a climatological peak in the negative bias between March and May (-1.12 K), which is significantly less than for FWD2 events. FWDs have cooler TEMP biases compared to the climatological average throughout most of the year, suggesting that the cool biases on FWDs are not solely caused by warm and dry events occurring before the spring green-up period. CBC helps to remove the majority of this bias, but cannot directly account for seasonal biases.

c. Using the Ensemble Kalman Filter to explore model bias and error on FWDs

The use of an Ensemble Kalman Filter (EnKF) to explore model performance on FWDs has never been attempted. In chapter 5, an EnKF is used to assess the rapid growth of model biases on FWDs, the impact of data assimilation for FWDs, and the effect of simultaneous state and parameter estimation (SSPE) on the forecast. Furthermore, SSPE is used to explore structural model errors within the parameterized PBL by examining the parameter variability in time. Although a few previous studies have used SSPE in PBL schemes to improve the model state or explore structural model errors, FWDs provide a unique case featuring a dry and well-mixed PBL. The parameters estimated in this dissertation include PVAR, which affects the magnitude and vertical distribution of eddy diffusivity in an unstable PBL, and KVVAR, which affects vertical mixing in a stable PBL.

Chapter 5 focuses on April of 2012, which features an extensive period of FWDs during a time of moderate drought. First, several sensitivity runs are conducted to ensure the EnKF is performing well using a 5 day simulation between 06 April 2012 and 11 April 2012. The sensitivity runs adjust the observational error variance and localization radius of influence for surface observations. The optimal sensitivity run consists of a halved observational error variance and tripled radius of influence from the default value in Zhang et al. 2014, which is used for the remainder of this dissertation.

To evaluate EnKF performance on FWDs, three EnKF sensitivity runs are conducted throughout April of 2012: 1) A control run (EnKF_control), 2) a multi-PBL run that uses the Yonsei University (YSU), Mellor-Yamada Nakanishi and Niino (MYNN), and ACM2 schemes (EnKF_multiPBL), and 3) a run that performs simultaneous state and parameter estimation (SSPE) within the ACM2 PBL scheme (EnKF_parsest). Cool and moist PBL model biases develop rapidly (within 6 hours) for both EnKF_control and EnKF_multiPBL, suggesting that post-processing is essential at any forecast hour. The EnKF is generally effective at removing bias in the update, even if the background field has large biases. Consistent with chapter 4, the negative background TEMP and positive SPHU model bias are greater in the PBL on FWDs, with little to no improvement in EnKF_multiPBL compared to EnKF_control.

The EnKF_parsest significantly reduces the background land-based surface negative TEMP (by 0.58 K) and positive SPHU (by 0.19 g kg-1) biases on FWDs, which results in significantly improved MAE for TEMP (by 0.32 K) but not SPHU. However, EnKF_parsest degrades the near-surface MAE over land on non-FWDs (by 0.35 K for TEMP and 0.10 g kg-1 for SPHU). This is likely caused by SSPE adjusting the parameters within the ACM2 PBL to drastically increase the eddy diffusivity. Therefore, the increased eddy diffusivity on FWDs could lead toward a deeper, drier and warmer PBL, which reduces the low level TEMP and RELH biases. However, on non-FWDs, the PBL is spuriously overmixed, which results in a significant positive WNDS bias (2.40 m s^{-1}) throughout the troposphere.

The EnKF_paresest results are encouraging and represent a good first step toward understanding and improving WRF based forecasts on FWDs. For instance, in a parameterized PBL without structural model error, the parameter estimates should converge toward an optimal value or be relatively stationary in time. The parameters observed in this dissertation do not converge and vary by about 400% throughout the April 2012 simulation. In addition, the parameter values of EnKF_parsest appear to vary with the near-surface model bias of the background field in the EnKF_control run. For instance, the KVVAR parameter has a significant negative relationship between increased eddy diffusivity and the negative TEMP bias of EnKF_control using MARINE flight data at 00 UTC. This means that in the presence of a large negative TEMP bias in EnKF_control, the parameter values of EnKF_parsest may adjust to overmix the PBL in an attempt to remove this bias. However, the dynamical impact of SSPE on the PBL should be treated with caution and a more comprehensive study is warranted.

6.3 Future Work

a. Extending the FWI to consider additional predictors

It is quite possible that the FWI can be improved by considering additional predictors in the binomial logistic regression model. However, the FWI is developed so that it could be used in ensemble models. Therefore the predictors in the binomial logistic regression must be somehow obtainable operationally and realistically represented. One potentially useful variable is an index representing preexisting drought, which likely exacerbates fire risk on a FWD. This was true for the 09 April 2012 case, where a moderate drought was already in place before a FWD developed. Since this variable would likely remain nearly static over the course of a mesoscale model simulation, a FWI that considers drought could be implemented operationally.

An additional predictor that considers the underlying dry fuels could also be added to the binomial logistic regression model. This addition would be trickier than a drought index since it is not immediately obvious how to represent dry fuels operationally in an operational ensemble mesoscale model. It is also possible that RELH serves as a good proxy for at least the top layer state of dry fuels, since latent heat fluxes from the ground into the atmosphere would prevent extremely low RELH values. Nonetheless, additional research is needed to determine a more ideal configuration for the FWI.

b. Exploring regime-based model bias on FWDs

This dissertation presents evidence that model performance differs between the typical FWD and the climatological average. However, a FWD can be associated with multiple synoptic flow patterns. For instance, Pollina et al. (2013) demonstrates that 78% of all fire occurrence within the NEUS coastal plain are associated with different configurations and orientations of a controlling high pressure near the surface; a "pre high" that is building into the NEUS from the northwest (30 %), an "extended high" configuration centered along the United States coastline (24 %), and the "back of high" centered off the eastern sea board (24 %).

FWD model biases may vary with the synoptic flow pattern or the position of the approaching high pressure system. To demonstrate this, model bias is calculated as a function of the nearest high pressure centroid location [using North American Regional Reanalysis (NARR) data] from NYC for all FWD1 events and presented as stacked angle histogram plot (Fig. 6.1). For the majority of FWD1 events, the high pressure center is to the northwest of NYC, which is in agreement with the "pre-high" results from Pollina et al. (2013), albeit over a different domain. Interestingly, model bias is more negative with the "pre-high" set up (-3.7 K) compared to the high pressure center displaced to the east or south of NYC (-2.6 K). Therefore, the fire synoptic classifications from Pollina et al. (2013) may have their own unique biases.

Another potential method for separating synoptic flow patterns statistically is to use empirical orthogonal function (EOF) analysis. From the EOF analysis, the principal component (PC) time series can be compared to model bias to ascertain if any variance explained can be attributed to model biases. As a proof of concept, EOF analysis is performed on the North American Regional Reanalysis (NARR) sea level pressure (SLP) data, and the resulting PC's are correlated to FWD1 2-m TEMP model biases over the NEUS. The first PC (explaining 15% of the variance) is correlated to TEMP biases for the raw SREF, SBC and CBC in Fig. 6.2. The EOF pattern in its positive phase appears to reflect an extra-tropical correlation between a high pressure center over the NEUS (perhaps a pre-high setup from Pollina et al. 2013) and a Greenland (Alaskan) low pressure (high pressure) anomaly. Correlations between PC 1 and the raw 2-m TEMP bias are significantly negative for 12 of the 21 SREF members. This suggests that the positive phase of EOF 1 (i.e. a positive SLP anomaly over the NEUS) has a statistically significant relationship with negative TEMP anomalies over the NEUS. Interestingly, there are still statistically significant correlations between PC 1 and 8 (9) of the SREF members after post-processing with SBC (CBC), suggesting that the mode of variability in EOF 1 is not fully captured by any post-processing method.

Planned future work will use EOF analysis on the NARR dataset to reduce the dimensionality of the data before applying cluster analysis (CA). The number of optimal clusters or flow regimes on FWDs is not inherently known, and techniques such as assessing the silhouette value (Rousseeuw et al. 1987) can be used to determine how tightly grouped the data is within each cluster. Verification and post-processing of FWDs can then be performed by cluster to determine statistically significant differences and the potential utility of applying a cluster analysis bias correction operationally.

c. Implementing an operational conditional post-processing technique for FWDs

Sectons 3.4 and 4.4 discuss the implementation of an operational CBC for FWDs. The main challenge with this approach is that FWDs must be known a priori. Due to the presence of model biases on FWDs, this poses a challenging forecast question: How accurately can FWDs be predicted in advance? This model verification question is somewhat different than what is addressed in chapters 3 and 4. One could use a thresholding technique based on typical model bias that "turns on" when the model becomes sufficiently favorable for FWDs. However, this method seems ad-hoc and somewhat dependent on the same post-processing methods to be employed after the thresholding technique.

A second method could be investigated which invokes a CA method similar to that proposed in section 6.1b. This will require a different type of verification that is inherently binomial; either a FWD did or did not occur. Essentially, CA will be tested in a predictive sense to determine if the model can predict the synoptic or mesoscale setup of FWDs. In addition, there will have to be a recognizable synoptic pattern for FWDs that CA is able to identify. This will likely require extensive testing with a variety of low and mid-level model state variables.

d. Additional research with simultaneous state and parameter estimation

Although encouraging, the results of SSPE from chapter 6 are very preliminary and open up several opportunities for future work. First, the effects of PVAR and KVVAR are difficult to separate within SSPE. The main complications from the SSPE run in chapter 5 occur during the day, when both KVVAR and PVAR have an effect on the WRF simulation. However, it is during the day when TEMP model biases are the most amplified and hence most relevant. Therefore, two additional EnKF simulations are proposed; one solely estimating PVAR while holding KVVAR fixed, and the other only estimating KVVAR. This allows for the individual effects from each parameter to be analyzed separately. The separate impacts of PVAR and KVVAR on MAE for FWDs and non-FWDs can then be examined.

The potential impact or relationship between parameter variability and model bias needs to be extended. For instance, Nielson-Gammon et al. (2010) stated the greatest impact of KVVAR was on WNDS. Therefore, the relationship between KVVAR (and PVAR) and WNDS should be explored in more detail. It is quite possible that the EnKF is simultaneously adjusting the parameter values (spuriously or correctly) and the analysis to remove model bias for a variety of model state variables. The relationship between MAE (or some other metric) and the parameter values should also be examined for the modeled surface and aloft.

` It might also be useful to connect the CA from section 6.3b and SSPE. Relating FWI to model bias may not be appropriate since there are a variety of flow-regimes that can create a FWD. It is possible that parameter estimates within SSPE are more sensitive to the flow-regime than they are to model biases near the surface. Once known, it would be relatively simple to extend the specific clusters determined in CA to the time series of parameter estimates. The synoptic setup and PBL structure with clusters that exhibit different parameter estimates could provide insight into structural model errors in the ACM2 PBL scheme. This would be particularly beneficial to examine for extreme FWDs, such as the 09 April 2012 case.

Finally, it is worthwhile to rule out potential sources of model bias not associated with the parameterized PBL. For instance, soil moisture, land-surface modeling, radiation scheme, clouds, or a combination of these processes could affect the PBL. Although these processes are linked in a complex and non-linear way, there are some relatively simple examinations that can be performed to rule out some non-PBL effects. The first is to ensure that the modeled soil moisture resembles reality. There is some soil moisture data available from the National Climatic Data Center (NCDC) and Meteorological Assimilation Data Ingest System (MADIS), as well as a variety of model derived soil fields. It would be worthwhile to compare the observed soil moisture to the modeled soil moisture for the April 2012 run to ensure there are no severe model biases. Secondly, the modeled clouds should be compared to the NARR or other reanalysis dataset to check for biases. It is possible that some biases in the model stem from the over-prediction of cumulus clouds during the afternoon. Third, critical cases studies should be analyzed in more detail, such as the differences in the modeled versus observed PBL development on 09 April 2012. Since biases develop rapidly, attention should be devoted to high resolution simulations focusing on short time steps where the biases originate in the model.

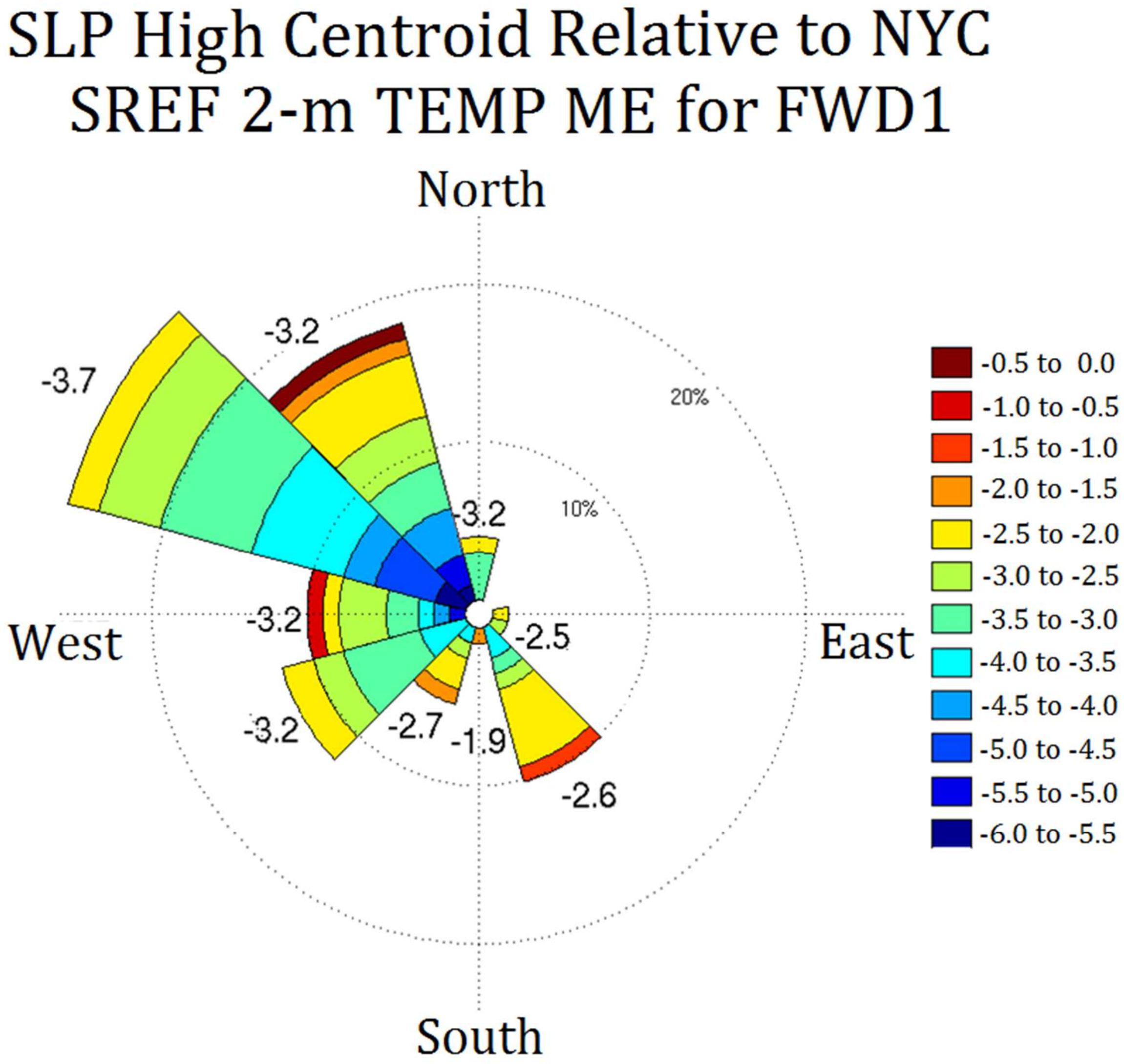

Figure 6.1: Stacked Angle histogram of 2-m TEMP model bias as a function of sea level pressure (SLP) high pressure centroid from NYC using FWD1 events. Numbers for each angle represent average model bias.

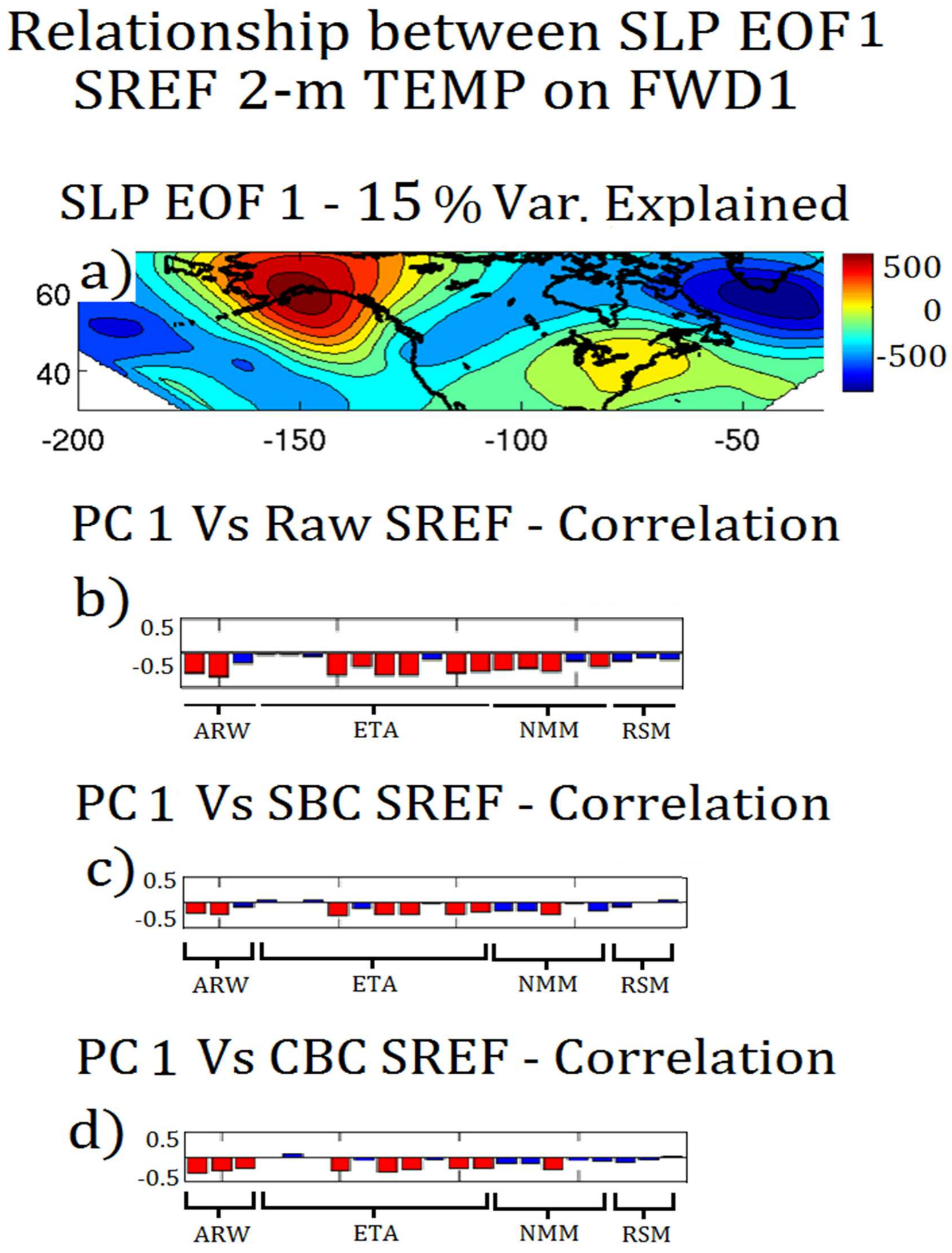

Figure 6.2: First empirical orthogonal function (EOF) of NARR SLP (a) and the correlation of the resulting PC1 to the SREF for raw (b), SBC (c) and CBC (d) 2-m TEMP bias. Red bars denote statistical significance at 95% for the correlations between PC time series and model bias.

References

Aksoy, A., F. Zhang, and J. W. Nielsen-Gammon, 2006a: Ensemble-based state and parameter estimation in a two-dimensional sea-breeze model. *Mon. Wea.Rev.*, **134**, 2951-2970.

Aksoy, A., F. Zhang, and J. W. Nielsen-Gammon, 2006b: Ensemble-based simultaneous state and parameter estimation with MM5. *Geophy. Res. Let.*, **33**, L12801.

Anderson, J., T. Hoar, K. Raeder, H. Liu, N. Collins, R. Torn, and A. Arellano, 2009: The Data Assimilation Research Testbed: A Community Facility. *Bulletin of the American Meteorological Society*, **90**, 1283-1296. doi:10.1175/2009BAMS2618.1

Anderson, J. L. and S. L. Anderson, 1999: AMonte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Wea. Rev.* **127**, 2741–2758.

Annan, J. D., and J. C. Hargreaves, 2004: Efficient parameter estimation for a highly chaotic system, *Tellus*, **56A**, 520-526.

Bakhshaii, A. and R. Stull, 2009: Deterministic Ensemble Forecasts Using Gene-Expression Programming. *Wea. Forecasting*, **24**:5, 1431-1451.

Barbero R, J. T. Abatzoglou, C. Kolden, K. Hegewisch, N. K. Larkin and H. Podschwit, 2014: Multi-scalar influence of weather and climate on very large-fires in the eastern United States. *Int. J. Climatol.* doi: 10.1002/joc.4090.

Barker, D. M., W. Huang, Y. R. Guo, A. J. Bourgeois, and Q. N. Xiao, 2004: A three-dimensional variational data assimilation system for MM5: Implementation and initial results. *Mon. Wea. Rev.*, **132**, 897–914.

Barth, M. F., P. A. Miller, and A. E. MacDonald, 2002: MADIS: The Meteorological Assimilation Data Ingest System. *In Symp. on Observations, Data Assimilation, and Probabilistic Prediction*, pages 20–25.

Benjamin, S. G., D. Devenyi, S. S. Weygandt, K. J. Brundage, J. M. Brown, G. A. Grell, D. Kim, B. E. Schwartz, T. G. Smirnova, T. L. Smith, and G. S. Manikin, 2004: An hourly assimilation/forecast cycle: The RUC. *Mon. Wea. Rev.*, **132**, 495-518

Benjamin, S., W. R. Moninger, S. R. Sahm, and T. L. Smith, 2007: Mesonet Wind Quality Monitoring Allowing Assimilation in the RUC and Other NCEP Models. Extended Abstract, *22nd Conf. Wea. Analysis Forecasting / 18th Conf. Num Wea. Pred.*, Park City, UT, Amer. Meteor. Soc., P1.33

Berrocal, V., A. E. Raftery, and T. Gneiting, 2007: Combining Spatial Statistical and Ensemble Information in Probabilistic Weather Forecasts. *Mon. Wea. Rev.* **135**, 1386-1402.

Bonavita, M., L. Torrisi, and F. Marcucci, 2008: The ensemble Kalman filter in an operational regional NWP system: Preliminary results with real observations. *Quart. J. Roy. Meteor. Soc*., **134**, 1733–1744.

Bradshaw, L. S., R. E. Burgan, J. D. Cohen, and J. E. Deeming. 1983: The 1978 National Fire-Danger Rating System: Technical Documentation. USDA Forest Service; Intermountain Forest and Range Experiment Station, General Technical Report INT-169, Ogden, Utah. 44 pp.

Brown, J., D. J. Seo, and J. Du, 2012: Verification of precipitation forecasts from NCEP's Short Range Ensemble Forecast (SREF) system with reference to ensemble stream flow prediction using lumped hydrologic models. *J. Hydrometeor*., **13**, 808-836.

Brundage, G., S. Manikin, G. DiMego, and B. Cosgrove, 2014: Hourly updated models: Rapid Refresh / HRRR review. NOAA NCEP/ERSL GSD. [Available online at http://ruc.noaa.gov/pdf/NCEP_PSR_2014_HRRR_COMBINED_lite.pdf]

Burgan, R. E., R. W. Klaver, and J. M. Klaver, 1998: Fuel models and fire potential from satellite and surface observations. *Int. J. of Wildland Fire*, **8**, 159-170.

Candille, G., 2009: The Multiensemble Approach: The NAEFS Example. *Mon. Wea. Rev.*, **137**, 1655–1665.

Cardil, A., D. M. Molina, J. Ramirez, and C. Vega-García, 2013: Trends in adverse weather patterns and large wildland fires in Aragón (NE Spain) from 1978 to 2010. *Nat. Hazards Earth Syst. Sci.*, **13**, 1393-1399.

Carrera, M. L., J. R. Gyakum, and C. A. Lin, 2009: Observational study of wind channeling within the St. Lawrence River Valley. *J. Appl. Meteorol. Climatol.*, **48**, 2341–2361.

Cartwright, T. J. and T. N. Krishnamurti, 2007: Warm Season Mesoscale Superensemble Precipitation Forecasts in the Southeastern United States. *Wea. Forecasting*, **22**, 873–886.

Charney, J. J. and D. Keyser, 2010: Mesoscale model simulation of the meteorological conditions during the 2 June 2002 Double Trouble State Park wildfire. *Int. J. Wildland Fire*, **19**, 427–448.

Chen, F., and J. Dudhia, 2001: Coupling an Advanced Land Surface Hydrology Model with the Penn State–NCAR MM5 Modeling System. Part I: Model implementation and sensitivity. *Mon. Wea. Rev.*, **129**, 569–585.

Clarke H., C. Lucas, and P. Smith, 2013: Changes in Australian fire weather between 1973 and 2010. *Int. J. Climatol.,* **33**, 931–944.

Coen, J. L., M. Cameron, J. Michalakes, E. G. Patton, P. J. Riggan, K. M. Yedinak, 2013: WRF-Fire: Coupled Weather–Wildland Fire Modeling with the Weather Research and Forecasting Model. *J. Appl. Meteor. Climatol.*, **52**, 16–38.

Collins, B. M., 2014: Fire weather and large fire potential in the northern Sierra Nevada. *Agricultural and Forest Meteorology*, **189-190**, 30-35.

Colle, B. A., J. B. Olson, and J. S. Tongue, 2003: Multi-season verification of the MM5: Part I, Comparison with the Eta over the Central and Eastern U.S. and impact of MM5 resolution*.* *Wea. Forecasting*, **18**, 431-457.

Deeming, J. E., J. W. Lancaster, M. A. Fosberg, R. W. Furman, and M. J. Schroeder, 1978: The National Fire-Danger Rating System. USDA Forest Service, Rocky Mountain Forest and Range Experiment Station, Research Paper RM-84, 165 pp.

Delle Monache, L., T. Nipen, X. Deng, Y. Zhou, and R. B. Stull, 2006: Ozone ensemble forecasts: 2. A Kalman filter predictor bias-correction. *J. Geophys. Res*, **111**, D05308, doi:10.1029/2005JD006311.

Delle Monache, L., T. Nipen, Y. Liu, G. Roux, and R. Stull, 2011: Kalman filter and analog schemes to post-process numerical weather predictions. *Mon. Wea. Rev.*, **139**, 3554-3570.

Du, J., G. DiMego, B. Zhou, D. Jovic, B. Ferrier, M. Pyle, G. Manikin, B. Yang, J. Wolff, and B. Etherton, 2012: New 16km NCEP Short-Range Ensemble Forecast (SREF) system: what we have and what we need? National Centers for Environmental Prediction. [Available online at http://www.dtcenter.org/events/workshops12/nuopc_2012/Presentations/SREF_2012ensembleworkshop_JDu.pdf].

Du, J., J. McQueen, G. DiMego, Z. Toth, D. Jovic, B. Zhou, and H. Chuang, 2006: New Dimension of NCEP Short-Range Ensemble Forecasting (SREF) System: Inclusion of WRF Members, Preprint, *WMO Expert Team Meeting on Ensemble Prediction System*, Exeter, UK, 5pp.

Dudhia, J., 1989: Numerical study of convection observed during the winter monsoon experiment using a mesoscale two-dimensional model. *J. Atmos. Sci*., **46**, 3077–3107.

Dumedah, G., and J. P. Walker, 2014: Intercomparison of the JULES and CABLE land surface models through assimilation of remotely sensed soil moisture in southeast Australia. *Advances in Water Resources* **74**, 231-244.

Eckel, F. A., M. S. Allen, and M .C. Sittel, 2012: Estimation of Ambiguity in Ensemble Forecasts. *Wea. Forecasting*, **27**, 50–69.

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

Engel, C. B, T. P. Lane, M. J. Reeder, M. Rezny, 2013: The meteorology of Black Saturday. *Q. J. R. Meteorol. Soc.* **139**, 585-599.

Erickson, M. J., B. A. Colle, and J. Charney, 2012: Impact of bias correction type and conditional training on Bayesian model averaging over the northeast United States. *Wea. Forecasting*, **27**, 1449-1469.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res*., **99 (C5)**, 10,143-10,162.

Mesinger, F., G. DiMego, E. Kalnay, K. Mitchell, P. C. Shafran, W. Ebisuzaki, D. Jović, J. Woollen, E. Rogers, E. H. Berbery, M. B. Ek, Y. Fan, R. Grumbine, W. Higgins, H. Li, Y. Lin, G. Manikin, D. Parrish, and W. Shi, 2006: North American Regional Reanalysis. *Bull. Amer. Meteor. Soc.*, **87**, 343–360.

Fisher, R. A., 1922: On the dominance ratio. *Proc. Roy. Soc. Edinb*, **42**, 321-341.

Fosberg M. A. 1978. Weather in Wildland Fire Management: The Fire Weather Index, *Conference on Sierra Nevada Meteorology.* Amer. Meteor. Soc., Lake Tahoe, CA.

Fox-Hughes, P., 2012: Springtime Fire Weather in Tasmania, Australia: Two Case Studies. *Wea. Forecasting*, **27**, 379–395.

Fraley, C., A. E. Raftery, and T. Gneiting, 2010: Calibrating Multi-Model Forecast Ensembles with Exchangeable and Missing Members using Bayesian Model Averaging. *Mon. Wea. Rev.,* **138**, 190-202.

Fujita T., D. J. Stensrud, and D. C. Dowell, 2007: Surface data assimilation using an ensemble Kalman filter approach with initial condition and model physics uncertainties. *Mon. Wea. Rev*., **135**, 1846-1868.

Gallus, W. A., Jr., M. E. Baldwin, and K. L. Elmore, 2007: Evaluation of probabilistic precipitation forecasts determined from Eta and AVN forecasted amounts. *Wea. Forecasting*, **22,** 207-215.

Gallus, W. A., Jr., and M. Segal, 2004: Does increased predicted warm season rainfall indicate enhanced likelihood of rain occurrence? *Wea. Forecasting*, **19**, 1127-1135.

Gaza, R. S., 1998: Mesoscale Meteorology and High Ozone in the Northeast United States. *J. Appl. Meteor.*, **37**, 961–977.

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

Glahn, B., M. Peroutka, J. Wiedenfeld, J. Wagner, G. Zylstra, B. Schuknecht, B. Jackson, 2009: MOS Uncertainty Estimates in an Ensemble Framework. *Mon. Wea. Rev*. **137**:1, 246-268.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. *Mon. Wea. Rev*., **133**, 1098–1118.

Golaz, J. C, V. E. Larson, J. A. Hansen, D. P. Schanen, and B. M. Griffin, 2007: Elucidating Model Inadequacies in a Cloud Parameterization by Use of an Ensemble-Based Calibration Framework. *Mon. Wea. Rev.*, **135**, 4077-4096.

Gopalakrishnan, S. G., S. B. Roy, and R. Avissar, 2000: An evaluation of the scale at which topographical features affect the convective boundary layer using large eddy simulations. *J. Atmos. Sci.*, **57,** 334–351.

Green, J. S., and L. S. Kalkstein, 1996: Quantitative analysis of summer air masses in the eastern United States and an application to human mortality. *Climate Research,* **7**, 43–53.

Grell, G. A., J. Dudhia and D. R. Stauffer, 1994: A Description of the Fifth-Generation Penn State/NCAR Mesoscale Model (MM5). NCAR Tech. Note. *NCAR/TN-398 + STR*, 128 pp.

Greybush, S. J., S. E. Haupt and G. S. Young, 2008: The Regime Dependence of Optimally Weighted Ensemble Model Consensus Forecasts of Surface Temperature. *Wea. Forecasting*, **23**, 1146–1161.

Haines, D. A., 1988: A Lower Atmosphere Severity Index for Wildland Fires. *Nat. Wea. Digest,* **3**, 23-27.

Hamill, T. M., 2007: Comments on "Calibrated Surface Temperature forecasts from the Canadian ensemble prediction system using Bayesian Model Averaging. *Mon. Wea. Rev.*, **135**, 4226-4230.

Hamill, T. M., and S. J. Colucci, 1997: Verification of Eta/RSM Short-Range Ensemble Forecasts. *Mon. Wea. Rev*., **125**, 1312-1327.

Hamill, T. M., and S. J. Colucci, 1998: Evaluation of Eta/RSM Ensemble Probabilistic Precipitation Forecasts. *Mon. Wea. Rev*., **126**, 711-724.

Hamill, T. M., J. S. Whitaker, and S. L. Mullen, 2006: Reforecasts: An Important Dataset for Improving Weather Predictions. *Bull. Amer. Meteor. Soc.*, **87**, 33–46.

Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: theory and application. *Mon. Wea. Rev.*, **134**, 3209-3229.

Hamilton C. and I. Ostapow, cited 2009: Long Island, NY—Central Pine Barrens. National Database of State and Local Wildfire Hazard Mitigation Programs, USDA Forest Service. [Available online at http://www.wildfireprograms.usda.gov/search.html?displayId=283.]

Hardy, C., R. D. Ottmar, J. Peterson, and J. Core, 2001: Smoke management guide for prescribed and wildland fire - 2000 edition. PMS 420-2. *NFES 1279.* Boise, ID, National Wildfire Coordination Group. 226 pp.

Heddinghaus, T. R. and P. Sabol, 1991: A review of the Palmer Drought Severity Index and where do we go from here? *Proceedings, 7th Conf. on Appl. Climatol.*, 10-13 September 1991, Boston: Amer. Meteor. Soc., 242-246.

Hegarty, J., H. Mao, and R. Talbot, 2007: Synoptic controls on summertime surface ozone in the northeastern United States, *J. Geophys. Res*., **112**, D14306, doi:10.1029/2006JD008170.

Hoadley, J. L., K. Westrick, S. A. Ferguson, S. L. Goodrick, L. Bradshaw, and P. Werth, 2004: The effect of model resolution in predicting meteorological parameters used in fire danger rating. *J. Appl. Meteor*., **43**, 1333–1347.

Hoadley, J. L., M. L. Rorig, L. Bradshaw, S. A. Ferguson, K. J. Westrick, S. L. Goodrick, and P. Werth, 2006. Evaluation of MM5 model resolution when applied to prediction of national fire-danger rating indexes. *Int. J. Wildland Fire,* **15**, 147–154.

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

Hong, S. Y., J. Dudhia, and S. H. Chen, 2004: A revised approach to ice microphysical processes for the bulk parameterization of cloud and precipitation. *Mon. Wea. Rev.*, **132**, 103–120.

Hu, X. M., J. W. Nielsen-Gammon, and F. Zhang, 2010a: Evaluation of Three Planetary Boundary Layer Schemes in the WRF Model. *J. Appl. Meteor. Climatol.*, **49**, 1831–1844.

Hu, X. M., F. Zhang, J. W. Nielsen-Gammon, 2010b: Ensemble-Based Simultaneous State and Parameter Estimation for Treatment of Mesoscale Model Error: A Real-data study. *Geophys. Res. Lett.*, **37**, L08802, doi:10.1029/2010GL043017.

Jin, Y., J. T. Randerson, N. Faivre, S. Capps, A. Hall, and M. L. Goulden, 2014: Contrasting controls on wildland fires in Southern California during periods with and without Santa Ana winds, *J. Geophys. Res. Biogeosci.*, **119**, 432–450.

Johnson, R. H., R. S. Schumacher, J. H. Ruppert Jr., D. T. Lindsey, J. E. Ruthford, and L. Kriederman, 2014: The Role of Convective Outflow in the Waldo Canyon Fire. *Mon. Wea. Rev.*, **142**, 3061–3080.

Jones, M., B. A. Colle, and J. Tongue, 2007: Evaluation of a short-range ensemble forecast system over the Northeast U.S., *Wea. Forecasting*, **22**, 36-55.

Demargne J., L. Wu, S. K. Regonda, J. D. Brown, H. Lee, M. He, D. J. Seo, R. Hartman, H. D. Herr, M. Fresch, J. Schaake, and Y. Zhu, 2014: The Science of NOAA's Operational Hydrologic Ensemble Forecast Service. *Bull. Amer. Meteor. Soc.*, **95**, 79–98.

Jung, Y., M. Xue, G. Zhang, 2010: Simultaneous Estimation of Microphysical Parameters and the Atmospheric State Using Simulated Polarimetric Radar Data and an Ensemble Kalman Filter in the Presence of an Observation Operator Error. *Mon. Wea. Rev.*, **138**, 539–562.

Kain, J. S., and J. M. Fritsch, 1990: A one-dimensional entraining/detraining plume model and its application in convective parameterization. *J. Atmos. Sci.*, **47**, 2784-2802.

Kaplan, M. L., C. Huang, Y. L. Lin, and J. J. Charney, 2008: The development of extremely dry surface air due to vertical exchanges under the exit region of a jet streak. *Meteor. Atmos. Phys*., **102**, 63-85.

Kleiber, W., A. E. Raftery, J. Baars, T. Gneiting, C. F. Mass, E. Grimit, 2011: Locally Calibrated Probabilistic Temperature Forecasting Using Geostatistical Model Averaging and Local Bayesian Model Averaging. *Mon. Wea. Rev.*, **139**, 2630–2649.

Krzysztofowicz, R. and W. B. Evans, 2008: Probabilistic Forecasts from the National Digital Forecast Database, *Wea. Forecasting*, **23**, 270-289.

Li, H., E. Kalnay, and T. Miyoshi, 2009: Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter. *Quart. J. Roy. Meteor. Soc.*, **135**, 523–533.

Libonati, R., I. Trigo, and C. C. DaCamara, 2008: Correction of 2 m-temperature forecasts using Kalman filtering technique. *Atmos. Res*., **87**, 183–197.

Y. Liu, Z. Liu, S. Zhang, X. Rong, R. Jacob, S. Wu, F. Lu, 2014a: Ensemble-Based Parameter Estimation in a Coupled GCM Using the Adaptive Spatial Average Method, *J. Climate*, **27:11**,4002-4014.

Y. Liu, Z. Liu, S. Zhang, R. Jacob, F. Lu, X. Rong, S. Wu, 2014b: Ensemble-Based Parameter Estimation in a Coupled General Circulation Model, *J. Climate*, **27:18**, 7151-7162.

Lombardo, K. A., and B. A. Colle, 2010: The spatial and temporal distribution of organized convective structures over the northeast United States and their ambient conditions. *Mon. Wea. Rev.,* **138,** 4456–4474.

Luke, R. H., and A. G. McArthur, 1978: *Bushfires in Australia*. Australian Government Publishing Service, Canberra, 359 pp.

Mass, C. F., J. Baars, G. Wedam, E. Grimit, and R. Steed, 2008: Removal of systematic model bias on a model grid. *Wea. Forecasting*, **23**, 438–459.

Mel, R., and P. Lionello, 2014: Storm Surge Ensemble Prediction for the City of Venice. *Wea. Forecasting*, **29**, 1044–1057.

Meng, Z., and F. Zhang, 2008a: Tests of an ensemble Kalman filter for mesoscale and regional-scale data assimilation. Part III: Comparison with 3DVAR in a real-data case study. *Mon. Wea. Rev.*, **136**, 522-540.

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

Miller, S. T. K., and B. D. Keim, 2003: Synoptic-Scale Controls on the Sea Breeze of the Central New England Coast. *Wea. Forecasting*, **18**, 236–248.

Mills G. A. 2005a: A re-examination of the synoptic and mesoscale meteorology of Ash Wednesday 1983. *Australian Meteorological Magazine,* **54**, 35–55.

Mills G. A. 2005b: On the subsynoptic-scale meteorology of two extreme FWDs during the eastern Australian fires of January 2003. *Australian Meteorological Magazine,* **54**, 265–290.

Mills G. A. 2008a: Abrupt surface drying and fire weather. Part 1: overview and case study of the South Australian fires of 11 January 2005. *Australian Meteorological Magazine,* **57**, 299–309.

Mills G. A. 2008b: Abrupt surface drying and fire weather Part 2: a preliminary synoptic climatology in the forested areas of southern Australia. *Australian Meteorological Magazine,* **57**, 311-328.

Mitchell, H. L., and P. L. Houtekamer, 2000: An Adaptive Ensemble Kalman Filter. *Mon. Wea. Rev*., **128**, 416-433.

Mofidi, A., I. Soltanzadeh, Y. Yousefi, A. Zarrin, M. Soltani, J. M. Samakosh, G. Azizi, S. T. K. Miller. 2014: Modeling the exceptional south Foehn event (Garmij) over the Alborz Mountains during the extreme forest fire of December 2005. *Natural Hazards*, **75**, 2489-2518.

Mondal, N., and R. Sukumar, 2014: Characterising weather patterns associated with fire in a seasonally dry tropical forest in southern India. *Int. J. Wildland Fire,* **23**, 196–201.

Miller, 2014: Modeling the exceptional south Foehn event (Garmij) over the Alborz Mountains during the extreme forest fire of December 2005. *Natural Hazards*. **75**, 2489-2518**.**

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

Mölders, N., 2008: Suitability of the Weather Research and Forecasting (WRF) model to predict the June 2005 fire weather for Interior Alaska. *Wea. Forecasting*, **23**, 953-973.

Mölders, N., 2010: Comparison of Canadian Forest Fire Danger Rating System and National Fire Danger Rating System fire indices derived from Weather Research and Forecasting (WRF) model data for the June 2005 Interior Alaska wildfires. *Atmos. Res.,* 95, 290-306.

Müller, M. D., 2011: Effects of Model Resolution and Statistical Postprocessing on Shelter Temperature and Wind Forecasts. *J. Appl. Meteor. Climatol.*, **50**, 1627–1636.

Murray, J. C., B. A. Colle, 2011: The Spatial and Temporal Variability of Convective Storms over the Northeast United States during the Warm Season. *Mon. Wea. Rev.*, **139**, 992–1012.

National Weather Service, 1998: Automated Surface Observing System: User's Guide. 72 pp. [Available online at http://www.nws.noaa.gov/asos/aum-toc.pdf.]

National Ice Center. 2008: updated daily. *IMS Daily Northern Hemisphere Snow and Ice Analysis at 4 km Resolution.* Boulder, Colorado USA: National Snow and Ice Data Center. [Available online at http://dx.doi.org/10.7265/N52R3PMC.]

Nielsen-Gammon, J. W., X. M. Hu, F. Zhang, and J. E. Pleim, 2010: Evaluation of Planetary Boundary Layer Scheme Sensitivities for the Purpose of Parameter Estimation. *Mon. Wea. Rev.*, **138**, 3400–3417.

Northeast States Emergency Consortium, cited 2014: Fires. [Available online at http:// http://nesec.org/fires/.]

Ott, E., and Coauthors, 2004: A local ensemble Kalman filter for atmospheric data assimilation. *Tellus*, **56A**, 415–428.

Pleim, J. E., 2007: A combined local and non-local closure model for the atmospheric boundary layer. Part 2: Application and evaluation in a mesoscale model, *J. Appl. Meteor. Climatol.*, **46**, 1396-1409.

Pollina, J. B., B. A. Colle, J. J. Charney, 2013: Climatology and Meteorological Evolution of Major Wildfire Events over the Northeast United States. *Wea. Forecasting*, **28**, 175–193.

Potter, B. E. 2012: Atmospheric interactions with wildland fire behaviour - I. Basic surface interactions, vertical profiles and synoptic structures. *Int. J. Wildland Fire*, **21**, 779-801.

Preisler, H. K., R. E. Burgan, J. C. Eidenshink, J. M. Klaver, and R. W. Klaver, 2009: Forecasting distributions of large federal-lands fires utilizing satellite and gridded weather information. *Int. J. Wildland Fire*, **18**, 508–516.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Mon. Wea. Rev.*, **133**, 1155-1174.

Rousseeuw, P. J. and A. M. Leroy, 1987: *Robust Regression and Outlier Detection*. WileyInterscience, New York (Series in Applied Probability and Statistics), ISBN 0-471-85233-3. 329 pp.

Salamanca, F., M. Georgescu, A. Mahalov, M. Moustaoui, M. Wang, and B. M. Svoma, 2013: Assessing summertime urban air conditioning consumption in a semiarid environment, *Environ. Res. Lett.,* **8**:3, 034022.

Schmeits, M. J., and K. J. Kok, 2010: A Comparison between Raw Ensemble Output, (Modified) Bayesian Model Averaging, and Extended Logistic Regression Using ECMWF Ensemble Precipitation Reforecasts. *Mon. Wea. Rev.*, **138**, 4199–4211.

Schroeder, M. J., M. Glovinsky, V. F. Hendricks, F. C. Hood, M. K. Hull, H. L. Jacobson, R. Kirkpatrick, D. W. Krueger, L. P. Mallory, A. G. Oertel, R. H. Reese, L. A. Sergius, and C. E. Syverson. 1964. Synoptic Weather Types Associated with Critical Fire Weather. U.S. Forest Service, Pacific Southwest Range and Experiment Station.

Schwitalla, T. H., H. S. Bauer, V. Wulfmeyer, G. Z. Zangl, 2008: Systematic errors of QPF in low-mountain regions as revealed by MM5 simulations. *Meteorol. Z.* **17**, 903–919.

Seaman, N. L. and S. A. Michelson, 2000: Mesoscale Meteorological Structure of a High-Ozone Episode during the 1995 NARSTO-Northeast Study. *J. Appl. Meteor.*, **39**, 384-398.

Simard, A. J., J. E. Eenigenburg, and S. L. Hobrla, 1987: Predicting extreme fire potential. *Proceedings of the Ninth Conference on Fire and Forest Meteorology*, Amer. Meteor. Soc., San Diego, California, 148-157.

Simpson, C. C, H. G. Pearce, V. Clifford, 2013a: High-resolution WRF simulation of fire weather associated with the Mt Cook Station fire. *20th International Congress on Modelling and Simulation*, Adelaide, Australia, 277-283.

Simpson C. C., H. G. Pearce., A. P. Sturman, P. Zawar–Reza, 2013b: Verification of WRF modelled fire weather in the 2009-10 New Zealand fire season. *Int. J. Wildland Fire*. **23(1),** 34-45.

Simpson C. C., H. G. Pearce., A. P. Sturman, P. Zawar–Reza, 2014: Behaviour of fire weather indices in the 2009−10 New Zealand wildland fire season. 2014: *Int. J. Wildland Fire.* **23(8)**,1147-1164.

Skamarock, W., J. B. Klemp, J. Dudhia, D. O. Gill, D. Barker, M. G. Duda, X. -Y. Huang, and W. Wang, 2008: A Description of the Advanced Research WRF Version 3. NCAR Technical Note NCAR/TN-475+STR, DOI: 10.5065/D68S4MVH.

Sloughter, J. M., A. E. Raftery, and T. Gneiting, 2007: Probabilistic Quantitative Precipitation Forecasting Using Bayesian Model Averaging. *Mon. Wea. Rev.*, **135**, 3209-3220.

Sloughter, J. M., T. Gneiting, and A. E. Raftery, 2010: Probabilistic Wind Speed Forecasting using Ensembles and Bayesian Model Averaging. *Journal of the American Statistical Association*, **105**, 25-35.

Stauffer, D. R., N. L. Seaman, and F. S. Binkowski, 1991: Use of four-dimensional data assimilation in a limited-area mesoscale model. Part II: Effects of data assimilation within the planetary boundary layer. *Mon. Wea. Rev.*, **119**, 734-754.

Stensrud, D. J., and N. Yussouf, 2003: Short-Range Ensemble Predictions of 2-m Temperature and Dewpoint Temperature over New England. *Mon. Wea. Rev.*, **131**, 2510–2524.

Stensrud, D. J., and N. Yussouf, 2005: Bias-corrected short-range ensemble forecasts of near surface variables. *Meteor. Appl.*, **12**, 217–230.

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

Stroud, J. R., and T. Bengtsson, 2007: Sequential State and Variance Estimation within the Ensemble Kalman Filter. *Mon. Wea. Rev.*, **135**, 3194–3208.

Svoboda, M. S., D. LeComte, M. Hayes, R. Heim, K. Gleason, J. Angel, B. Rippey, R. Tinker, M. Palecki, D. Stooksbury, D. Miskus, and S. Stephens, 2002: The Drought Monitor. *Bull. Amer. Meteor. Soc.*, **83**, 1181–1190.

Takle, E. S., D. J. Bramer, W. E. Hellman, and M. R. Thompson, 1994: A synoptic climatology for forest fires in the NEUS and future implications from GCM simulations. *Int. J. Wildland Fire*, **4**, 217–224.

Talagrand, O. and P. Courtier, 1987: Variational Assimilation of Meteorological Observations With the Adjoint Vorticity Equation. I: Theory. *Quart. J. Roy. Meteor. Soc.*, **113**, 1311–1328.

Tanamachi, R. L., H. B. Bluestein, M. Xue, W.-C. Lee, K. A. Orzel, S. J. Frasier, and R. M. Wakimoto, 2013: Near-surface vortex structures in a tornado and a sub-tornado-strength, convective-storm vortex observed by a mobile, W-band radar during VORTEX2. *Mon. Wea. Rev.*, **141**, 3661-3690.

Tong, M., and M. Xue, 2008a: Simultaneous estimation of microphysical parameters and atmospheric state with simulated radar data and ensemble square root Kalman filter. Part I: Sensitivity analysis and parameter identifiability. *Mon. Wea. Rev*., **136**, 1630–1648.

Tong, M., and M. Xue, 2008b: Simultaneous estimation of microphysical parameters and atmospheric state with simulated radar data and ensemble square root Kalman filter. Part II: Parameter estimation experiments. *Mon. Wea. Rev*., **136**, 1649–1668.

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

Van Wagner, C. E., 1987: Development and structure of the Canadian Forest Fire Weather Index System. Canadian Forestry Service, Headquarters, Ottawa. Forestry Technical Report 35. 35 pp.

Vrugt, J. A., C. G. H. Diks, and M. P. Clark, 2008: Ensemble Bayesian model averaging using Markov chain Monte Carlo sampling. *Environ. Fluid Mech*., **134**, 1–17.

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

Weygandt, S. S., M. Hu, T. G. Smirnova, H. Lin, J. Olsen, E. James, J. D. Brown, D. Dowell, G. A. Grell, J. Kenyon, I. Jankov, B. Moninger, T. L. Smith, B. Jamison, S. Peckham, K. J., K. Brundage, G. Manikin, G. DiMego, B. Cosgrove, 2014: Hourly updated models: Rapid Refresh / HRRR review. NCEP Production Suite Overview. [Available online at: http://ruc.noaa.gov/pdf/NCEP_PSR_2014_HRRR_COMBINED_lite.pdf.]

Whitaker, J. S. and T. M. Hamill, 2002: Ensemble Data Assimilation without Perturbed Observations. *Mon. Wea. Rev.*, **130**, 1913-1924.

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

Wilks, D.S., 2011: *Statistical Methods in the Atmospheric Sciences*, 3rd Ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.

Wilks, D.S., 2009: Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteorological Applications* **16**:3, 361-368.

Wilson, L. J., S. Beauregard, A. E. Raftery, and R. Verret, 2007: Calibrated Surface Temperature Forecasts from the Canadian Ensemble Prediction System Using Bayesian Model Averaging. *Mon. Wea. Rev.*, **135**, 1364-1385.

Xue, M., Y. Jung, and G. Zhang, 2010: State estimation of convective storms with a two-moment microphysics scheme and an ensemble Kalman filter: Experiments with simulated radar data. *Quart. J. Roy. Meteor. Soc*., **136**, 685-700.

Yahya, K., K. Wang, M. Gudoshava, T. Glotfelty, and Y. Zhang, 2015: Application of WRF/Chem over North America under the AQMEII Phase 2: Part I. Comprehensive evaluation of 2006 simulation, *Atmos. Environ.*, online first, doi:10.1016/j.atmosenv.2014.08.063, *in press*.

Yarnal, B., 1993: *Synoptic Climatology in Environmental Analysis: A Primer*. Belhaven Press, 256 pp.

Yulia R. G., 2007: Comparative Analysis of the Local Observation-Based (LOB) Method and the Nonparametric Regression-Based Method for Gridded Bias Correction in Mesoscale Weather Forecasting. *Wea. and Forecasting*, **22**:6, 1243-1256.

Yussouf N., E. R. Mansell, L. J. Wicker, D. M. Wheatley, and D. J. Stensrud, 2013: The Ensemble Kalman Filter Analyses and Forecasts of the 8 May 2003 Oklahoma City Tornadic Supercell Storm Using Single- and Double-Moment Microphysics Schemes. *Mon. Wea. Rev.*,**141**, 3388–3412.

Zhang, F., C. Snyder, and J. Sun, 2004: Impacts of initial estimate and observation availability on convective-scale data assimilation with an ensemble Kalman filter. *Mon. Wea. Rev.,* **132**, 1238-1253.

Zhang, Y., F. Zhang, M. Zhiyong, and D. Stensrud, 2014: Ensemble Assimilation and Predictability of a Tornadic Supercell Event. *The 6th EnKF Workshop*. Buffalo, NY [Available online at: http://hfip.psu.edu/fuz4/EnKF2014/Day4/Moore_Tornado_DA_Computation_EnKF2014.pptx]

Zsoter, E., R. Buizza, and D. Richardson, 2009: 'Jumpiness' of the ECMWF and UK Met Office EPS control and ensemble-mean forecasts'. *Mon. Wea. Rev.*, **137**, 3823-3836.