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

MICHAEL J. ERICKSON AND BRIAN A. COLLE

School of Marine and Atmospheric Sciences, Stony Brook University, Stony Brook, New York

JOSEPH J. CHARNEY

Northern Research Station, USDA Forest Service, East Lansing, Michigan

(Manuscript received 6 December 2011, in final form 14 May 2012)

ABSTRACT

The performance of a multimodel ensemble over the northeast United States is evaluated before and after applying bias correction and Bayesian model averaging (BMA). The 13-member Stony Brook University (SBU) ensemble at 0000 UTC is combined with the 21-member National Centers for Environmental Prediction (NCEP) Short-Range Ensemble Forecast (SREF) system at 2100 UTC. The ensemble is verified using 2-m temperature and 10-m wind speed for the 2007–09 warm seasons, and for subsets of days with high ozone and high fire threat. The impacts of training period, bias-correction method, and BMA are explored for these potentially hazardous weather events using the most recent consecutive (sequential training) and most recent similar days (conditional training). BMA sensitivity to the selection of ensemble members is explored. A running mean difference between forecasts and observations using the last 14 days is better at removing temperature bias than is a cumulative distribution function (CDF) or linear regression approach. Wind speed bias is better removed by adjusting the modeled CDF to the observation. High fire threat and ozone days exhibit a larger cool bias and a greater negative wind speed bias than the warm-season average. Conditional bias correction is generally better at removing temperature and wind speed biases than sequential training. Greater probabilistic skill is found for temperature using both conditional bias correction and BMA compared to sequential bias correction with or without BMA. Conditional and sequential BMA results are similar for 10-m wind speed, although BMA typically improves probabilistic skill regardless of training.

1. Introduction

Operational ensembles are used to quantify forecast uncertainty by employing different initial conditions, physical parameterizations, and model cores. In this manner, the spread of model solutions creates a representative sample of all possible future outcomes from which probabilistic forecasts can be derived. Unfortunately, many ensemble members exhibit biases (Colle et al. 2003; Eckel and Mass 2005; Jones et al. 2007) that remain in the non-bias-corrected ensemble mean. Various bias-correction techniques are effective at removing the average ensemble bias; including the running mean bias removal (Stensrud and Yussouf 2003; Stensrud and Yussouf 2005; Eckel and Mass 2005), the multivariate regression model (Glahn and Lowry 1972; Mao et al. 1999; Wilson and Vallée 2002; Wilson and Vallée 2003; 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).

Planetary boundary layer (PBL) physics errors contribute significantly to lower atmospheric model error (Pleim 2007; Hu et al. 2010; Nielsen-Gammon et al. 2010), suggesting that bias correction may be important when forecasting phenomena sensitive to near-surface meteorological conditions. However, even after bias correction, many ensembles exhibit insufficient spread [i.e., are underdispersed; Eckel and Mass (2005); Jones

Corresponding author address: Michael Erickson, School of Marine and Atmospheric Sciences, Stony Brook University, Stony Brook, NY 11794-5000.
E-mail: mjaerickson@gmail.com

DOI: 10.1175/WAF-D-11-00149.1

© 2012 American Meteorological Society

et al. (2007)], which is worse in the lower atmosphere (Hamill and Colucci 1997). There is a long history of ensemble postprocessing 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). For this study, BMA is used to calibrate a multimodel ensemble, since it is competitive with other calibration methods (Marrocu and Chessa 2008). BMA, which will be described in more detail below, corrects for ensemble underdispersion by estimating each member's contributing weight and forecast uncertainty.

High wildfire threat and air pollution transport and diffusion are examples of weather phenomena sensitive to near-surface variables over the northeast United States. Although wildfires are less common in the northeast United States compared to other parts of the country (National Interagency Fire Center 2010), they pose a threat to lives and property due to a higher population density. Summer high ozone in the densely populated northeast U.S. area can be harmful to humans (Kinney and Ozkaynak 1991). Accurate simulation of the near-surface weather conditions associated with high fire threat and ozone days is desirable, since meteorological model fields are used to produce fire weather forecasts and to drive air quality models (AQMs).

Previous fire weather modeling studies typically focused on particular events, such as the New Jersey Double Trouble State Park wildfire (Kaplan et al. 2008; Charney and Keyser 2010) and Australian wildfire events (Mills 2005a,b; Mills 2008a,b), or over an entire fire season for a region such as the Pacific Northwest (Hoadley et al. 2004, 2006) and Alaska (Mölders 2008, 2010). While Mölders (2008, 2010) investigates the use of ensemble techniques to enhance fire weather forecasts, bias correction is not an element of either study. Schroeder et al. (1964) and Takle et al. (1994) investigate the synoptic atmospheric patterns associated with high fire threat, but without employing numerical weather prediction models for a more comprehensive analysis. Potential differences in meteorological model biases between these synoptic patterns and the seasonal averages have not been investigated.

There is a long history of employing meteorological models to drive AQMs in the literature [see, e.g., Seaman (2000) for a review of previous studies]. Several recent AQM studies demonstrate how ensemble-forecasting techniques coupled with bias correcting the pollutant of interest can increase forecasting skill (McKeen et al. 2005; Delle Monache et al. 2006; Wilczak et al. 2006; Kang et al. 2008; Delle Monache et al. 2008; Kang et al. 2010). Furthermore, high ozone days over the northeast United States are linked to key synoptic surface circulation patterns (Hegarty et al. 2007), such as an offshore Bermuda high. However, ensemble performance between high ozone days and the seasonal average are largely unknown.

There is growing evidence that model biases may be related to the atmospheric flow pattern. Greybush et al. (2008) find 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-square 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 Weather Research and Forecasting Model (WRF) forecasts to the current forecast to improve 10-m wind speed predictions by 20%–25% over the western United States compared to nonanalog bias-correction methods.

Although BMA generates calibrated forecasts when verified over several consecutive months (Raftery et al. 2005), its performance has not been tested exclusively for specific types of weather phenomena. Training BMA with the most recent few weeks may not remove model biases or improve ensemble dispersion if these phenomena are sensitive to the synoptic flow regime. This is particularly concerning for fire weather and poor air quality days, which are likely associated with specific flow patterns over the northeast United States (Takle et al. 1994; Gaza 1998; Hegarty et al. 2007). Furthermore, BMA has not been tested over the highly populated and meteorologically complex northeast U.S. region. This region has variable land-use characteristics and is frequently influenced by different source air masses (Green and Kalkstein 1996). Additionally, the northeast United States 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). The effectiveness of a bias-correction method may also depend on the model variable being considered.

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). The multimodel ensemble in this study considers several model cores, different model physics, and different initial conditions. However, 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.

This study will address the following motivational questions:

* How do three different bias-correction methods perform for 10-m wind speed and 2-m temperature over a portion of the northeast United States?
* Can ensemble error be reduced for high ozone and high fire threat events using a similar day (conditional) bias-correction approach?
* Can conditional training with BMA improve results on high ozone and fire threat days?
* How sensitive is BMA to the subset of members selected from the total ensemble?

Although other calibration methods may give different results, a major goal of this paper is to compare different training period approaches rather than calibration methods on high fire threat and ozone days.

Section 2 of this paper describes the multimodel ensemble used, the selection criterion for high fire threat and ozone days, and the postprocessing methods. Section 3 presents the bias correction and BMA results for the hazardous weather days and the sensitivity of postprocessing to the members selected in a multimodel ensemble. Section 4 summarizes the results and makes suggestions for future work.

2. Data and methods

a. Selection of high-impact weather events

High fire threat days are classified using 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 data are 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. For this study, an FPI value of 50 or an NFDRS rating of ''high'' is considered to be a high fire threat event. A high fire threat day must have at least 10% of the inner domain (Fig. 1) experiencing high fire threat, while the remainder of the inner domain has moderate fire threat (FPI 21–50). Although it would be useful operationally to select a more extreme fire threat category, this would require several extra years of data to determine statistical significance. High ozone days are identified using in situ measurements from the Environmental Protection Agency (EPA) Aerometric Information Retrieval Now (AIRNow; EPA 1999) network over the northeast United States (Fig. 1). A high ozone day must have 10% or more of the regional stations measuring an air quality index (AQI) of yellow or above [daily maximum 8-h averaged ozone concentrations exceed 59 parts per billion (ppb)], while all other stations must exceed 30 ppb.

b. Multimodel ensemble

The Stony Brook University (SBU) and National Centers for Environmental Prediction (NCEP) Short Range Ensemble Forecast (SREF) ensembles are evaluated over the northeast U.S. domain shown in Fig. 1 for the 2007–09 warm seasons (April–September). The SBU ensemble consists of seven fifth-generation Pennsylvania State University–National Center for Atmospheric Research (NCAR) Mesoscale Model (MM5; Grell et al. 1994) members and 6 WRF (Skamarock et al. 2008) members at 36- and 12-km grid spacing domains covering the eastern two-thirds of the United States and the northeast United States, 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 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 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° 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 Advanced Research WRF (ARW)]. Initial condition perturbations for the SREF implement a breeding technique similar to that in Toth and Kalnay (1997), and each core uses different physics (Table 1). Except for the Eta and RSM, SREF members sharing the same core can be considered exchangeable.

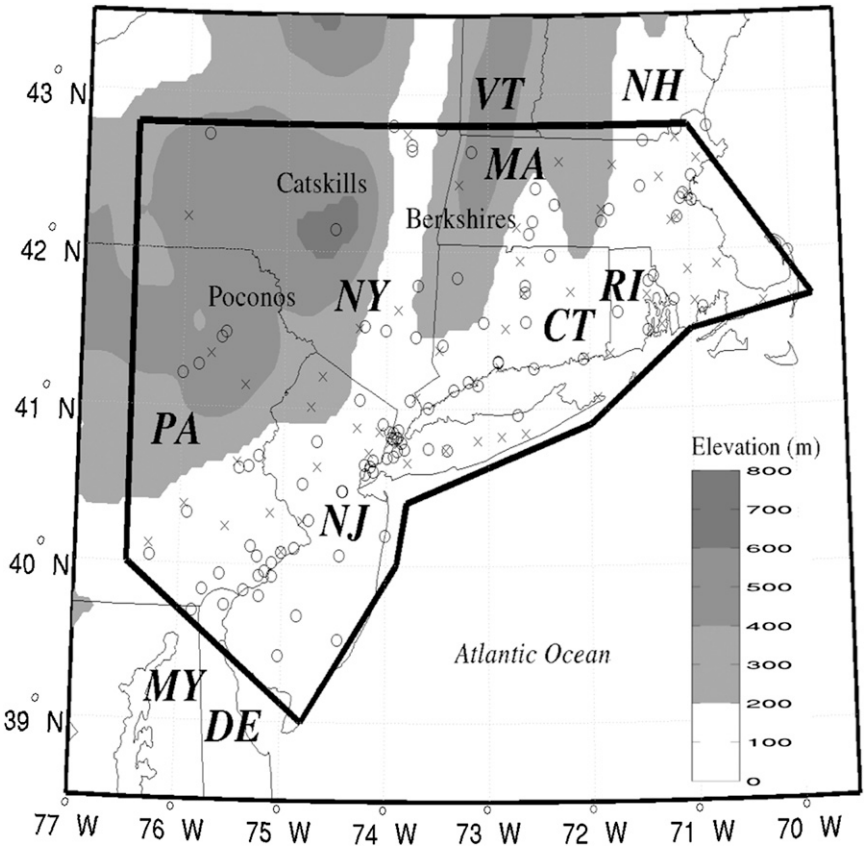

FIG. 1. Domain used for the verification of high ozone and fire threat days over the northeast United States (black border) and the ASOS stations (exes) and AIRNow stations (circles). Shaded contours denote elevation while the text labels the mountainous regions of the Poconos, Catskills, and Berkshires. 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).

The SBU and SREF ensembles are verified for the 0000 and 2100 UTC run cycles, respectively, for the region shown in Fig. 1. A subset of the northeast United States is selected based on its high population density and neighboring woodland area (i.e., Catskills, Poconos, Berkshires; Fig. 1). The predicted 2-m temperature and 10-m wind speed are verified with Automated Surface Observing System (ASOS) observations. The 2-m specific humidity is verified for the SBU ensemble only (archived moisture is not available for SREF at the time of this study).

The verification and postprocessing of high-impact weather focuses on the daytime period (1200–0000 UTC), since this is when peak high fire threat or ozone conditions typically occur. Two adjacent forecast cycles are used; the SBU (SREF) 36–48-h (39–51-h) forecast 1 day prior to the high-impact event and the 12–24-h (15–27 h) forecast on the day of the event. Model forecast data at the four grid points surrounding the observation sites are bilinearly interpolated to each ASOS location (Jones et al. 2007).

c. *Bias-correction methods*

Three different bias-correction techniques are compared using the ensemble forecasts: linear regression, additive, and cumulative distribution function (CDF). 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 mean error 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 mean absolute error (MAE) by 5%–10%. 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

TABLE 1. Description of the SBU and SREF ensembles, including model used, microphysical schemes, PBL schemes, radiation schemes, cumulus schemes, and initial conditions.

SBU and SREF						
Members	Model	Microphysics	PBL scheme	Radiation	Cumulus	Initial condition
1	MM5	Simple ice	MY	Chemistry–Climate Model 2 (CCM2)	Betts–Miller	GFS
2	MM5	Reisner	MY	Cloud radiation	Grell	WRF-NMM
3*	MM5	Simple ice	MY	CCM2	Kain–Fritsch	Canadian Meteorological Centre (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	Navy Operational Global Atmospheric Prediction System (NOGAPS)
7*	MM5	Simple ice	Blackadar	CCM2	Grell	GFS
8	WRF-ARW	WRF six-moment, three class (WSM3)	MYJ	Rapid Radiative Transfer Model (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	Geophysical Fluid Dynamics Laboratory (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	Simplified Arakawa–Schubert/ relaxed Arakawa-Schubert (SAS SAS/RAS	GFS

* Includes those members for BMA-only experiments; SREF uses the control member only.

bins the data by terrain elevation (at thresholds between 0 and 50, 50 and 100, 100 and 200, and greater than 200 m) and dominant land-use categories (i.e., urban, mixed forest, water, etc.) to find the 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%–3%.

The sensitivity of the model state variable to bias correction method is explored for temperature and wind speed 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 1 day at a time by incrementally dropping the oldest day from the training period while adding the most recent event. Comparing training lengths of 7, 14, 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 (National Weather Service 1998), and the modeled wind speed. For consistency, all modeled wind speed 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 artificially high model bias.

d. 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

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

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 study applies BMA similar to Raftery et al. (2005), including the assumption that temperature is normally distributed. However, this study separates bias correction from BMA and chooses the best bias-correction method prior to BMA for each model state variable.

BMA for wind speed is implemented in manner similar to that of Sloughter et al. (2007) for precipitation and Sloughter et al. (2010) for maximum wind speed, but with a few adjustments. Sloughter et al. (2007, 2010) assume 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 study. Sloughter et al. (2010) assumes the gamma distribution's mean follows a linear relationship with the forecasted wind speed. This study uses the bias-corrected modeled wind speed 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 wind speed is calm, since it results in unacceptable zero-valued gamma means. Therefore, all zero-valued gamma means are replaced with the average observed wind speed for each ensemble member when the model-forecasted wind speed is calm. In this manner, BMA can still calculate an unbiased posterior PDF.

As in Sloughter et al. (2007), the wind speed 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 wind speed being calm given the forecast. A power transformation is used to make the wind speed data more Gaussian. The BMA fit to observations over the northeast United States is optimized during the daytime hours when the modeled wind speed data are 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. Our study uses the Differential Evolution Adaptive Metropolis (DREAM) Markov Chain Monte Carlo (MCMC) algorithm developed by Vrugt et al. (2008). 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 25–30-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 and 0000 UTC (SBU 12–24- 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, five members are selected from each of the SBU and SREF ensembles. The five best SBU members are selected in terms of lowest seasonally averaged MAE within each PBL scheme (Table 1). The SREF control members are used, since the perturbed members had higher MAEs for temperature and wind speed (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. One thousand 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. Results

a. Bias-correction methodology

Figures 2 and 3 show the performance of different bias correction methods for 2-m temperature and 10-m wind speed, respectively, averaged over all members during the warm season. Ensemble mean 2-m temperature has an increasingly negative mean error (ME) from −1.19 to −1.96 K between the 289- and 304-K thresholds (Fig. 2a). All bias-correction techniques improve the cool bias for all thresholds. However, the linear regression

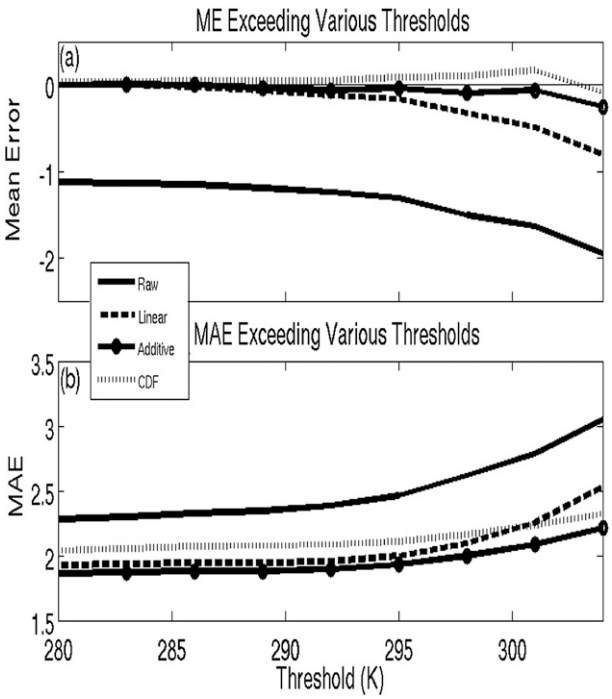

FIG. 2. Surface 2-m temperature (K) for (a) ME and (b) MAE by threshold for raw (solid), linear bias correction (dashed), additive bias correction (filled circles), and CDF bias correction (dotted). Thin black horizontal line in (a) denotes zero bias.

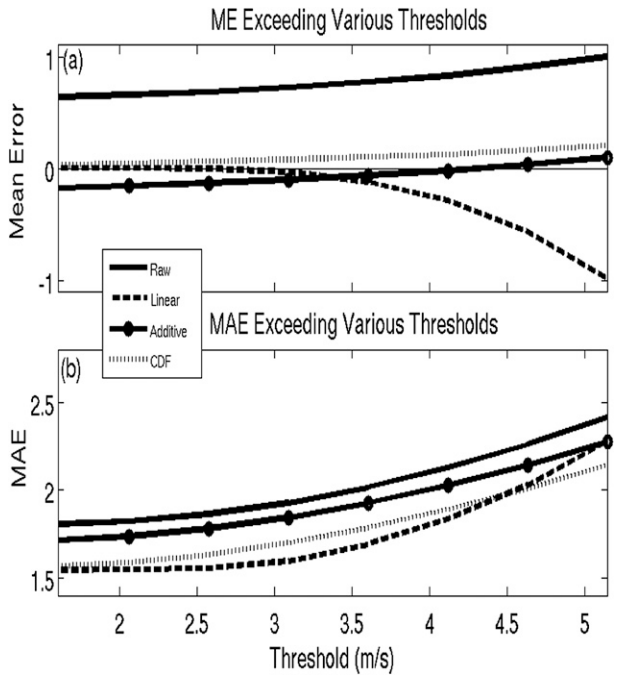

FIG. 3. As in Fig. 2, but for 10-m wind speed (m s^{-1}).

overcorrects for the warmest temperatures 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). The additive bias correction has lower average MAE (1.92 K) than the CDF (2.10 K) and linear techniques (2.02 K; Fig. 2b). The differences in MAE between the linear, CDF, and additive techniques are all statistically significant.

Ensemble mean 10-m wind speed has a positive bias that increases from 0.64 to 1.00 m s^{-1} for thresholds between 1.5 and 5 m s^{-1} (Fig. 3a). The WRF Mellor–Yamada Janjić (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. 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. 3a) 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 in the variance problem as temperature, causing high MAE (2.28 m s^{-1}) at 5 m s^{-1}. Although the linear method performs best in terms of MAE for most thresholds, the CDF method is preferred due to its better performance at higher thresholds. In general, similar results are achieved when using contingency table bias to measure model bias, and equitable threat scores (ETS) in addition to the logarithmic score to measure model skill (Wilks 2011). However, the linear bias-correction method performs worse for wind speed in terms of ETS or logarithmic score than MAE, with the CDF method performing the best for wind speed (not shown).

Accurate forecasts at higher thresholds are important, since strong winds can contribute to the generation and spread of wildfires (Charney and Keyser 2010) and affect the dispersion and advection of pollutants (Seaman and Michelson 2000). Given the results above, the remainder of this paper will bias correct temperature (wind speed) with the additive (CDF) method.

b. *Bias correction applied to fire weather days*

Temperature ME and MAE values are presented for high fire threat days (Fig. 4). Results are grouped according to PBL scheme within the SBU ensemble or SREF model core and averaged for several different periods: the warm season, fire threat days, fire threat days with sequential bias correction, and fire threat days with conditional bias correction. The ensemble mean has a warm-season cool bias of −1.09 K (Fig. 4a), with

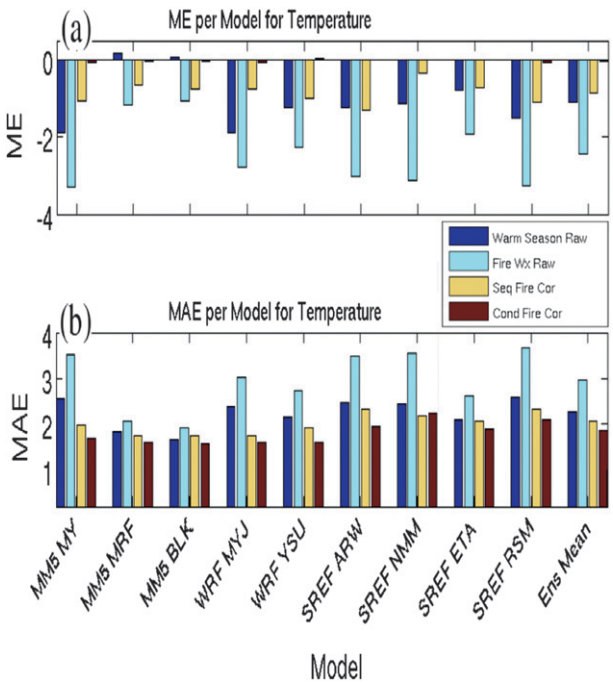

FIG. 4. Histograms of (a) ME and (b) MAE for 2-m temperature (K) for each ensemble member subgroup that shares the same PBL scheme in MM5–WRF and the model core in SREF showing the raw warm season (dark blue), raw high fire threat (light blue), high fire threat with sequential bias correction (yellow), and high fire threat with conditional bias correction (red).

a bias of −2.45 K on fire threat days. The SBU's MM5 MY and WRF MYJ members have the largest SBU cold bias on high fire threat days (−3.29 and −2.77 K, respectively), while the SBU's MM5 Medium-Range Forecast (MRF) model and MM5 Blackadar (BLK) members have cool biases of −1.18 and −1.05 K, respectively. The sequential bias correction on fire threat days has an ensemble mean cool bias (−0.85 K), while conditional bias correction has almost no bias (−0.02 K). The temperature MAE for high fire threat days is reduced for conditional bias correction (1.85 K) compared to sequential training (2.06 K; Fig. 4b). The improvements in the ensemble mean ME and MAE for conditional training are statistically significant at greater than 95% confidence compared to sequential training.

Ensemble-averaged mean error is plotted spatially for the warm-season average (Fig. 5a), high fire threat days (Fig. 5b), sequentially corrected high fire threat days (Fig. 5c), and conditionally corrected high fire threat days (Fig. 5d). This spatial plot is created by interpolating model biases for each station onto a 0.25° × 0.25° latitude–longitude grid. The model and observation must exceed the 298-K threshold for the mean error to be calculated, although results for other thresholds are similar. Warm-season biases are generally negative (−4.0 to 0.4 K), but cool biases are larger on high fire threat days (−7.2 to −1.5 K). Sequential bias correction reduces the spread of potential biases (−4.1 to −0.2 K), but conditional bias correction removes the magnitude and spread of the negative bias more effectively (−2 to 1.2 K).

Warm-season biases with 2-m specific humidity (Fig. 6) vary between 0.04 g kg^{-1} for the WRF Yonsei University (YSU) PBL members to 4.40 g kg^{-1} with the MM5 MY PBL (Fig. 6a). Schwitalla et al. (2008) noted a strong daytime wet bias in specific humidity using the MM5 MY PBL and attributed this bias to weaker vertical mixing in the boundary layer. High fire threat days 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 sequential (conditional) training. The difference between the sequential and conditionally bias-corrected ensemble mean ME and MAE is statistically significant, exceeding 95% confidence levels.

Warm-season model biases vary between ensemble members for 10-m wind speed (Fig. 7). 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 high fire threat days with sequential bias correction has an ME of −0.55 m s^{-1}, which is corrected with conditional training (~0 m s^{-1}). The improvement in MAE between conditional (1.49 K) and sequential bias correction (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 3d).

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

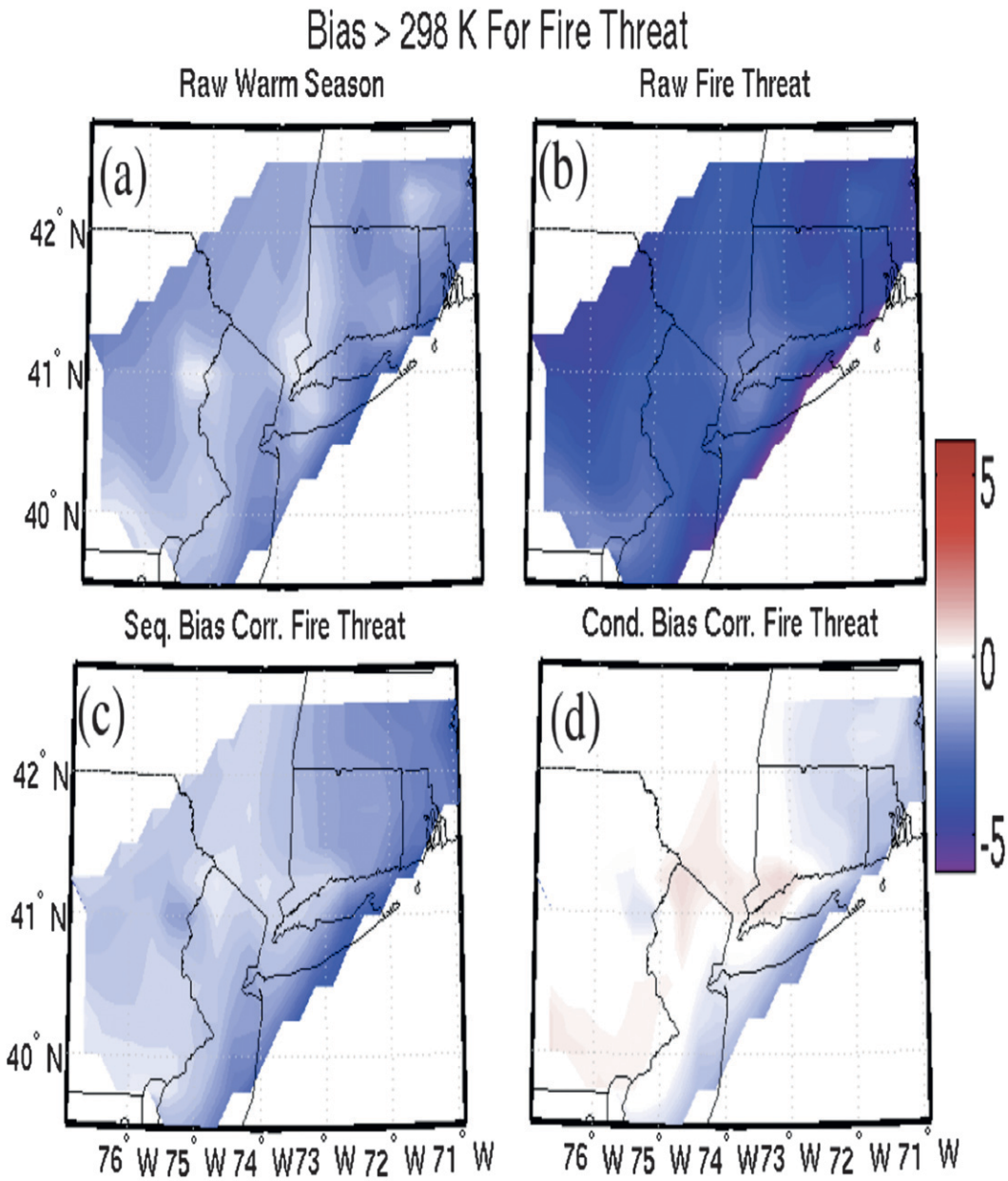

FIG. 5. Spatial 2-m temperature ensemble mean error for the threshold exceeding 298 K for the (a) raw warm season, (b) raw fire threat days, (c) fire threat days with sequential bias correction, and (d) fire threat days with conditional bias correction.

c. *Bias correction applied to high ozone days*

High ozone days have a larger ensemble mean cool bias for 2-m temperature (−1.52 K) than the warm season average (−1.09 K; Fig. 9a). Sequential bias correction results in an ensemble mean bias of −0.42 K, while conditional bias correction removes all but +0.03 K. This difference in ensemble ME is statistically significant above the 95% confidence level, but the ensemble MAE is degraded by 0.01 K (Fig. 9b; i.e., is not significant). The bias in 10-m wind speed on high ozone days shows a statistically significant improvement between conditional and sequential bias correction (by 0.12 m s^{-1}), but not for MAE (degrades by 0.02 m s^{-1}; not shown). Spatial temperature (wind speed) biases for high ozone days are similar to the high fire threat days in Fig. 5 (Fig. 8; not shown).

High fire threat and high ozone datasets are largely independent, with the former occurring predominately in the spring and the latter almost entirely in summer. Of

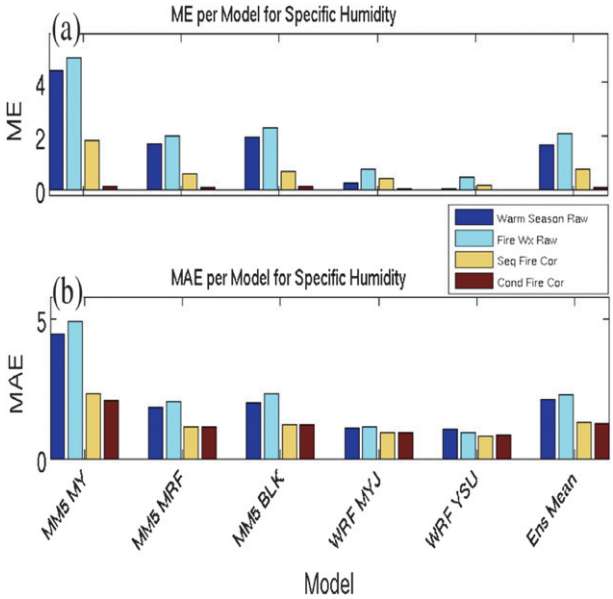

FIG. 6. As in Fig. 4, but for 2-m specific humidity using only the SBU ensemble on high fire threat days.

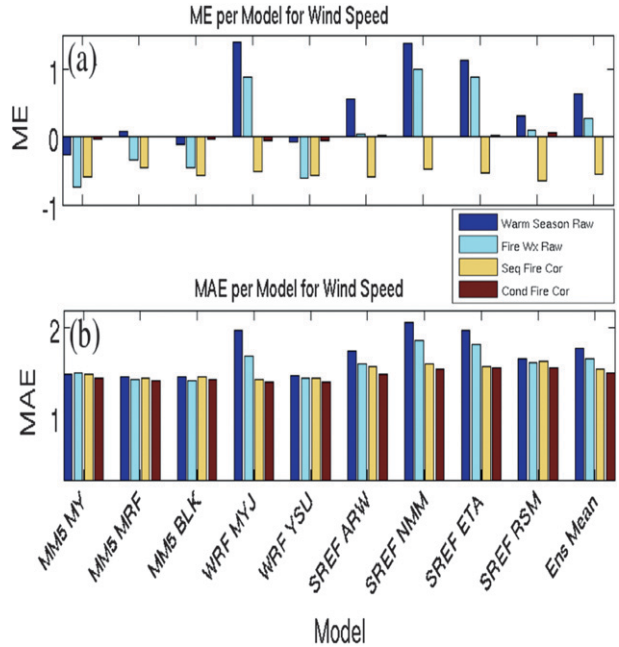

FIG. 7. As in Fig. 4, but for 10-m wind speed on high fire threat days.

the 80 high fire threat model runs and 134 high ozone model runs analyzed, there are only 16 model runs that overlap.

d. Bayesian model averaging

BMA is applied to the hazardous weather days 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 high fire threat and ozone events with BMA-ST and BMA-CT are compared to bias-correction-only rank histograms for 2-m temperature (Fig. 10) and 10-m wind speed (Fig. 11). 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 underdispersed (U-shaped) temperature and wind speed forecasts for all high fire threat and ozone events. Sequential bias correction has a backward L-shaped histogram, indicative of a negative bias. Applying BMA after bias correction removes most of the underdispersion for temperature (Fig. 10) and wind speed (Fig. 11) and also greatly lowers RI index values (by 34.8% for temperature and 55.7% for wind speed, on average). However, BMA-ST is still negatively biased (Figs. 10b,d and 11b) compared to BMA-CT (Figs. 10a,c and 11a) for all high fire threat days and high ozone days with temperature. This is reflected in the lower RI values for conditional versus sequential training, suggesting that BMA cannot correct for lingering model biases.

The influence of BMA is shown probabilistically using reliability plots for high fire threat days exceeding the 4 m s^{-1} and 6 m s^{-1} thresholds (Fig. 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 (Fig. 12). However, the negative bias caused by sequential training on high fire threat days (Figs. 7a and 11b) results in less accurate probabilistic forecasts for BMA-ST (Figs. 12b,d) compared to BMA-CT (Figs. 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 sequential bias correction with no BMA for multiple thresholds. Warm-season high fire threat and high ozone days are considered between 286 and 300 K for temperature and 1.5 and 5.2 m s^{-1} for wind speed. The stations, models, and days considered are identical for bias correction, BMA-ST, and BMA-CT, allowing for a direct comparison among all three postprocessing methods. Most thresholds for temperature (Fig. 13) and all thresholds for wind speed (Fig. 14) have BMA BSS values greater than zero at 95% confidence, demonstrating that BMA improves

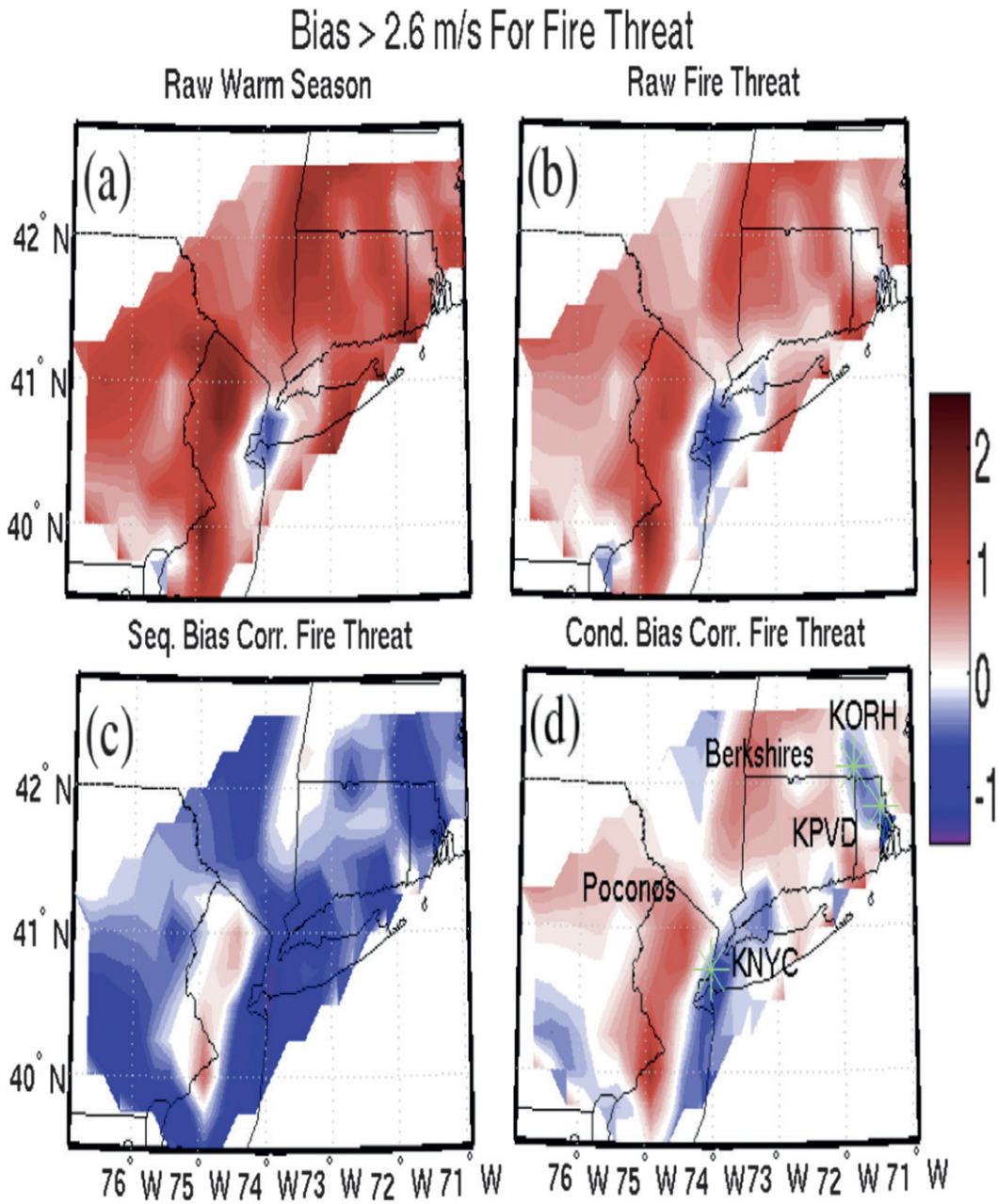

FIG. 8. As in Fig. 5, but for 10-m wind speed on high fire threat days. In Fig. 8d, the green asterisks indicate the locations of KNYC, KPVD, and KORH. The Poconos and Berkshire Mountains are also labeled.

probabilistic scores regardless of training period. Results for temperature on high fire threat days reveal statistically significant improvement in BMA-CT compared to BMA-ST for all thresholds (Fig. 13a). High ozone days have a similar benefit with BMA-CT, although only the 294- and 295-K thresholds are statistically significant (Fig. 13b). There is little difference in wind speed between BMA-CT and BMA-ST on high ozone or fire threat days (Fig. 14).

e. Sensitivity of BMA to forecast hour

The impact of forecast hour on BMA is explored on high fire threat and ozone days by rerunning BMA for 2-m temperature between SBU forecast hours 3 and 48 (6–51-h SREF). To obtain two diurnal cycles for high fire threat, two subsequent model runs are postprocessed and combined. BSS is used to assess BMA performance for 2-m temperature by hour with sequential bias

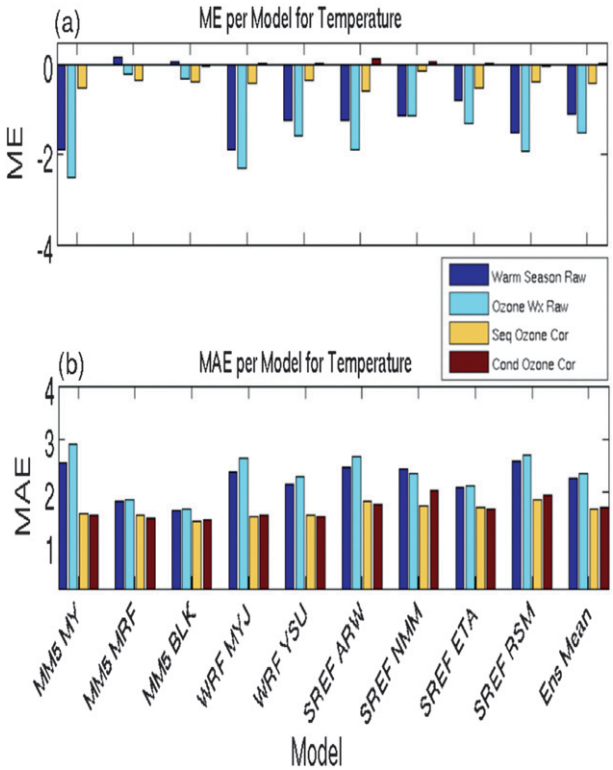

FIG. 9. As in Fig. 4, but for 2-m temperature (K) on high ozone days.

correction as a reference (Fig. 15). The BSS for BMA-CT on high fire threat days varies diurnally and is nearly out of phase with BMA-ST (Fig. 15a). As a result, the probabilistic results with BMA-CT on high fire threat days 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 30 and 33 (both exceeding 95% confidence). The poor performance of BMA-CT compared to BMA-ST between hours 30 and 33 is a result of better performance with the conditional (sequential) bias correction during the day (night). Regardless, both BMA-ST and BMA-CT generally improve upon sequential bias correction. High ozone days with BMA-CT are not statistically significantly different than BMA-ST for 2-m temperature for most hours (Fig. 15b), although both BMA methods are generally better than bias correction alone.

f. 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 sections 3d and 3e may have useful information for BMA.

To test BMA's sensitivity to member selection, five SBU and five SREF members are chosen from the total ensemble as described in section 2 (B10). This ensemble is compared to five SBU and five SREF members that are randomly selected from the total ensemble 1000 times (R10). R10 and B10 are compared for 2-m temperature and 10-m wind speed after bias correction for the 2007–09 warm seasons.

The difference between B10 and R10 in terms of MAE, variance, and BSS (referenced against R10) is calculated by threshold for temperature and wind speed (Fig. 16). B10 has significantly lower MAE for both temperature and wind speed (Figs. 16a,b), but is more underdispersed (Figs. 16c,d). This affects the probabilistic results (Figs. 16e,f), particularly for wind speed, where the R10 ensemble has better probabilistic skill. Comparing B10 and R10 is an interesting test case for ensemble calibration on the hazardous weather days, since BMA should correct for underdispersion while preserving the ensemble's skill.

The sensitivity of BMA to member selection for each hazardous weather day 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 temperature for high fire threat (Fig. 17a) and high ozone days (Fig. 17b) referenced against the sequential bias correction of R10–BMA. B10–BMA is significantly better than R10–BMA on high fire threat days for thresholds between 287 and 289 and 292 K (Fig. 17a). For high ozone days, B10–BMA is significantly better than R10–BMA for all thresholds between 288 and 299 K. 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 five SBU members (B5–SBU–BMA), the five SREF members (B5–SREF–BMA) and a combined ensemble (B5–ALL–BMA) that for each hazardous weather day randomly draws two SBU members, two SREF members, and one member that has an equal chance of being either SBU or SREF.

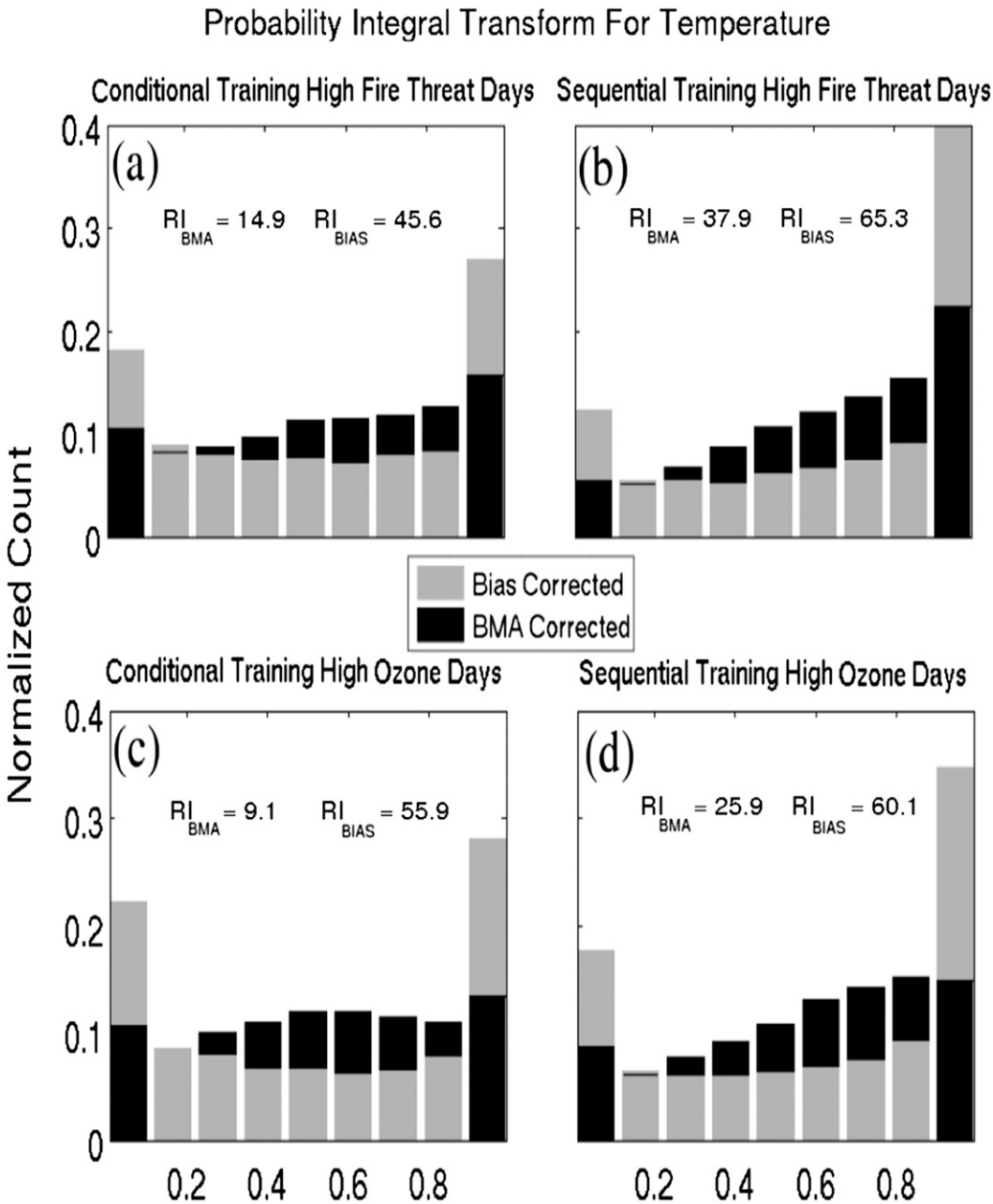

FIG. 10. PIT for 2-m temperature showing bias- (gray) and bias+BMA- (black) corrected prediction during (a),(b) high fire threat and (c),(d) high ozone days using (a),(c) conditional and (b),(d) sequential training. The reliability index for each rank histogram (see text for full description) is also included.

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 values by threshold for high fire threat (Fig. 18a) and high ozone (Fig. 18b) are plotted by threshold for 2-m temperature 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. High fire threat days benefit from a combined SBU–SREF ensemble, except at higher thresholds (greater than 298 K), while the SREF (SBU) performs better than the combined ensemble at lower (higher) thresholds for high ozone days (Fig. 18b).

In general, selecting skillful and unique members from both the SBU and SREF ensembles for use in BMA

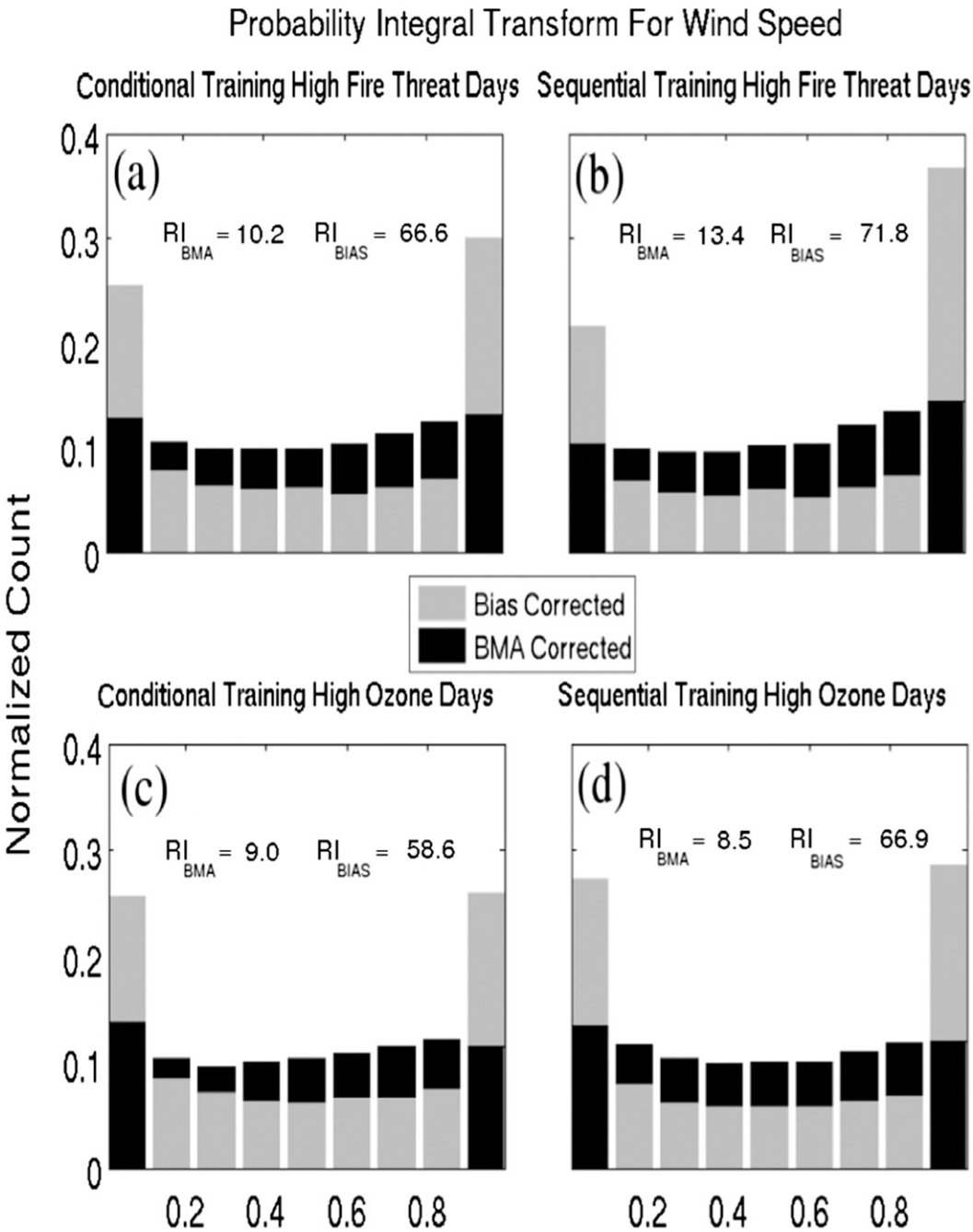

FIG. 11. As in Fig. 10, but for 10-m wind speed.

improves the probabilistic results. This suggests that model diversity is important, even with a calibration method like BMA. Given previous results, the members selected for B10 are 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.

4. Summary and conclusions

The postprocessing of 2-m temperature and 10-m wind speed for the combined 13-member Stony Brook

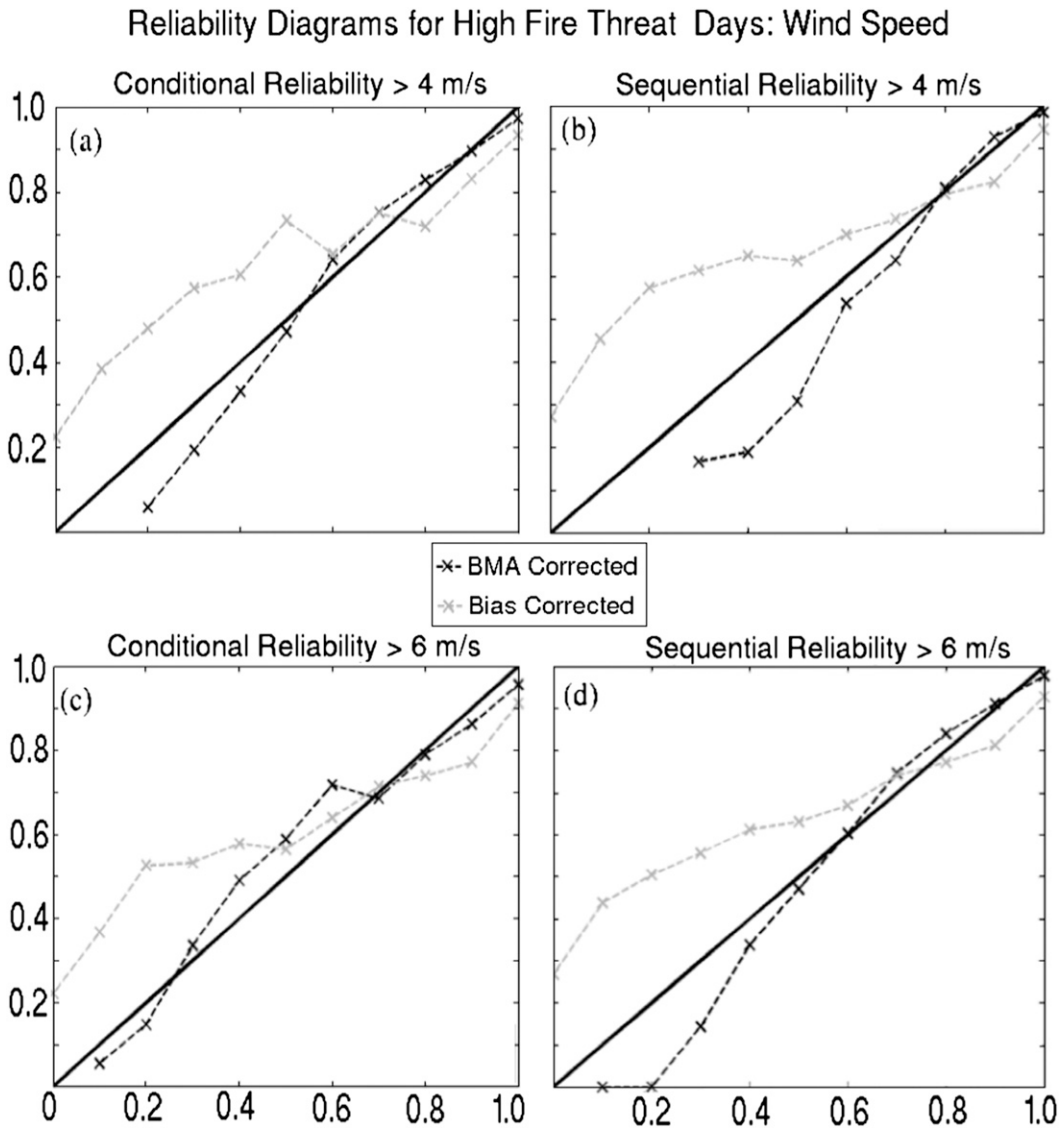

FIG. 12. Reliability diagrams for 10-m wind speed exceeding 4 m s^{-1} for (a) conditional and (b) sequential training and 6 m s^{-1} for (c) conditional and (d) sequential training. The BMA-corrected (black dashed) and bias-corrected (gray dashed) results are compared with the 1:1 line (solid black).

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 postprocessing are explored for high fire threat and ozone days over a subset of the northeast United States. Postprocessing with 2-m specific humidity (SBU ensemble only) is also explored.

High fire threat and ozone days have different model biases for 2-m temperature and 10-m wind speed compared to the warm-season average, which can degrade attempts to calibrate an ensemble. Therefore, bias correction and BMA are implemented using a training window of the 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 (1200–0000 UTC) for two subsequent model runs initialized at 0000 UTC (2100 UTC for the SREF) on the day of and the day before the hazardous weather event. Additionally, BMA uses a smaller ensemble consisting of the five best performing SBU members (in terms of mean absolute error) with a unique PBL and the five SREF control members.

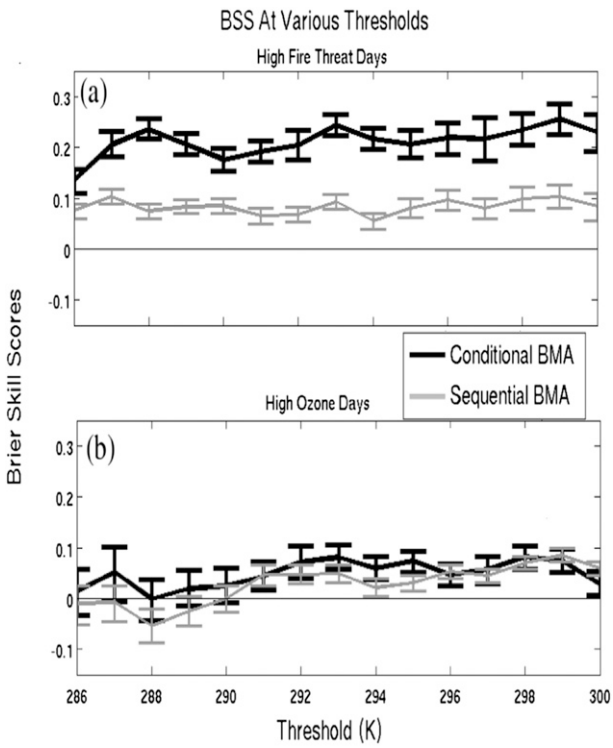

FIG. 13. BSSs for 2-m temperature by threshold for (a) high fire threat days and (b) high ozone days. Sequential (gray) and conditional (black) BMA results are referenced against sequential bias correction. Thin black horizontal line denotes zero BSS.

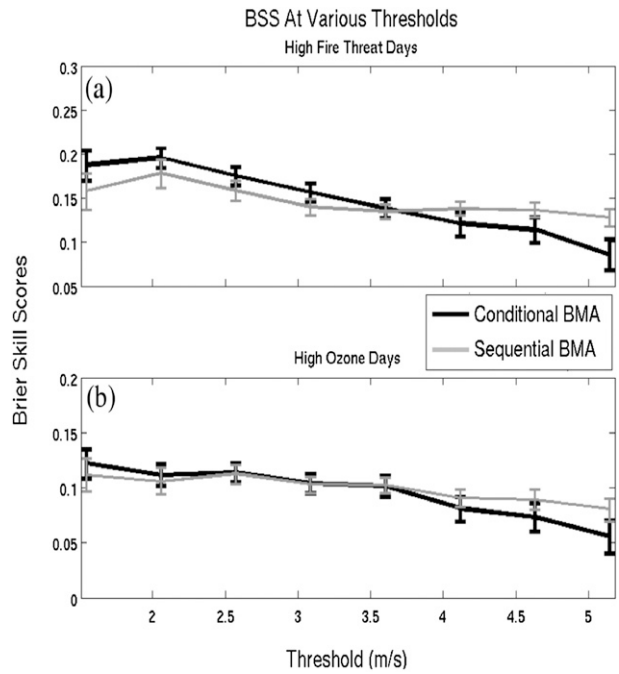

FIG. 14. As in Fig. 13, but for 10-m wind speed.

The optimal bias correction depends on the model state variable. The CDF (additive) bias correction performs best for 10-m wind speed (2-m temperature). On average, simulated high fire threat days are cooler, moister, and less windy than the typical warm-season average bias. High ozone days exhibit a cooler temperature bias than the warm-season average. As a result, conditional bias correction results in significant improvements in mean error over sequential training for high ozone (with temperature and wind speed) and fire threat days (with temperature, wind speed, and specific humidity). There are also statistically significant improvements in MAE on high fire threat days for 2-m temperature and 2-m specific humidity.

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 high fire threat days for 2-m temperature. Wind speed results with BMA-CT and BMA-ST are similar probabilistically for high fire threat and ozone events. Future work is needed to determine the lack of any conditional probabilistic benefit for wind speed on high fire threat days, which might be improved by adjusting how wind speed is implemented with BMA. Regardless of the model state variable or high-impact event, BMA-ST and BMA-CT almost always improve the ensemble probabilistic value compared to sequential

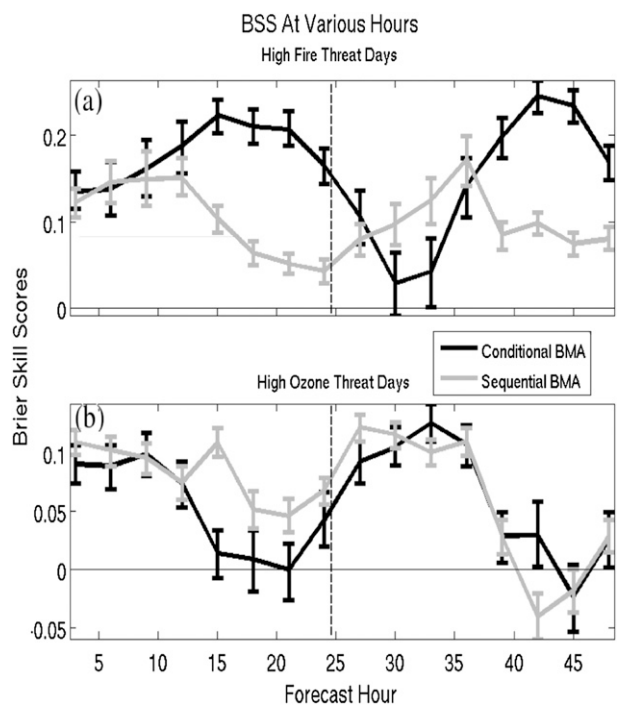

FIG. 15. BSSs for 2-m temperature by forecast hour for (a) high fire threat days and (b) high ozone days. Sequential (gray) and conditional (black) BMAs are referenced against sequential bias correction. The thin black horizontal line denotes zero BSS, while the dashed black vertical line separates the day 0 forecast from the day 1 forecast.

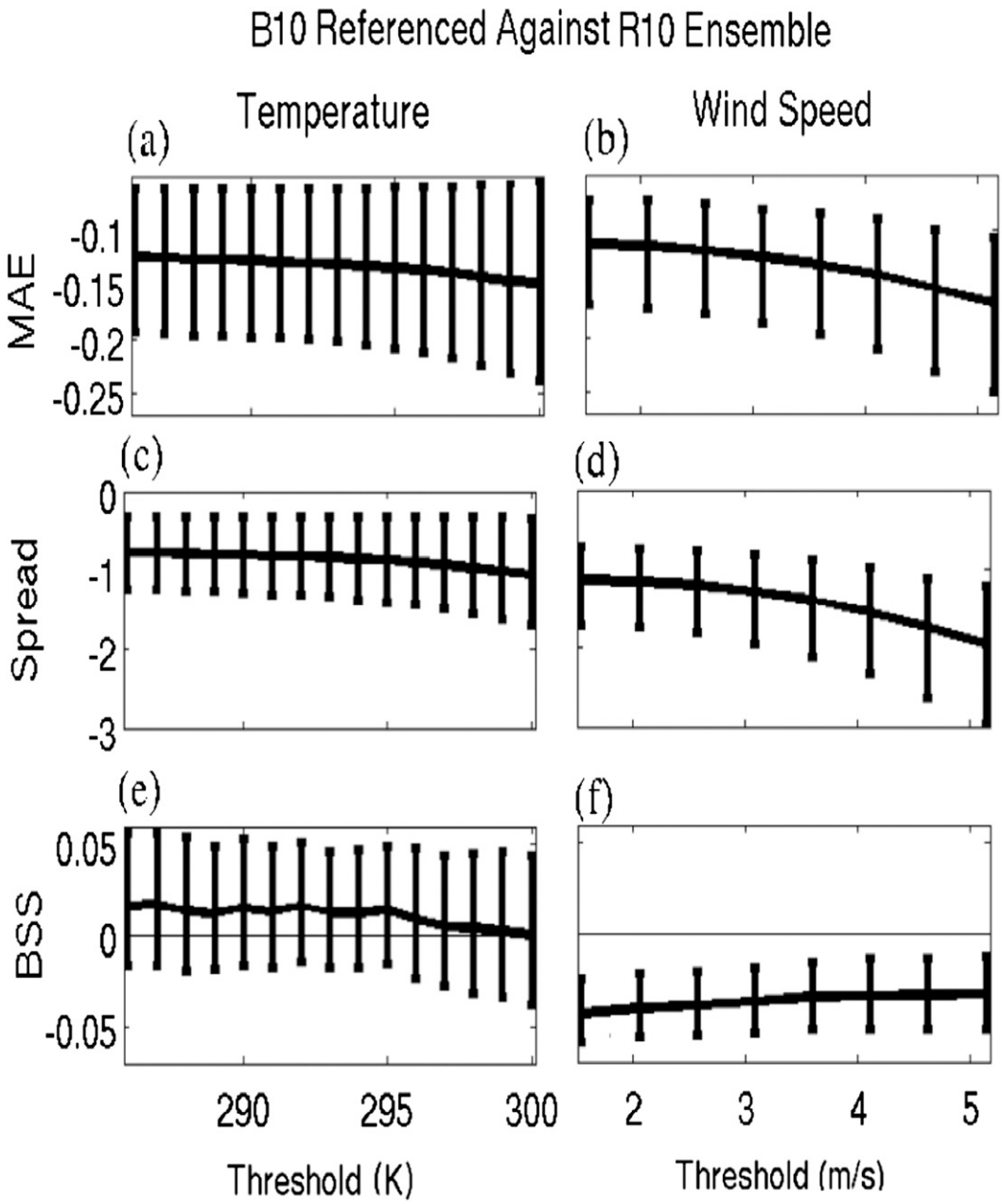

FIG. 16. Difference between R10 and B10 for (a) MAE, (c) ensemble spread, and (e) BSS (referenced against R10) for warm-season 2-m temperature. (b),(d),(f) As in (a),(c),(e) but for 10-m wind speed.

bias correction 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 high fire threat and ozone days varies as a function of forecast hour. For high fire threat days, this is likely caused by a more effective conditional bias correction during the daytime hours. This suggests a potential physics problem in the lower level of the modeled atmosphere in the presence of shortwave radiation, perhaps caused by the parameterized planetary boundary layer, land surface model, or radiation scheme. For high ozone days, the difference between BMA-CT and BMA-ST is generally not statistically significant. 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).

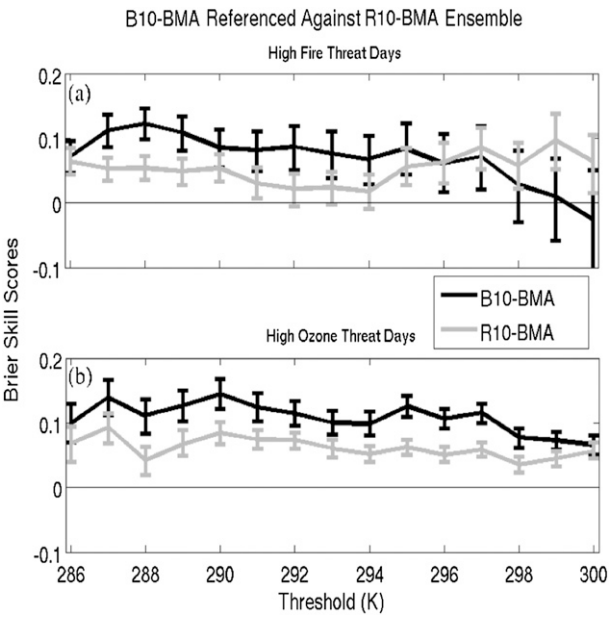

FIG. 17. BSSs for warm-season 2-m temperature by threshold for (a) high fire threat days and (b) high ozone days. B10–BMA (black) and R10–BMA (gray) are referenced against R10 with sequential bias correction. The thin black horizontal line denotes zero BSS.

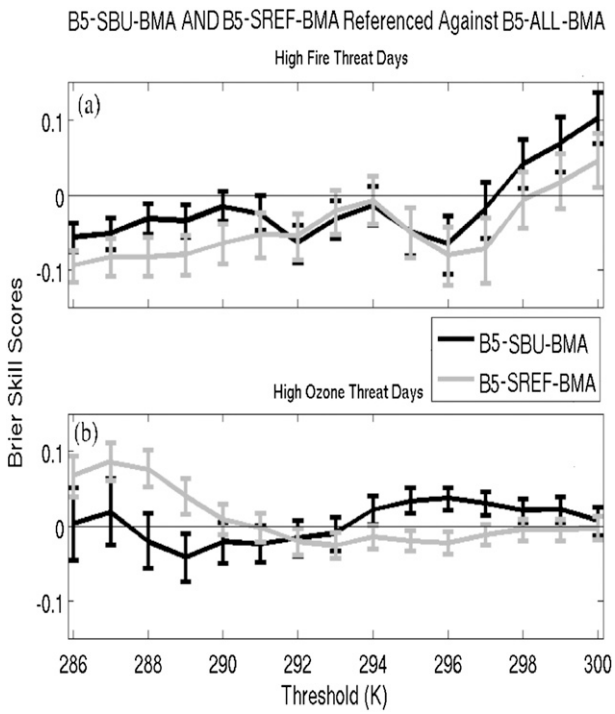

FIG. 18. BSSs for warm-season 2-m temperature by threshold for (a) high fire threat days and (b) high ozone days. B5–SBU–BMA (black) run with BMA and B5–SREF–BMA run (gray) with BMA are referenced against B5–ALL–BMA. The thin black horizontal line denotes zero BSS.

Over the warm-season average, B10 performs better in terms of MAE for wind speed and temperature, 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 high fire threat and ozone day (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 postprocessing.

The SBU and SREF ensembles are combined in this study, since additional model cores are expected to bring about greater probabilistic accuracy. This hypothesis is tested by comparing two ensembles consisting of five SBU and five 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). High fire threat postprocessing with temperature benefits from the combined SBU and SREF ensemble, while results are more mixed with high ozone days.

The presence of conditional model biases on days important to human health and safety underscores the importance of similar day postprocessing. Although complex event-based postprocessing has been developed (Hamill and Whitaker 2006; Mass et al. 2008; Delle Monache et al. 2011), methods designed to capture high fire threat or ozone days may be more beneficial than selecting days of similar temperature and time of year. It is important to be able to predict high fire threat and ozone days operationally and understand why model biases vary temporally.

Fortunately, atmospheric models when coupled with air quality models (AQMs) show promising skill for predicting ozone, especially after bias correction (McKeen et al. 2005; Wilczak et al. 2006; Delle Monache et al. 2006; Delle Monache et al. 2008; Kang et al. 2008; Kang et al. 2010). For instance, Kang et al. (2010) shows a category hit rate greater than 90% for bias-corrected ozone exceeding 59 ppb (the high ozone day threshold in this study) over the United States. Furthermore, probabilistic ozone concentration can be inferred through ensemble modeling of AQMs (McKeen et al. 2005; Delle Monache et al. 2006; Wilczak et al. 2006; Delle Monache el al. 2008).

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 high fire threat bias-correction method that ''turns on'' when high fire threat conditions (i.e., defined by widespread low humidity and strong winds) are predicted within the next 48 h. This threshold would have to be adjusted based on the original model biases.

Since high fire threat days 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 high fire threat days 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 will be a topic of future work.

Acknowledgments. This work was supported by a research joint venture agreement between Stony Brook University and the U.S. Forest Service (08-JV-11242306-093) and the New York State Energy Research and Development Authority (Agreement 10599). We appreciate Christian Hogrefe's suggestions to improve this manuscript, Prakash Draiswamy for a list of high ozone days over the northeast United States, Jasper Vrugt for providing his BMA code, Jun Du for his help with the SREF system, and Larry Bradshaw for his help with the WFAS data.

REFERENCES

Bakhshaii, A., and R. Stull, 2009: Deterministic ensemble forecasts using gene-expression programming. *Wea. Forecasting,* **24,** 1431–1451.

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

Candille, G., 2009: The multiensemble approach: The NAEFS example. *Mon. Wea. Rev.,* **137,** 1655–1665.

Carrera, M. L., J. R. Gyakum, and C. A. Lin, 2009: Observational study of wind channeling within the St. Lawrence River Valley. *J. Appl. Meteor. 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.

Colle, B. A., J. B. Olson, and J. S. Tongue, 2003: Multiseason verification of the MM5. Part I: Comparison with the Eta Model over the central and eastern United States 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, 1972: The National Fire-Danger Rating System. Rocky Mountain Forest and Range Experiment Station Research Paper RM-84, USDA Forest Service, 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.

——, and Coauthors, 2008: A Kalman-filter bias correction of ozone deterministic, ensemble-averaged, and probabilistic forecasts. *Tellus,* **60B,** 238–249.

——, T. Nipen, Y. Liu, G. Roux, and R. Stull, 2011: Kalman filter and analog schemes to postprocess numerical weather predictions. *Mon. Wea. Rev.,* **139,** 3554–3570.

Du, J., J. McQueen, G. DiMego, Z. Toth, D. Jovic, B. Zhou, and H.-Y. Chuang, 2006: New dimension of NCEP Short-Range Ensemble Forecasting (SREF) system: Inclusion of WRF members. Preprints, *Expert Team Meeting on Ensemble Prediction System,* Exeter, United Kingdom, WMO, 5 pp.

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

——, M. S. Allen, and M. C. Sittel, 2012: Estimation of ambiguity in ensemble forecasts. *Wea. Forecasting,* **27,** 50–69.

EPA, 1999: Air quality index reporting. 40 CFR Part 58/RIN 2060-AH92, 21 pp. [Available online at http://www.epa.gov/ttn/oarpg/t1/fr_notices/airqual.pdf.]

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

Fraley, C., A. E. Raftery, and T. Gneiting, 2010: Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Mon. Wea. Rev.,* **138,** 190–202.

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.

——, 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.

Gaza, R. S., 1998: Mesoscale meteorology and high ozone in the northeast United States. *J. Appl. Meteor.,* **37,** 961–977.

Glahn, B., M. Peroutka, J. Wiedenfeld, J. Wagner, G. Zylstra, B. Schuknecht, and B. Jackson, 2009: MOS uncertainty estimates in an ensemble framework. *Mon. Wea. Rev.,* **137,** 246–268.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.,* **11,** 1203–1211.

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.

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 Res.,* **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.

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.

——, and S. J. Colucci, 1997: Verification of Eta–RSM short-range ensemble forecasts. *Mon. Wea. Rev.,* **125,** 1312–1327.

——, and ——, 1998: Evaluation of Eta–RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.,* **126,** 711–724.

——, and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.,* **134,** 3209–3229.

——, ——, and S. L. Mullen, 2006: Reforecasts: An important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.,* **87,** 33–46.

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.

——, 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.

Hu, X. M., J. W. Nielsen-Gammon, and F. Zhang, 2010: Evaluation of three planetary boundary layer schemes in the WRF model. *J. Appl. Meteor. Climatol.,* **49,** 1831–1844.

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

Kang, D., R. Mathur, S. T. Rao, and S. Yu, 2008: Bias adjustment techniques for improving ozone air quality forecasts. *J. Geophys. Res.,* **113,** D23308, doi:10.1029/2008JD010151.

——, ——, and ——, 2010: Real-time bias-adjusted O_3 and PM2.5 air quality index forecasts and their performance evaluations over the continental United States. *Atmos. Environ.,* **44,** 2203–2212.

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.

Kinney, P. L., and H. Ozkaynak, 1991: Associations of daily mortality and air pollution in Los Angeles County. *Environ. Res.,* **54,** 99–120.

Kleiber, W., A. E. Raftery, J. Baars, T. Gneiting, C. F. Mass, and 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.

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

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.

Mao, Q., R. T. McNider, S. F. Mueller, and H. M. Juang, 1999: An optimal model output calibration algorithm suitable for objective temperature forecasting. *Wea. Forecasting,* **14,** 190–202.

Marrocu, M., and P. A. Chessa, 2008: A multimodel/multianalysis limited-area ensemble: Calibration issues. *Meteor. Appl.,* **15,** 171–179.

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.

McKeen, S., and Coauthors, 2005: Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004. *J. Geophys. Res.,* **110,** D21307, doi:10.1029/2005JD005858.

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. *Aust. Meteor. Mag.,* **54,** 35–55.

——, 2005b: On the subsynoptic-scale meteorology of two extreme fire weather days during the eastern Australian fires of January 2003. *Aust. Meteor. Mag.,* **54,** 265–290.

——, 2008a: Abrupt surface drying and fire weather. Part 1: Overview and case study of the South Australian fires of 11 January 2005. *Aust. Meteor. Mag.,* **57,** 299–309.

——, 2008b: Abrupt surface drying and fire weather. Part 2: A preliminary synoptic climatology in the forested areas of southern Australia. *Aust. Meteor. Mag.,* **57,** 311–328.

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.

——, 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., and 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 Interagency Fire Center, cited 2010: Fire information—Wildland fire statistics. [Available online at http://www.predictiveservices.nifc.gov/intelligence/2010_statssumm/fires_acres.pdf.]

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

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.

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

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.

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., and Coauthors, 1964: Synoptic weather types associated with critical fire weather. Pacific Southwest Range and Experiment Station, U.S. Forest Service, 492 pp.

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

Seaman, N. L., 2000: Meteorological modeling for air-quality assessments. *Atmos. Environ.,* **34,** 2231–2259.

——, 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.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp.

Sloughter, J. M., A. E. Raftery, and T. Gneiting, 2007: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.,* **135,** 3209–3220.

——, T. Gneiting, and A. E. Raftery, 2010: Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *J. Amer. Stat. Assoc.,* **105,** 25–35.

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.

——, and ——, 2005: Bias-corrected short-range ensemble forecasts of near surface variables. *Meteor. Appl.,* **12,** 217–230.

——, and ——, 2007: Reliable probabilistic quantitative precipitation forecasts from a short-range ensemble forecasting system. *Wea. Forecasting,* **22,** 3–17.

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

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

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.

Wilczak, J., and Coauthors, 2006: Bias corrected ensemble and probabilistic forecasts of surface ozone over eastern North America during the summer of 2004. *J. Geophys. Res.,* **111,** D23S28, doi:10.1029/2006JD007598.

Wilks, D. S., 2009: Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteor. Appl.,* **16,** 361–368.

——, 2011: *Statistical Methods in the Atmospheric Sciences.* 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.

Wilson, L. J., and M. Vallée, 2002: The Canadian Updateable Model Output Statistics (UMOS) system: Design and development tests. *Wea. Forecasting,* **17,** 206–222.

——, and ——, 2003: The Canadian Updateable Model Output Statistics (UMOS) system: Validation against perfect prog. *Wea. Forecasting,* **18,** 288–302.

——, 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.

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. Forecasting,* **22,** 1243–1256.

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