Evaluation and Postprocessing of Ensemble Fire Weather Predictions over the Northeast United States

MICHAEL J. ERICKSON[a] AND BRIAN A. COLLE

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

JOSEPH J. CHARNEY

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

(Manuscript received 26 June 2017, in final form 6 February 2018)

ABSTRACT

The Short-Range Ensemble Forecast (SREF) system is verified and bias corrected for fire weather days (FWDs) defined as having an elevated probability of wildfire occurrence using a statistical Fire Weather Index (FWI) over a subdomain of the northeastern United States (NEUS) between 2007 and 2014. The SREF is compared to the Rapid Update Cycle and Rapid Refresh analyses for temperature, relative humidity, specific humidity, and the FWI. An additive bias correction is employed using the most recent previous 14 days [sequential bias correction (SBC)] and the most recent previous 14 FWDs [conditional bias correction (CBC)]. Synoptic weather regimes on FWDs are established using cluster analysis (CA) on North American Regional Reanalysis sea level pressure, 850-hPa temperature, 500-hPa temperature, and 500-hPa geopotential height. SREF severely underpredicts FWI (by two indices at FWI = 3) on FWDs, which is partially corrected using SBC and largely corrected with CBC. FWI underprediction is associated with a cool (ensemble mean error of −1.8 K) and wet near-surface model bias (ensemble mean error of 0.46 g kg^{-1}) that decreases to near zero above 800 hPa. Although CBC improves reliability and Brier skill scores on FWDs, ensemble FWI values exhibit underdispersion. CA reveals three synoptic weather regimes on FWDs, with the largest cool and wet biases associated with a departing surface low pressure system. These results suggest the potential benefit of an operational analog bias correction on FWDs. Furthermore, CA may help elucidate model error during certain synoptic weather regimes.

1. Introduction

Wildfires in the northeastern United States (NEUS) burn an average of 13 633 acres annually (Pollina et al. 2013), which represents 0.27% of the total acres burned by wildfires in the contiguous United States. However, NEUS wildfires are often high impact phenomena because of the region's high population density. Recent examples of high-impact wildfires include the 7000 acre Sunrise fire (August 1995, New York) that closed a highway and stopped railroad service, effectively cutting the Hamptons off from the rest of Long Island, New York, for multiple days (McFadden 1995); the 1300 acre Double Trouble State Park wildfire (June 2002, New Jersey) that forced the closure of the Garden State Parkway at a high-volume traffic time and damaged or destroyed 36 homes and outbuildings (Charney and Keyser 2013); and the 2000 acre Manorville event (April 2012, New York) that closed roads and railways and destroyed 9 homes and outbuildings (Yan et al. 2012). While the aforementioned events represent some of the largest wildfire cases in the NEUS, even a small wildfire (<100 acres) can be disruptive given the population density in the region.

Fire managers in the NEUS depend on fire weather meteorologists to provide forecasts of the meteorological parameters that are conducive to fire occurrence, and they are particularly interested in weather forecasts on

[a] Current affiliation: Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, Colorado, and National Oceanographic and Atmospheric Administration Weather Prediction Center, College Park, Maryland.

Corresponding author: Michael Erickson, mjaerickson@gmail.com

DOI: 10.1175/JAMC-D-17-0180.1

days that exhibit an elevated probability for fire occurrence. Fire weather meteorologists, in turn, depend on numerical weather prediction (NWP) models to provide information about the current and future state of the atmosphere when preparing fire weather forecasts. For example, the Storm Prediction Center routinely produces operational forecasts of several fire weather indices using output from the Short-Range Ensemble Forecast (SREF; Du et al. 2012) system.

However, fire weather meteorologists must use NWP output cautiously because models can exhibit significant near-surface biases in meteorological variables relevant to fire weather applications. For example, Simpson et al. (2014a,b) analyzed the Weather Research and Forecasting (WRF) Model for the 2009/10 New Zealand fire season and reported negative 2-m temperature mean error (ranging from −4.4 to 0.0 K), negative and positive 2-m relative humidity mean error (ranging from −17.0% to +13.8%), positive 10-m wind speed mean error (averaging 1.4 $m s^{-1}$), and positive precipitation mean error (averaging 0.35 mm day^{-1}). Erickson et al. (2012) compared the SREF and the Stony Brook University NWP ensemble to climatology in the NEUS and reported a 2-m temperature negative mean error (ensemble mean of −2.5 K) and 2-m positive specific humidity mean error (ensemble mean of 2.1 g kg^{-1}) on fire threat days. The modeled cool and moist biases are greater on fire threat days than the climatological average, contributing to an underestimation of model-derived fire threat. Furthermore, Erickson et al. (2012) found a large removal of mean error (averaging −2.45 K for 2-m temperature) is possible on fire threat days in the SREF and Stony Brook University ensembles with postprocessing, suggesting the possibility that model performance may vary with the synoptic weather regime.

To evaluate model mean error on days with elevated fire threat, a robust method is needed to effectively capture and separate these anomalous days from the climatological average. Case studies (e.g., Kaplan et al. 2008; Charney and Keyser 2010 for the NEUS) assess model performance for a single event but typically do not produce statistically significant results and are not conducive to a more general assessment. Other studies, such as Hoadley et al. (2004), Hoadley et al. (2006), Mölders (2008), and Simpson et al. (2014a,b), have analyzed and verified model performance for an entire fire season. However, a fire season can vary greatly from location to location and from dataset to dataset, and not all days within a fire season are necessarily conducive to the occurrence of a wildfire that requires a management response. In the NEUS, this phenomenon can be particularly problematic because even during the most active months of a fire season, only a small percentage of days can be classified as fire occurrence days (Pollina et al. 2013). Furthermore, days that experience meteorological conditions conducive to the occurrence of a large fire, but do not experience ignitions, are not represented in a fire weather occurrence database.

To isolate the meteorological conditions that are conducive to fire ignition, Erickson et al. (2012) defined a ''fire threat day'' by combining the fire potential index from the Wildland Fire Assessment System (Burgan et al. 1998) with the National Fire Danger Rating System (Deeming et al. 1972). However, the Erickson et al. (2012) fire threat day definition is subjective and difficult to apply to other applications since it uses arbitrary thresholds from two uniquely different and complex rating systems. To address this deficiency, Erickson et al. (2016) employed a binomial logistic regression model to establish that 2-m relative humidity and 2-m temperature are the most effective statistical predictors of wildfire occurrence in the NEUS. A statistical fire weather index (FWI) is developed from the independent, skillful, and reliable predictions of wildfire occurrence and compared to the seasonal climatological probability of wildfire occurrence. Finally, the FWI threshold is used to define a fire weather day (FWD) in the NEUS as having a 30% or greater probability of wildfire occurrence from the binomial logistic regression model. The objective and straightforward characteristics of the FWI make it ideal for exploring NWP model performance on FWDs. It is important to note that the FWI used in this study significantly differs from the more complex FWI component within the Canadian Forest Fire Weather Index System (Van Wagner 1987), which considers the effects of fuel moisture, fire behavior, and meteorological conditions.

In this study, the FWI from Erickson et al. (2016) is applied to define FWDs in the NEUS and investigate the performance of an ensemble of NWP models. The goals of this paper include 1) quantifying the difference in ensemble model biases (i.e., vertically, horizontally, and by model) on FWDs and non-FWDs verified against a gridded analysis, 2) establishing the effectiveness of a training period on bias correction of the model fields, and 3) exploring how ensemble model biases vary with synoptic flow regimes on FWDs.

The paper is organized as follows: section 2 details how the FWI from Erickson et al. (2016) is applied to gridded data, including a description of the gridded analysis and ensemble model data. Results are presented in section 3, with comparisons of the gridded FWI to the original FWI and verification and postprocessing results for the FWI, temperature, and specific humidity. Section 4 explores how model bias varies with regional

atmospheric weather regimes using the FWI and section 5 concludes with a discussion of the results and future directions for ensemble NWP research related to FWDs.

2. Data and methods

a. Gridded analysis and ensemble model datasets

The ensemble model data for this study originate from the SREF system between 1 April 2007 and 30 June 2014. The SREF is run to forecast hour 87 four times daily (0300, 0900, 1500, and 2100 UTC) by the National Centers for Environmental Prediction (NCEP). There have been three major upgrades to the SREF in the period studied here, as detailed below:

1) SREF2007: Run between 1 April 2007 and 26 October 2009 with four unique cores [3 Advanced Research version of WRF (WRF-ARW), 3 WRF Nonhydrostatic Mesoscale Model (WRF-NMM), 10 Eta Model, and 5 Regional Spectral Model (RSM)] at 32–45-km grid spacing.
2) SREF2009: Run between 26 October 2009 and 21 August 2012 with four unique cores (5 WRF-ARW, 5 WRF-NMM, 6 Eta Model, and 5 RSM) at 32–35-km grid spacing.
3) SREF2012: Analyzed between 21 August 2012 and 30 June 2014 with three unique cores [7 WRF-ARW, 7 WRF-NMM, and 7 WRF B-grid of the NMM model (NMMB)] at 16-km grid spacing.

For additional details on the SREF physics and setup, see Erickson et al. (2012; their Table 1) and Du et al. (2012; slide 15). For simplicity, data from the SREF2007 and SREF2009 are combined into one period since the number of model cores remains consistent even though the number of members within each core changes. Although there are some physics changes between SREF2007 and SREF2009 within some of the model cores (particularly within the Eta model), these changes are assumed to have a minor effect on the overall performance metrics of the members. The specific model variables and levels analyzed are further explained in section 2d.

The datasets employed in this study to evaluate the SREF consist of hourly analyses from the Rapid Update Cycle (RUC; Benjamin et al. 2004), the Rapid Refresh (RAP; Benjamin et al. 2016), and the North American Regional Reanalysis (NARR; Mesinger et al. 2006). The RUC dataset is employed at 13-km grid spacing between 1 April 2007 and 1 May 2012, and replaced by the RAP at 13-km grid spacing from 21 August 2012 to 30 June 2014. While physics changes and upgrades have occurred within the RUC and RAP periods (Weygandt et al. 2013), this study assumes the impact of these changes to the analysis field (i.e., model hour zero) is minor. The specific atmospheric variables and levels examined in this study are described in sections 2c–e.

b. The fire weather index

This study uses the FWI developed in Erickson et al. (2016), which is based on a binomial logistic regression model with the predictor being the probability of fire occurrence and the predictands consisting of 2-m temperature and 2-m relative humidity. The functional form for the binomial logistic regression model is

$$\ln\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1\mathrm{TEMP}_i + b_2\mathrm{RELH}, \qquad (1)$$

where p is the probability of a wildfire occurring in the domain, i is each data sample, TEMP is the daily maximum of hourly 2-m temperature, RELH is the daily minimum of hourly 2-m relative humidity, and the three b are the regression coefficients. The binomial logistic regression model produces independent reliable probabilistic values of wildfire occurrence within a subdomain of the NEUS using Automated Surface Observing System (ASOS) station data.

Erickson et al. (2016) establish categories for the FWI based on thresholds of wildfire occurrence probabilities: less than 30% is assigned a value of zero, between 30% and 40% is assigned a value of one, 40% and 50% is given a value of two, and greater than 50% being given a value of three. For this study, all days with an FWI greater than zero are considered to be an FWD, which provides a large sample size of days while still analyzing events with a statistically elevated probability of fire occurrence. Since the FWI in Erickson et al. (2016) is developed using standardized anomalies of point observations, consideration must be given to how the gridded data should be standardized.

c. Developing a climatology for the gridded FWI

To perform a regression analysis, the gridded data must be standardized using a suitably long archive of the gridded mean and standard deviation (std dev) for temperature and relative humidity at each grid point. In other words, the RUC/RAP anomaly must be referenced to its own internal climate (Hamill et al. 2013), which is likely to exhibit different biases than in situ observations and other analyses. What constitutes a ''suitably long'' climatology is somewhat subjective, and it is possible that the 7-yr archive of the RUC/RAP is too short compared to a more typical 30-yr climatology used by the National Weather Service (NWS). Therefore, the shorter RUC/RAP analysis and the longer NARR analysis climatologies are compared.

Although NARR mean errors for the near-surface have been evaluated (Mesinger et al. 2006), they have not been studied on a subset of FWDs. Differences are calculated between the RUC/RAP and NARR for 2-m relative humidity and 2-m temperature during the overlapping periods of 2007–14. The mean relative humidity within the NARR is considerably higher than in the RUC/RAP (averaging 10%–15%), particularly with relative humidity values below 40% (not shown). This conditional positive relative humidity mean error in the NARR climatology would result in spuriously low FWI values. Based on this analysis, the NARR is an inappropriate climatological dataset for the near-surface variables, particularly for FWDs with low relative humidity values. Not all variables within the NARR reanalysis may be inappropriate to study fire weather applications, but caution should be used when analyzing near-surface variables, particularly moisture. While the RUC/RAP may not be the most conventional choice, it is deemed the most appropriate for this study given the NARR is biased with respect to the RUC/RAP for the variables analyzed. Ideally, this analysis should be revisited as additional years of RUC/RAP analyses become available to verify that the results presented here are not biased by the short duration of the climatology.

Before evaluating the FWI based on the gridded RUC/RAP climatology, the consistency of the gridded FWI is compared to the FWI using point values from ASOS observations, as in Erickson et al. (2016). The ASOS stations from 40.5° to 42°N and −74.5° to −71.5°W (Fig. 1) are used, which is consistent with domain 1 from Erickson et al. (2016; their Fig. 1). The binomial logistic regression model parameters are assumed to be spatially invariant and hence are applied to all points in the model grid and at each observation location. This assumption is based on the minor variations in parameter estimates between the mid-Atlantic and New York City domain found in Erickson et al. (2016), and the minor spatial variations in parameter estimations found within each domain (not shown). Although this assumption is not likely to hold over larger areas, this study assumes it is acceptable within the NEUS subdomain.

The same binomial logistic regression model used in Erickson et al. (2016) is applied here point-by-point over the domain in Fig. 1, resulting in one 2-m temperature and one 2-m relative humidity input for each point in the model domain. A spatial median of FWI is computed for all Fig. 1 points to determine the final probability of fire occurrence. 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 Snow and Ice Data Center 2008) to find and exclude these days.

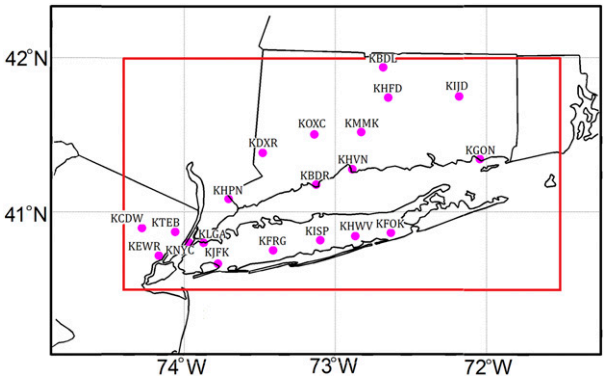

FIG. 1. Domain used (red box) and ASOS stations used for developing the grid-based FWI.

d. *Ensemble verification and postprocessing*

Verification of the SREF on FWDs is separated into two unique periods based on the data availability of the RUC and RAP. SREF1 consists of the SREF2007 and SREF2009 between 1 April 2007 and 1 May 2012 verified with the RUC analysis. SREF2 consists of the SREF2012 between 21 August 2012 and 30 June 2014 verified using the RAP analysis. The time period between 2 May 2012 and 20 August 2012 is not considered to keep the newer (older) SREF version verification consistent with the RAP (RUC) analysis.

To calculate the FWI, the daily maximum 2-m temperature and daily minimum 2-m relative humidity are obtained for each SREF member's day 1 to day 3 forecasts initialized at 0900 UTC, and then averaged. For these calculations, a day is defined as the time span from 0000 to 2300 UTC, but all maximum FWI values used for this analysis occur between 1200 and 2300 UTC. All RUC/RAP analyses are bilinearly interpolated to the SREF grid before verification or calculation of the FWI.

The ensemble analyses are verified against the RUC/RAP analysis by analyzing systematic bias and nonsystematic error metrics for all FWDs. In this case, an FWD is defined as having an FWI of greater than zero in either the RUC analysis or SREF forecast. In addition, meteorological variables important to fire weather are assessed for FWDs to quantify the three-dimensional structure of the bias. Ensemble model biases and error for each SREF core are assessed by calculating mean error (ME) or mean absolute error (MAE) by threshold (Wilks 2011). Reliability plots (Wilks 2011) comparing the average observed probability for select forecast probability bins are produced for several thresholds. Brier skill scores (BSS; Wilks 2011) are calculated to

explore the skill of the ensemble probabilities relative to some other reference, which will be elaborated in section 3e.

A bias correction is applied and evaluated for the SREF ensemble using a spatially variant additive approach (Wilson et al. 2007; Erickson et al. 2012). The additive approach to bias correction generally performs best for normally distributed variables such as temperature and daily minimum relative humidity. As in Erickson et al. (2012), two types of postprocessing are explored: one with sequential bias correction (SBC; uses the previous 14 days to train the bias correction) and another with conditional bias correction (CBC; uses 14 previous FWDs to correct future FWDs). For the purposes of this study, SBC can be thought of as a typical bias-correction approach, while CBC is similar to a simple analog bias correction. When the FWI is being verified with CBC or SBC, relative humidity and temperature are postprocessed before the index is calculated.

Training and verification periods are created by bootstrapping (Wilks 2011) the original dataset with replacement 1000 times to reconstruct a new dataset with the same number of FWD events as the original dataset. The purpose of the bootstrapping methodology is to assess statistical significance by gathering multiple data samples. In the case of the CBC data, all FWDs are simply resampled, which rearranges the chronological order of the original data. In the case of SBC, the FWDs are resampled but the chronological order of the previous 14 days (which may or may not be an FWD) are preserved. The training and verification windows are always independent with postprocessing being applied using a sliding window approach (i.e., iteratively advancing the training window one day forward after verification). All results with the raw, SBC, and CBC contain data in the independent verification window only. Error bars in all plots represent the 25th and 975th permillile of the resampled dataset. All references to ''statistically significant'' or ''significant'' indicate confidence exceeding the 95% threshold using the bootstrapped datasets.

e. *Using cluster analysis to explore regime-based model mean error on FWDs*

Cluster analysis (CA) is a common method for separating events associated with different synoptic weather regimes (Huth et al. 2008). Typically, principal component analysis (PCA) is applied before CA to reduce the dimensionality of the data (Huth et al. 2008). This study applies T-mode (Wilks 2011) PCA separately to the 1800 UTC NARR standardized sea level pressure (SLP), 850-hPa temperature, 500-hPa temperature, and 500-hPa geopotential height on all FWDs. The 1800 UTC time period is selected since this is when model near-surface temperature and specific humidity mean error is maximized (not shown). Although it was previously found that the NARR exhibits greater near-surface temperature and relative humidity mean error compared to the RUC (section 2b), mean error reduces to near zero above the PBL (not shown) and is not expected to significantly impact SLP or variables above 850 hPa.

The NARR is selected for CA since it has a larger domain size than the RUC and RAP. Although a smaller domain size is presented in this study, the sensitivity of CA to larger domain sizes has been explored (not shown). The domain presented here spans from 32° to 48°N latitude and 82° to 68°W longitude (see domain in Fig. 14). This region captures the ambient atmosphere surrounding the NEUS while excluding the more variable synoptic details upstream and downstream of the region. Loadings that explain 90% of the variance are retained and included as separate variables for the CA, resulting in five SLP loadings, three 850-hPa temperature loadings, three 500-hPa temperature loadings, and three 500-hPa geopotential height loadings. A Varimax (Wilks 2011) orthogonal rotation is applied to facilitate interpretation. Thereafter, a *k*-means clustering (Lloyd 1982) on the retained PCs is used to minimize the distance between each object and the cluster centroid over all clusters. Since the optimal total number of clusters is not known a priori with *k*-means clustering, the optimal number of clusters is selected by maximizing the average silhouette value (Rousseeuw and Leroy 1987), which is a common measure of intracluster similarity and intercluster dissimilarity among all the points. Model performance in terms of ME is explored and compared for each cluster. The relationships between the datasets (model and analysis data), statistical procedures (bias correction, binomial logistic regression, and cluster analysis), and SREF verification presented in sections 2d and 2e are shown in Fig. 2.

3. Results

a. *Creating a gridded statistical FWI*

When developing an FWI climatology using the RUC/RAP analyses, the RUC/RAP climatological distributions of temperature and relative humidity must be compared to ASOS observations (as in Erickson et al. 2016), to assess whether the effects of model cores, model physics, and data assimilation affect the FWI at initialization (Weygandt et al. 2013). To investigate these effects, four RUC/RAP FWI climatologies are produced using different methodologies, and each is compared to the ASOS-based FWI climatology. The first climatology uses only the RUC to normalize the RUC/RAP data

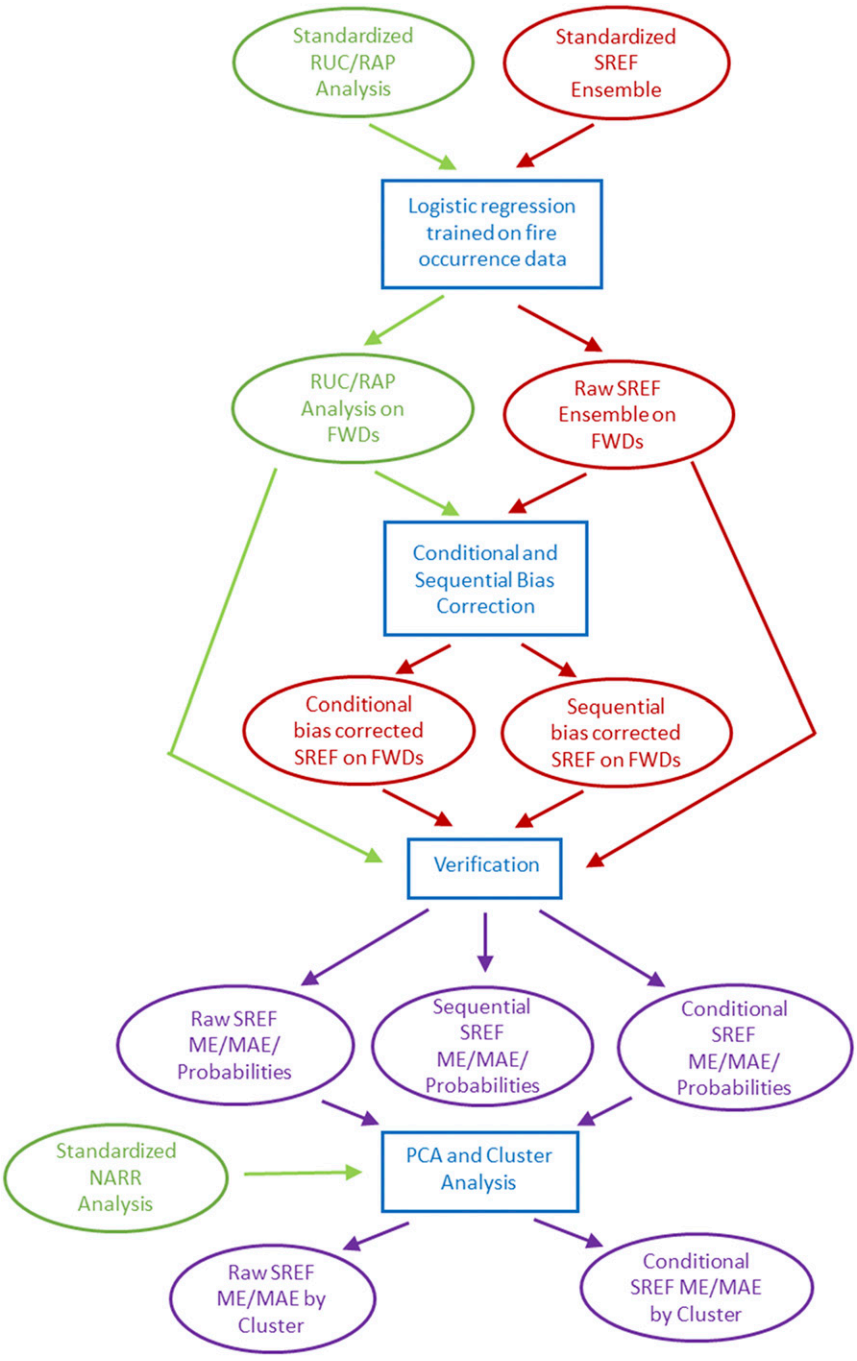

FIG. 2. Flowchart demonstrating the datasets, statistical techniques, and relationships between them in analyzing the SREF members. Ovals represent datasets while rectangles represent statistical techniques. Green indicates analysis data, red is model data, and purple is verification output.

(RUC_ALL), the second uses the RAP for the RUC/RAP data (RAP_ALL), the third uses the combined RUC/RAP climatology for the entire period (RR_ALL), and the fourth applies the RUC and RAP normalization to the RUC and RAP periods separately (RR_SEP).

The four RUC/RAP FWI climatologies are compared to the ASOS FWI climatology for all thresholds using ME (Fig. 3a) and MAE (Fig. 3b) metrics. An underprediction of the gridded FWI is apparent for most thresholds using RUC_ALL (ME = −0.01 at FWI ≥ 1),

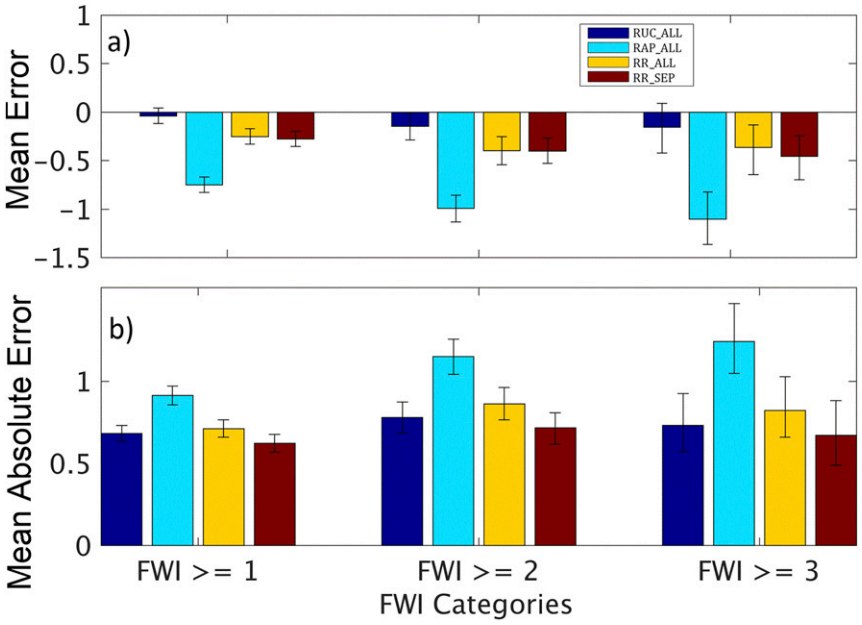

FIG. 3. (a) ME and (b) MAE for RUC/RAP-derived FWI compared to ASOS stations–derived FWI using RUC climatology (RUC_ALL; blue), RAP climatology (RAP_ALL; cyan), RUC and RAP combined climatology (RR_ALL; yellow), and RUC climatology for the RUC analysis and RAP climatology for the RAP analysis (RR_SEP; red). Error bars represent the 25th and 975th permillile of the 1000 resampled datasets.

RAP_ALL (ME = −0.74 at FWI ≥ 1), RR_ALL (ME = −0.23 at FWI ≥ 1), and RR_SEP (ME = −0.25 at FWI ≥ 1). In addition, RAP_ALL has statistically significantly lower ME (difference averaging 0.57 at FWI ≥ 1) and greater MAE (difference averaging 0.25 at FWI ≥ 1) than the other climatologies. RUC_ALL has a better ME than all other methodologies for FWI ≥ 1. In terms of MAE, RUC_ALL, RR_ALL, and RR_SEP are not significantly different from each other in the resampled datasets at any FWI threshold. Although the gridded FWI exhibits an underprediction compared to ASOS FWI, all the climatologies perform equally well, with the exception of RAP_ALL. The gridded FWI exhibits skill in identifying FWDs, which suggests the FWI can be used to investigate the three-dimensional structure of model error using the gridded analyses. Given that all climatologies except RAP_ALL perform well, separating the RUC and RAP climatologies makes the most intuitive sense. Hence, RR_SEP is selected as the default climatology for normalizing the gridded FWI in this study.

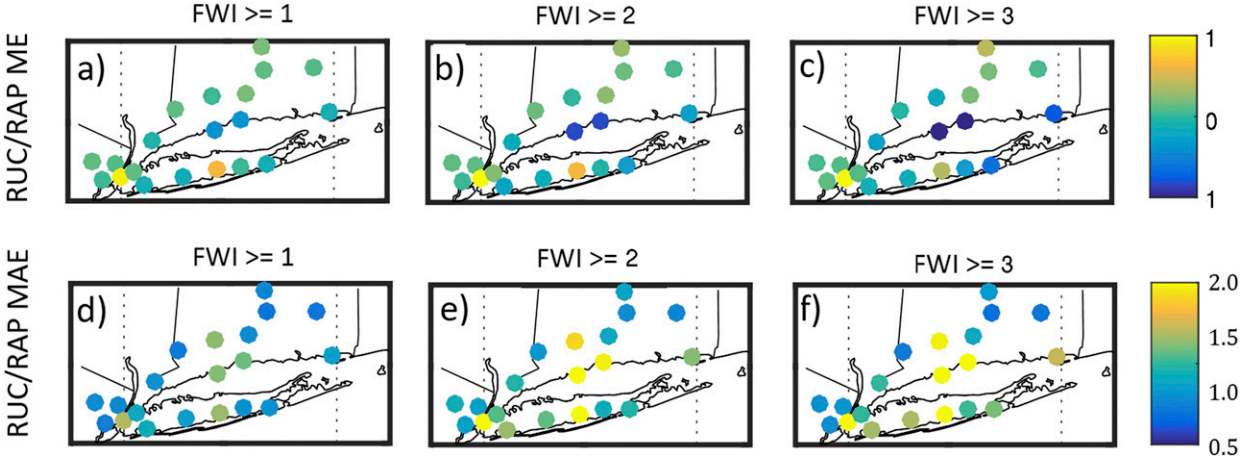

FIG. 4. Spatial (a)–(c) ME and (d)–(f) MAE for RUC/RAP-based FWI compared to ASOS-based FWI by threshold.

To evaluate spatial consistency with the surface observations, the gridded FWI is interpolated to each ASOS station and compared to the ASOS FWI using ME and MAE scores (Fig. 4). As discussed in section 2c, the final gridded FWI is derived from the spatial median of the output from the binomial logistic regression model. However, for the spatial comparisons discussed here, the final step of computing the spatial median is skipped. ME values at FWI ≥ 1 are close to zero with the exception of KNYC (Fig. 1; Central Park, New York; ME = 0.94) and KISP (Islip, New York; ME = 0.63). KNYC has considerably more data missing (75%) compared to the average station (averaging 19.4%), which might affect the results, while KISP may be influenced by representativeness errors given the bilinear interpolation of land and water points near the coast. In general, coastal boundary representativeness errors produces ME values for many coastal locations (KFOK, KHWV, KISP, KFRG, KJFK, KLGA, and KHPN in New York; KBDR, KHVN, and KGON in Connecticut) that are substantially higher than ME values at inland locations (KBDL, KHFD, KIJD, KMMK, and KDXR in Connecticut; KEWR, KTEB, and KCDW in New Jersey) at FWI ≥ 1 (difference averaging 0.18). However, the effect of the representativeness errors on the mean error of the gridded FWI (i.e., Fig. 3a) is relatively small when averaged over the domain, which is consistent with the ASOS FWI.

Figure 5 compares the average number of FWDs per year for the RUC/RAP analysis and ASOS observations stacked by FWI. The FWI climatologies are qualitatively similar, with a primary peak in April and a smaller secondary peak in July. However, the gridded FWI climatology identifies fewer FWDs than the ASOS FWI, particularly in March, which is consistent with the overall analysis presented in Fig. 3. Figure 5 emphasizes the rarity of an FWI = 3 event, which occurs on average 4.3 days yr^{-1} using the RUC/RAP analysis. Since the gridded and ASOS FWI climatologies are comparable, the gridded technique is considered suitable for identifying FWDs in the SREF and RUC/RAP analysis. It is apparent that FWDs are still being selected with the gridded approach, albeit less frequently than the ASOS FWI. The differences between gridded and ASOS FWI could be caused by minor nonnormal deviations in the distribution of the RUC/RAP climatology compared to ASOS observations that are not corrected through normalization.

b. *Ensemble mean verification and bias correction of SREF-derived FWI*

ME is computed for all FWI thresholds and averaged across all model cores for the SREF1 and SREF2 (Fig. 6).

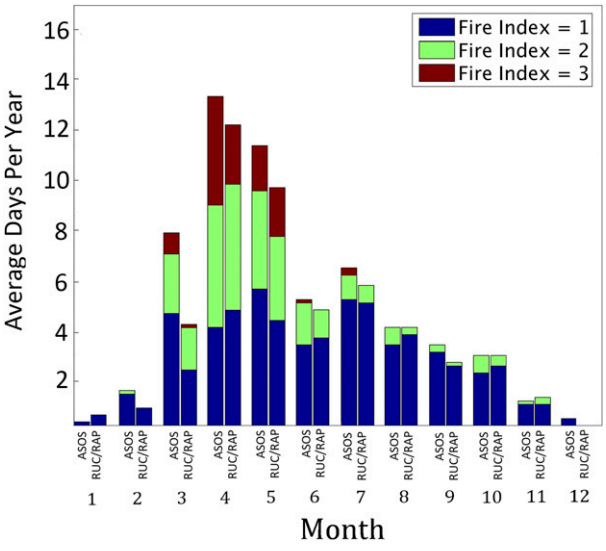

FIG. 5. FWI average days per year by index value (blue, green, and red for FWI values of 1, 2, and 3, respectively) grouped by ASOS stations (left bars) and RUC/RAP (right bars) between 2007 and 2014.

For an FWI ≥ 1, the values are underpredicted for the SREF1 (ME = −1.29) and SREF2 (ME = −1.34). This underprediction generally grows with FWI value (e.g., SREF1 has an average value of −2.00 and SREF2 an average value of −2.73 for FWI = 3). Ensemble average ME at FWI ≥ 1 is substantially improved with SREF1 SBC (ME = −0.70) and SREF2 SBC (ME = −0.56) for all SREF cores. There is additional improvement at FWI ≥ 1 for SREF1 CBC (ME = −0.02) and SREF2 CBC (ME = 0.02). However, in most cases the CBC overcorrects the mean error at FWI = 3, with the SREF1 CBC averaging 0.56 and SREF2 CBC averaging 0.78. The performance of the CBC-based FWI suggests that near-surface atmospheric biases differ significantly for FWDs compared to the annual average.

Figure 7 shows MAE for all FWI thresholds averaged across the ensemble cores. MAE is significantly reduced compared to raw data for the SREF1 SBC (improvement averaging 0.42) and SREF2 SBC (improvement averaging 0.51) at an FWI ≥ 1. However, CBC does not result in any statistically significant improvements over SBC for any thresholds analyzed. This suggests that although postprocessing is generally effective at removing mean error, there is still considerable day-to-day variability that is difficult to correct with CBC. The performance of individual SREF cores is discussed in section 3d.

Figure 7 can also be used to compare improvements in ensemble performance associated with model upgrades. For instance, the 2012 SREF upgrade (i.e., SREF2) invoked a newer WRF version (from version 2.2 to

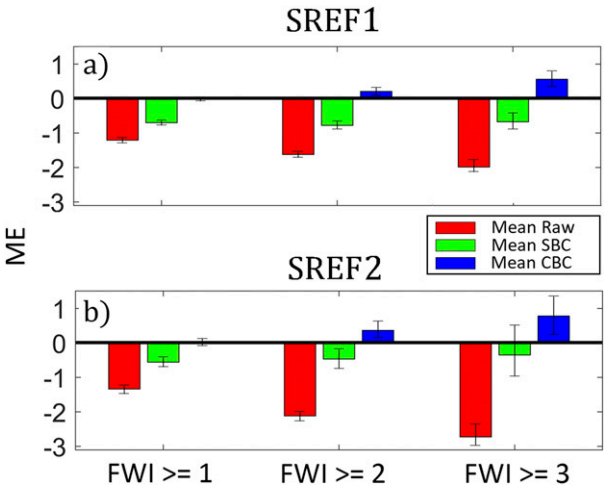

FIG. 6. SREF ensemble-mean-based raw (red), SBC (green), and CBC (blue) ME by FWI threshold for (a) the SREF between 1 Apr 2007 and 1 May 2012 (SREF1) and (b) the SREF between 21 Aug 2012 and 30 Jun 2014 (SREF2). Error bars represent the 25th and 975th permillile of the 1000 resampled datasets.

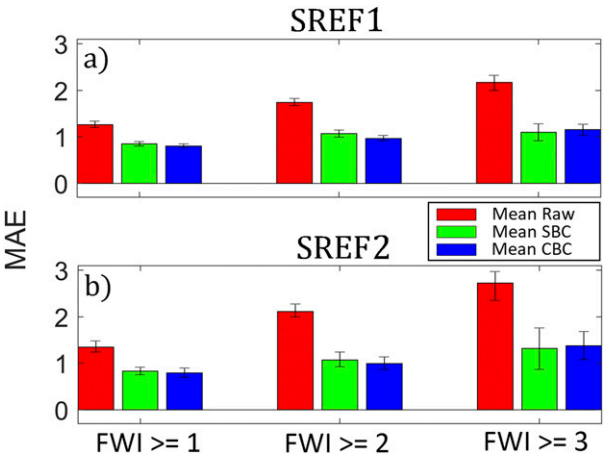

FIG. 7. As in Fig. 6, but for MAE.

version 3.3), increased the model resolution (from ~32 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). The raw MAE is statistically significantly worse in SREF2 compared to SREF1 for all thresholds except FWI = 1. SBC and CBC bias correction improves upon the MAE, with no significant differences between the SREF1 and SREF2. The potential sources of this bias are explored in sections 3c and 3d by analyzing the model variables that went into the FWI. Note that the slightly larger error bars for the SREF2 is related to the limited training period available (total of 86 FWDs), compared to SREF1 (total of 235 FWDs). The differences between SREF1 and SREF2 emphasize the importance of retraining the statistical methods employed in this paper after model upgrades.

Spatial ensemble ME and MAE are shown for the SREF1 and SREF2 at FWI $\geq$ 1 after CBC is applied (Fig. 8). The ensemble mean still exhibits significant spatial variability for both ME and MAE despite the use of a spatially additive bias correction. For instance, spatial MAE values range from 0.74 to 1.21 in SREF1 and from 0.68 to 1.22 in SREF2. ME values are more negative over the Long Island Sound and in the New York Bight region for SREF1 and SREF2, which degrades the MAE for these locations. This contrast along the coast could be associated with representativeness errors between the higher resolution RUC/RAP analysis and coarser SREF model grid. For instance, relative humidity and temperature differences between ocean and atmosphere during FWDs can be as large as 70% and 20 K, respectively (not shown).

c. *Ensemble mean verification and bias correction of SREF temperature and specific humidity*

From section 3b, raw SREF model output exhibits a large underprediction of FWI, particularly for the SREF2. Therefore, it is important to understand how model biases in temperature and specific humidity affect the FWI and explore model performance for additional meteorological variables of interest to the fire weather community. The vertical profile of ensemble mean temperature ME is shown for the raw, SBC, and CBC in Fig. 9. Temperature exhibits a significant cool mean error maximized at 1000 hPa for SREF1 (averaging −2.41 K) and SREF2 (averaging −1.30 K) that decays to a ME near zero above 800 hPa. ME is significantly improved compared to the raw temperature for SREF1 (SREF2) SBC below 825 hPa (925 hPa). Likewise, CBC significantly improves on ME over SBC in both ensembles below 850 hPa. In contrast to the FWI results from Fig. 6, the upgraded SREF2 statistically significantly improves raw ME below 700 hPa (by 1.21 K at 1000 hPa) compared to the SREF1. This benefit is also apparent after SBC, where the SREF2 exhibits significantly less mean error below 850 hPa compared to the SREF1.

The ensemble mean vertical structure of SREF MAE is analyzed for the SREF1 and SREF2 (Fig. 10). SREF1 SBC results in a significant improvement below 875 hPa compared to the raw data (at 1000 hPa averaging 1.07 K). However, SREF2 SBC only yields statistically significant improvement below 925 hPa, suggesting that bias correction is less impactful in this case for the newer version. Similar to Fig. 9, the raw SREF2 exhibits significantly lower MAE than the raw SREF1 below 850 hPa.

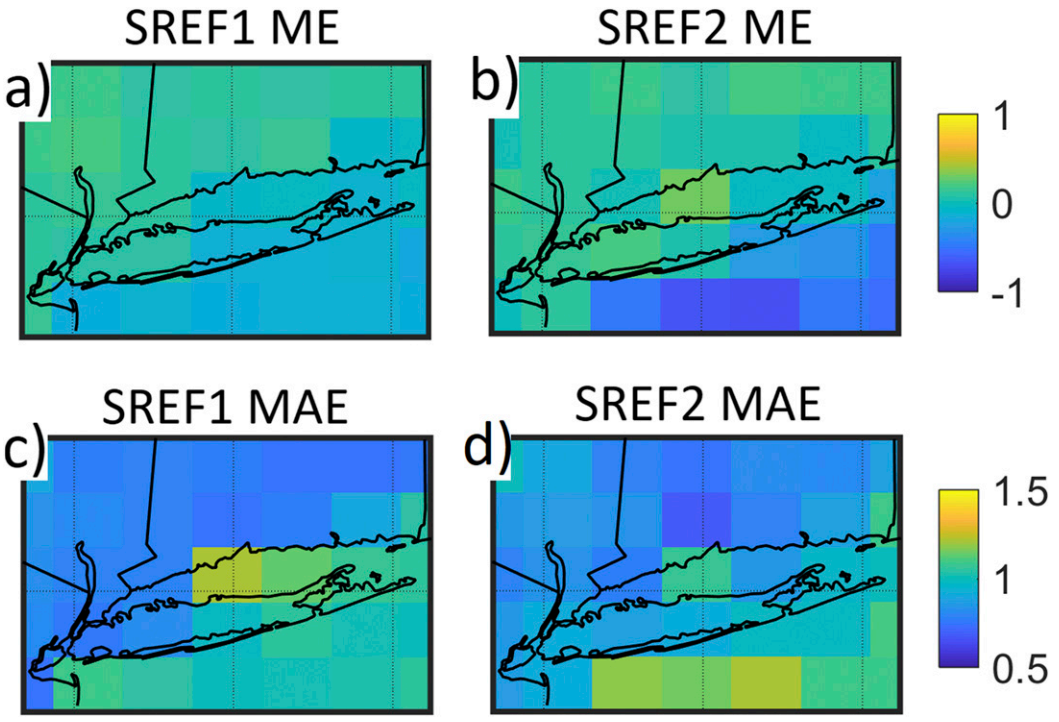

FIG. 8. Ensemble-mean-derived spatial (a) SREF1 ME, (b) SREF2 ME, (c) SREF1 MAE, and (d) SREF2 MAE after applying CBC for the FWI ≥ 1.

As with temperature, specific humidity has a statistically significant positive moisture mean error for the SREF1 (SREF2) raw data below 825 hPa (875 hPa) that is maximized at 975 hPa (1000 hPa) for the ensemble mean (Fig. 11). SREF1 (SREF2) SBC significantly improves ME below 850 hPa (900 hPa) with additional significant improvements for CBC below 875 hPa (975 hPa). The SREF upgrade in 2012 significantly improves the high biased raw specific humidity forecasts between 850 and 975 hPa.

Overall, the underprediction of FWI in the raw SREF model data is associated with large negative temperature (Fig. 9) and positive specific humidity (Fig. 11) biases that generally occur below 875 hPa. The result is a relative humidity mean error of greater than 10% below 875 hPa (950 hPa) in SREF1 (SREF2; not shown). Bias correction can effectively remove these PBL biases and marginally improve model error metrics like MAE.

d. Performance of individual SREF cores on FWDs

SREF performance among the model cores is analyzed for ME at FWI = 1, 2-m temperature, and 2-m specific humidity (Fig. 12). Two-meter temperature raw ME is significantly more negative for the SREF1 within the WRF-ARW and RSM cores. Likewise, 2-m specific humidity ME is significantly greater within the Eta and RSM cores. The effect of these biases is apparent on the raw FWI values, with the WRF-NMM performing the best (raw ME = 0.8) and the RSM performing the worst (raw ME = 1.4). Raw ME is more consistent across each core for all variables analyzed within SREF2. There are no significant differences in raw ME between any of the SREF2 averaged cores. Except for SREF1 WRF-ARW, SREF1 WRF-NMM, and SREF1 RSM specific humidity, SBC significantly improves ME over the raw values. In all instances, CBC improves upon SBC values of ME.

There is an inconsistency between the lower atmosphere (1000 hPa) improvement in raw SREF2 temperature (Fig. 9) and specific humidity (Fig. 11) compared to SREF1, and the more negatively biased FWI forecasts at the near-surface for SREF2 (Fig. 6b) versus SREF1 (Fig. 6a). Since near-surface variables are used to compute the FWI rather than 1000-hPa values, Fig. 12 provides perspective into the degraded raw FWI values in Fig. 6. Examining differences in the model upgrade, SREF2 raw 2-m temperature mean error has a statistically significant improvement in the WRF-ARW (0.7 K) with little change in the WRF-NMM performance. However, there is a statistically significant increase in SREF2 2-m specific humidity mean error for the WRF-ARW (by 0.56 g kg^{-1}) and WRF-NMM (by 0.44 g kg^{-1}) when compared to the SREF1. This large positive 2-m specific humidity mean error in SREF2 negatively

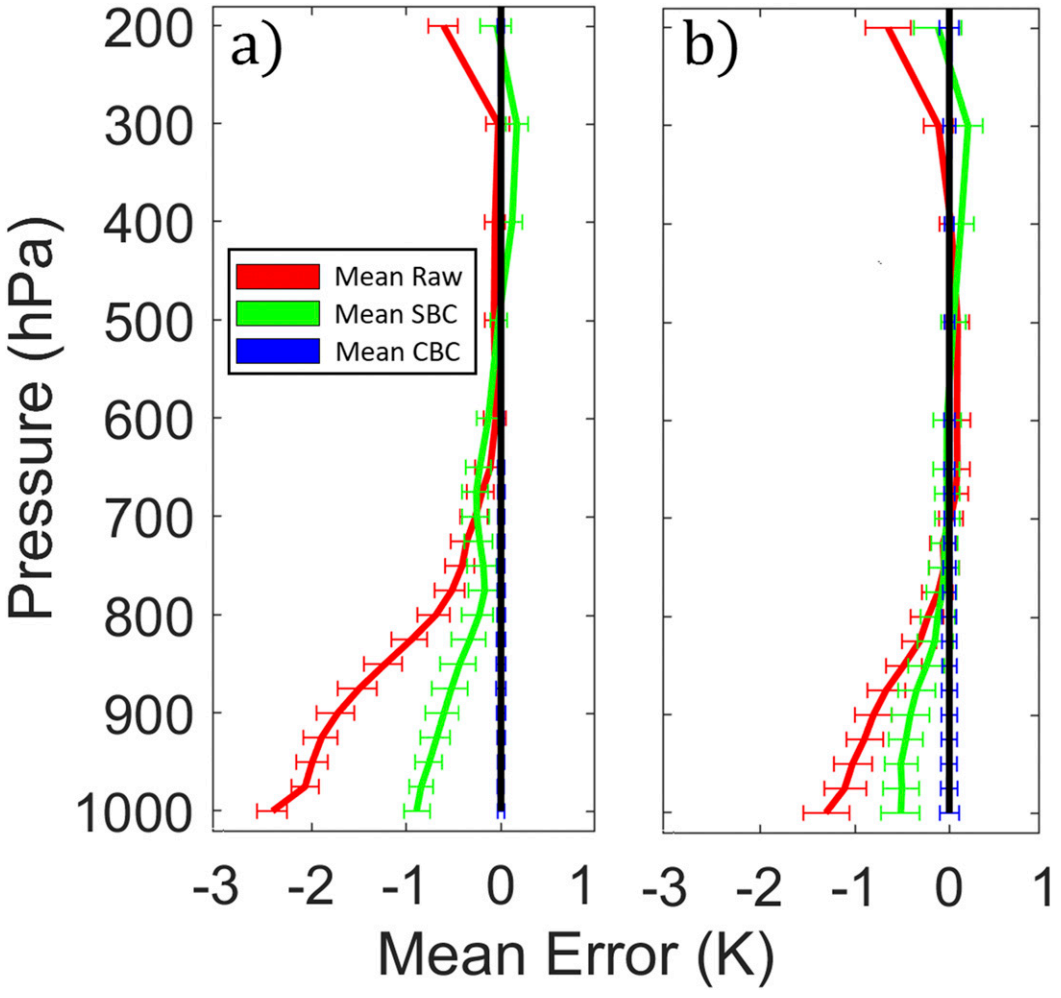

FIG. 9. ME-derived SREF ensemble mean temperature profiles of raw (red), SBC (green), and CBC (blue) for (a) SREF1 and (b) SREF2. Error bars represent the 25th and 975th permillile of the 1000 resampled datasets.

impacts the raw FWI and emphasizes the importance of careful postprocessing, even when ensembles experience critical upgrades. Since this increase in SREF2 specific humidity mean error is not found at 1000 hPa, there may be issues within the PBL scheme or land surface model when interpolating the lowest model sigma level to the near surface for the SREF2 WRF cores.

e. Probabilistic SREF verification of the FWI

Figure 13 shows the SREF1 and SREF2 ensemble reliability (section 2d) at FWI ≥ 2 for raw (red), SBC (green), and CBC (blue). Raw ensemble FWI is underpredicted, with the SREF2 never predicting raw FWI probabilities greater than 10%. Bias correction improves this mean error, particularly with CBC, which is consistent with section 3b. However, low (high) CBC-derived FWI probabilities verify too low (high) compared to observations, especially for SREF2 CBC.

BSSs (Wilks et al. 2011) are used to compare the probabilistic accuracy of the SBC and CBC FWI with raw FWI as the reference (Fig. 14). The BSS is statistically significantly greater than zero for FWI ≥ 1 and FWI ≥ 2, indicating that the probabilistic accuracy of both bias corrections improves upon the raw model despite the ensemble underdispersion (Fig. 13). In addition, both bias-correction methods exhibit average BSS > 0 at FWI ≥ 3, although this result is not statistically significant. Compared to SBC, CBC significantly improves upon BSS at FWI ≥ 1, but does not significantly change BSS at higher thresholds.

4. Sensitivity of model bias to cluster analysis on FWDs

As described in section 2e, the optimal number of CA clusters is three for temperature by using the silhouette

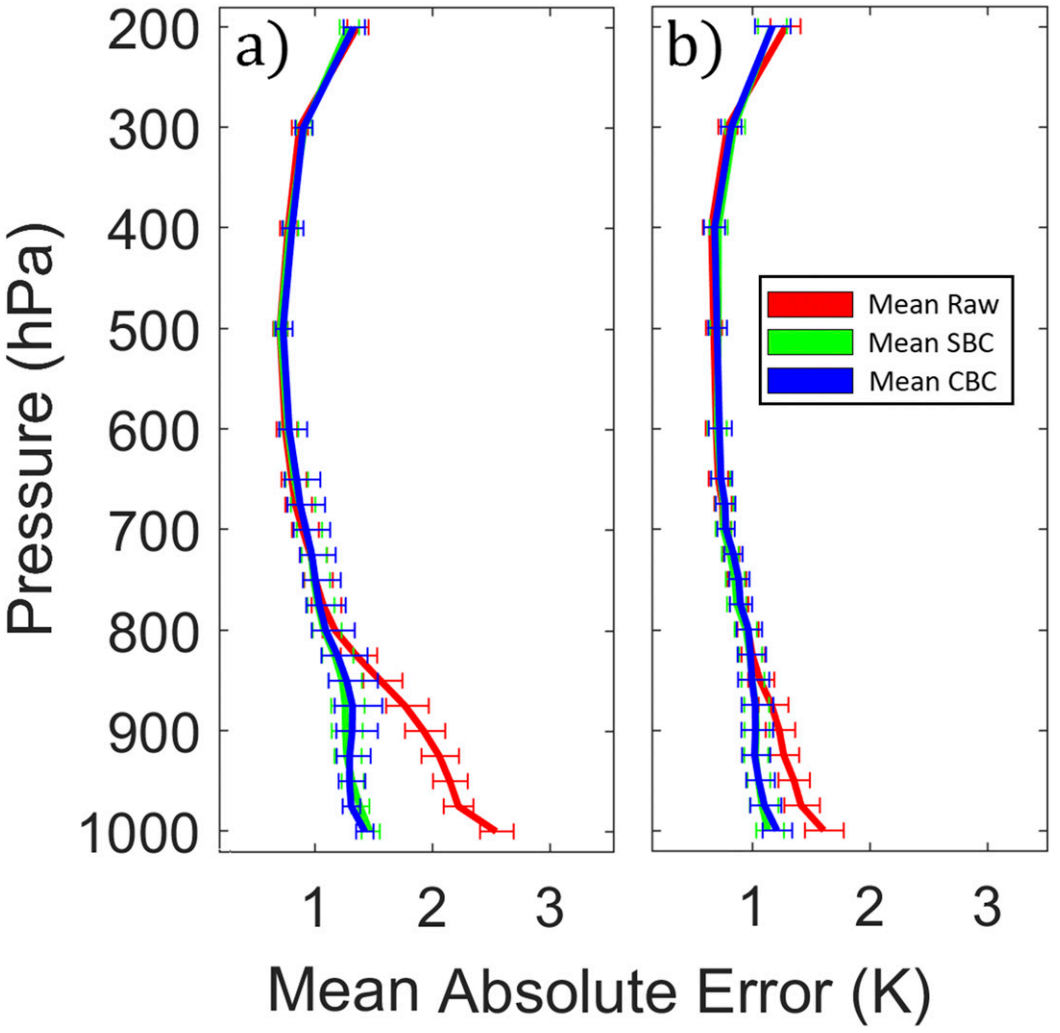

FIG. 10. As in Fig. 9, but for temperature MAE.

technique (Rousseeuw and Leroy 1987). To examine the regional characteristics of each cluster, the composite of each cluster's 500-hPa height anomaly and SLP anomaly is shown in Fig. 15. Since some events do not fit clearly into any assigned cluster, all days with a silhouette value below 0.1 are excluded from the composite. Cluster 1 appears to be associated with a departing surface low pressure system and advancing surface high pressure system from Canada. This weather regime is similar to the "pre-high" regime from Pollina et al. (2013). Cluster 2 is associated with an advancing 500-hPa height ridge from the west and a SLP maximum centered over the region or to the south. This weather regime is a combination of the "extended high" and "high to the south" regimes from Pollina et al. (2013). The third cluster is associated with a 500-hPa ridge and SLP maximum to the southwest of the region, and is identical to the "back of high" regime described in Pollina et al. (2013). Cluster 1 is typically associated with cool and dry surface conditions with a low-level northwest wind, while clusters 2 and 3 typically have warmer surface conditions with light and southwest low-level winds, respectively.

The ensemble mean 2-m temperature ME for all FWDs are separated in boxplots by cluster for the SREF1 and SREF2 raw and CBC (Fig. 16). Clusters 1 and 2 have a greater negative temperature mean error than cluster 3 for both the raw SREF1 and SREF2. Using a Kolmogorov–Smirnov (K-S; Wilks 2011) test at 95% confidence, the differences between cluster 3 and clusters 1 and 2 are statistically significant for both raw SREF1 and SREF2. This is also true for relative humidity (not shown), where cluster 3 exhibits significantly smaller positive moisture mean error than clusters 1 and 2. These statistically significant differences in model

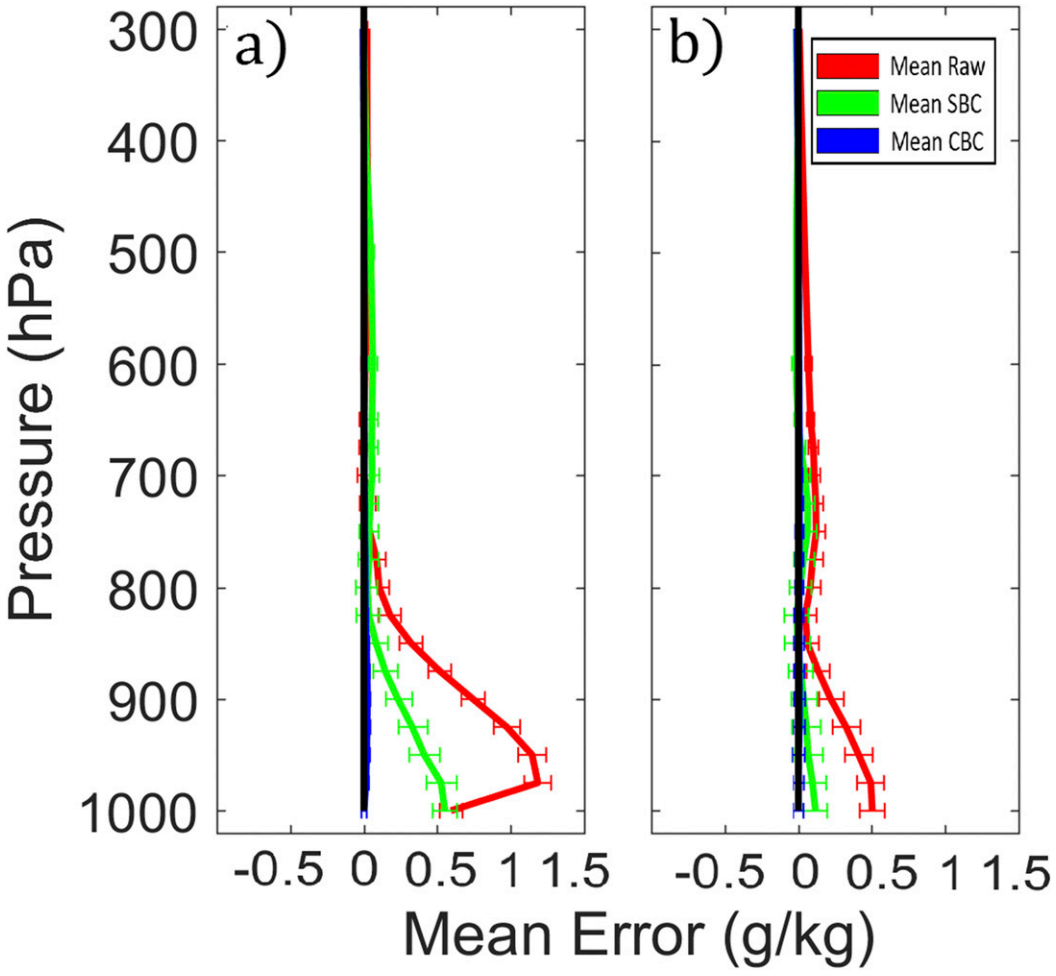

FIG. 11. As in Fig. 9, but for specific humidity ME.

mean error between clusters remain even after CBC is applied.

Model mean error by FWI strength and cluster is presented for the SREF1 and SREF2 raw and CBC (Fig. 17) temperature fields. Model mean error for temperature appears to be dependent on both cluster and FWI strength. For instance, FWDs with a higher FWI generally have a colder model mean error. CBC is not completely effective at removing model mean error conditional on cluster or FWI for either the SREF1 (Fig. 17b) or SREF2 (Fig. 17d), suggesting that there is potential to improve the additive bias correction with intelligent subsetting of FWDs combined with thresholding techniques.

5. Conclusions

This study explores model error characteristics on fire weather days (FWDs) by comparing the Short-Range Ensemble Forecast (SREF) system to the Rapid Update Cycle (RUC) and Rapid Refresh (RAP) analyses. The Fire Weather Index (FWI) from Erickson et al. (2016) is used to define a consistent and reliable quantification of the atmospheric conditions that constitute an FWD. The effectiveness of bias correction on FWDs is explored using the previous 14 days (sequential bias correction) and most recent 14 FWDs (conditional bias correction). Mean error, mean absolute error, reliability plots, and Brier skill scores are calculated on FWDs for the SREF. Finally, cluster analysis (CA) is applied to North American Regional Reanalysis (NARR) data on FWDs to explore the relationship between regional flow pattern and model error. An overview of the datasets, statistical techniques, and relationships between them is presented in Fig. 2.

SREF severely underpredicts FWI (mean error of 1.3 indices at FWI = 1) on FWDs, which is partially corrected with sequential bias correction (mean error of 0.5

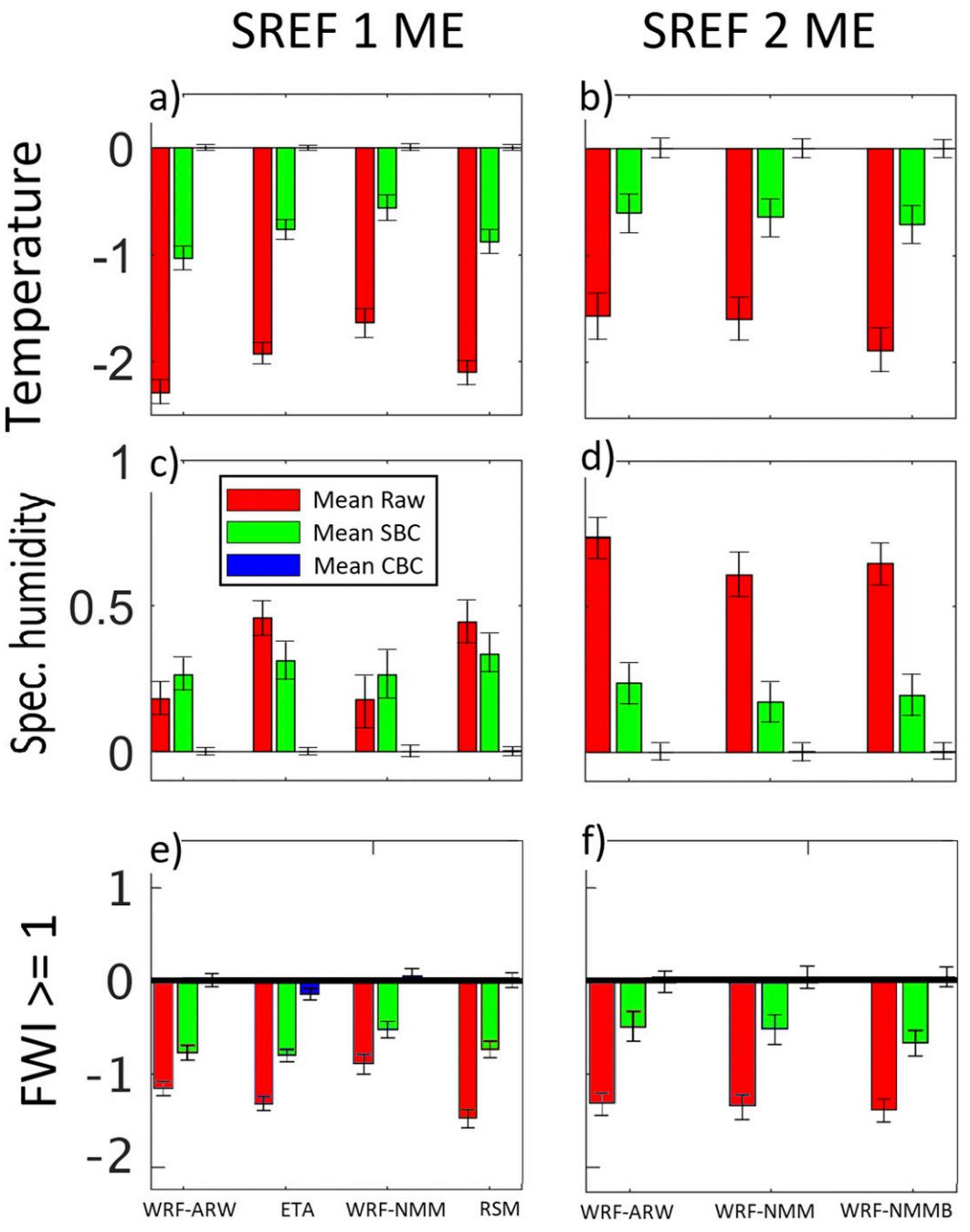

FIG. 12. Raw (red), SBC (green), and CBC (blue) ME averaged by model core for the (a),(c),(e) SREF1 and (b),(d),(f) SREF2 2-m temperature [in (a) and (b)], 2-m specific humidity [in (c) and (d)], and values of the FWI ≥ 1 [in (e) and (f)].

indices at FWI = 1) and largely corrected with conditional bias correction. This is caused by a near-surface cool (ensemble mean error of -1.8 K) and wet (ensemble mean error of 0.46 g kg^{-1}) bias. The cool temperature mean error and effectiveness of conditional bias correction are consistent with the findings from Erickson et al. (2012; their Fig. 4). In terms of statistical significance, sequential bias correction significantly (defined as exceeding 95% confidence using statistical bootstrapping) improves forecasts of the FWI. Additional significant improvement occurs with conditional bias correction for FWI values less than 3.

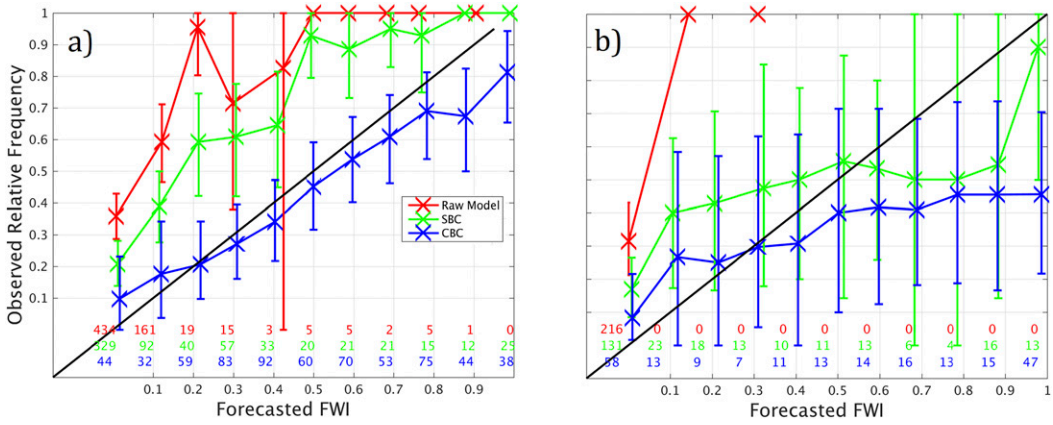

FIG. 13. Average reliability for the raw (red), SBC (green), and CBC (blue) at FWI ≥ 2 using (a) SREF1 and (b) SREF2. Black line shows the 1:1 probability between forecast and observation, numbers at the bottom of the image show the sample size for each bin and error bars denote the 95% confidence intervals.

Regarding the vertical structure of model bias, raw temperature (specific humidity) mean error is more negative (positive) near the surface, which gradually reduces to a mean error near zero above 700 hPa. Sequential bias correction significantly improves the temperature and specific humidity mean error below 900 hPa. Conditional bias correction is more effective at improving mean error than sequential bias correction for low-level temperature and specific humidity. Conditional bias correction has a minor positive impact on mean absolute error compared to sequential bias correction. Overall, any bias-correction method is better than using raw model data on FWDs. This is encouraging, since it is straightforward in practice to implement sequential bias correction when predicting FWDs, although conditional bias correction is recommended. These results are similar to the findings of Erickson et al. (2012) but extend verification above the surface, suggesting that the moist model bias exists in later versions of both WRF cores, regardless of the physics configurations, and propagates upward throughout the PBL.

Variations in spatial mean error and mean absolute error exist within the Northeast U.S. domain subset, even after conditional bias correction is applied. For instance, temperature mean error averages 0.07 K away from the coast and −0.13 K near the coastal plan. While this may be related to representativeness errors associated with interpolation, a more complex spatially varying additive bias correction may be beneficial.

Furthermore, the ensemble exhibits differences in biases related to model core. For instance, raw 2-m temperature WRF-ARW is significantly more negatively biased (by −0.7 K) than the WRF-NMM prior to 2012. The Eta and RSM exhibit significantly greater 2-m specific humidity mean error (averaging 0.30 g kg^{-1} ME) than the WRF-ARW and WRF-NMM prior to 2012. There are no significant differences between each of the cores of the ensemble after 2012. The raw 2-m specific humidity biases are significantly greater after 2012 compared to before 2012 (averaging 0.40 g kg^{-1}), which increases the FWI raw average bias (by 0.2 indices at FWI = 1). This degradation occurs only at the surface, with a significant improvement found after 2012 for specific humidity between 750 and 975 hPa and for temperature between 825 and 1000 hPa. These results suggest that 2-m specific humidity values should be used with caution in the latter SREF period.

SREF reliability plots for the FWI are underdispersed even after bias correction, although probabilistic skill increases with conditional bias correction. This ensemble underdispersion is consistent with Erickson et al. (2012), which used Bayesian model averaging (BMA; Raftery et al. 2005) to inflate the ensemble predictions and improved ensemble reliability on high fire risk days. Bayesian

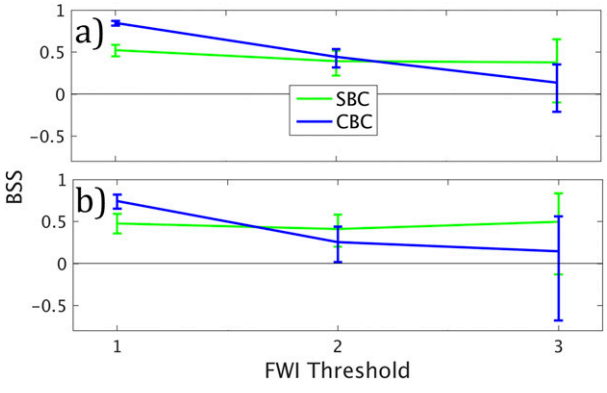

FIG. 14. Brier skill scores by FWI threshold for the SBC (green) and CBC (blue) referenced against the raw values for (a) SREF1 and (b) SREF2. Error bars denote the 95% confidence intervals.

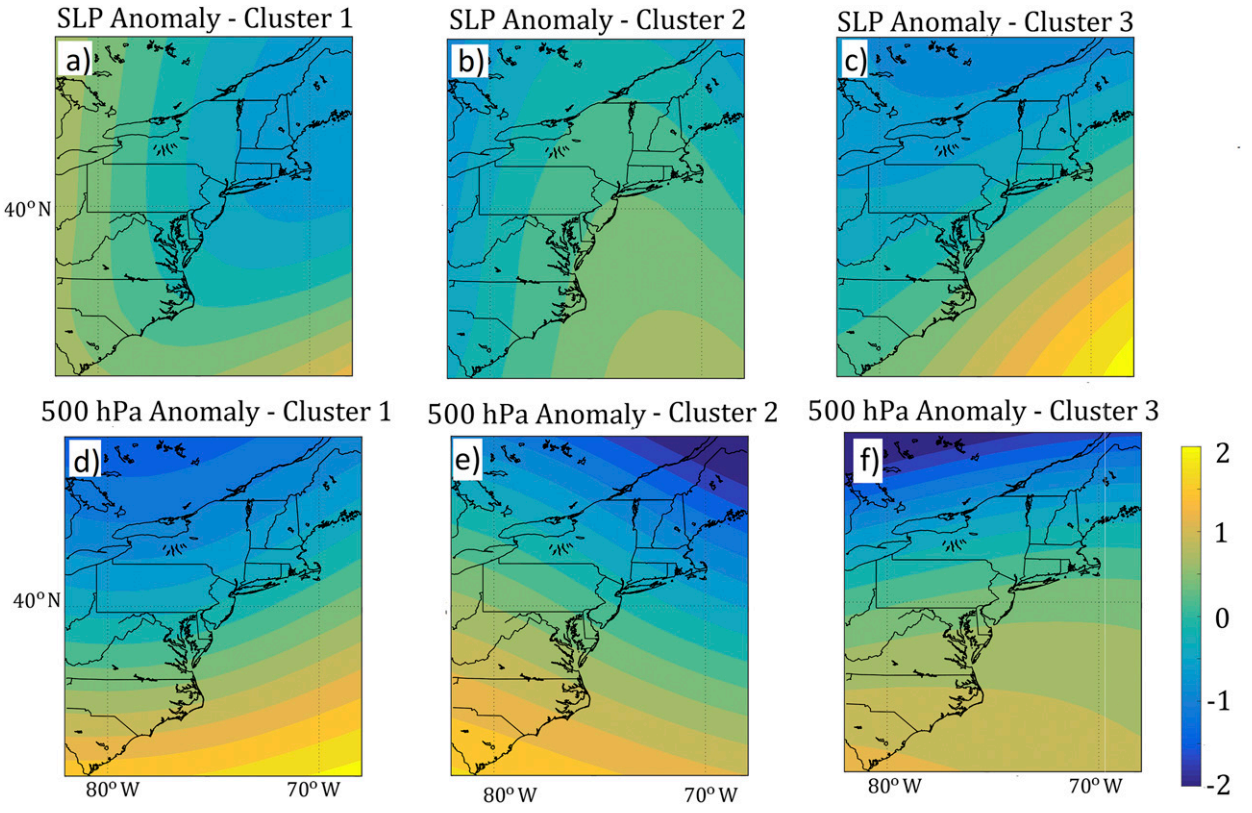

FIG. 15. Composites by cluster for (a)–(c) sea level pressure and (d)–(f) 500-hPa height anomaly over the NEUS. Note that events having an intracluster silhouette value of greater than 0.1 are retained for the composites.

model averaging could also be applied to the results in this study, but would have to be adapted to gridded data (Berrocal et al. 2007) beforehand.

Cluster analysis results suggest that the 1000-hPa temperature model bias appears to be sensitive to the regional weather regime, with less negative biases during an FWD with a high pressure system centered to the southeast of Long Island. This is similar to the ''back of high'' regime described in Pollina et al. (2013). The magnitude of model bias also appears to increase as the FWI value increases for the clusters analyzed.

Although the FWI should be used with caution outside of the Northeast U.S. subdomain in this study, it could be useful when applied in context with other more familiar indices such as the National Fire Danger Rating System and Canadian Forest Fire Danger Rating System. It is important to note that the FWI values have a specific statistical relationship to fire occurrence within the Northeast U.S. subdomain, and this index can be readily tuned to other domains where fire occurrence data are available. A variant of the logistic regression employed here can be used to isolate FWDs, and possibly even wildfire size, during a fire season over more active regions of the world. This would allow for conditional model biases during FWDs to be isolated from the bulk model biases presented in previous studies (e.g., Hoadley et al. 2004, 2006; Mölders 2008; Simpson et al. 2014a,b).

The results from this study present a general idea of what a conditional bias correction could achieve operationally. Several improvements could be made to make the bias correction more effective. The most obvious adjustment is to bias correct each variable based on an optimized threshold or cumulative distribution function (also known as quantile matching) bias-correction method. The bias correction could be adapted to consider additional conditional model biases such as season and month, which do exist but are not presented in this study.

Given the intercluster variability in model bias, the cluster analysis technique can be used to develop an operational FWD analog bias correction. While this bias-correction approach would be of great benefit to the fire weather community, it would require several years of model data for training, preferably from an up-to-date reforecast. The analog bias correction would require two steps: first, determine if an upcoming day appears to be an FWD; second, assign the upcoming FWD into the proper weather regime. The first step can be as simple as applying a binomial threshold (i.e., if temperature and

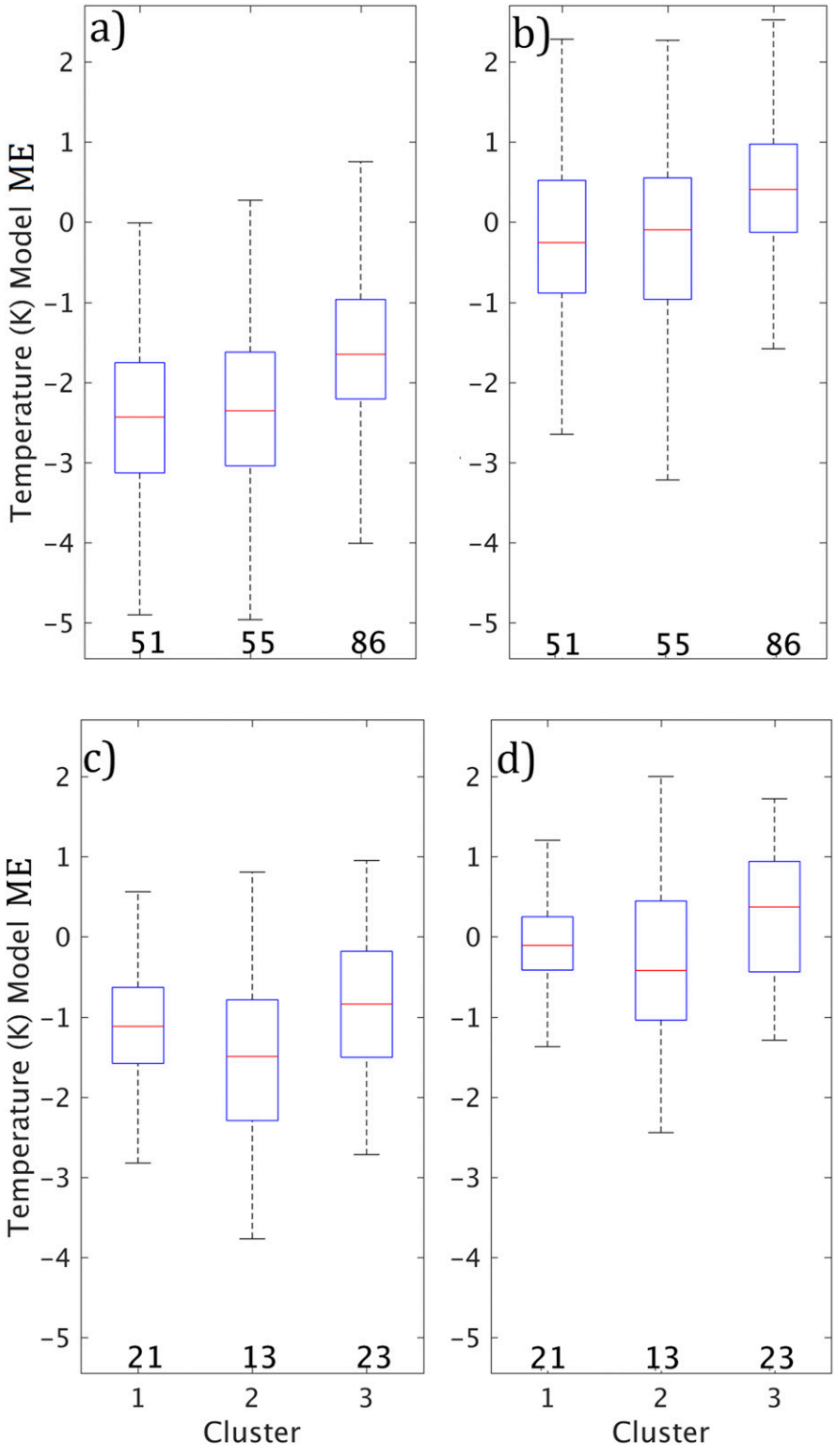

FIG. 16. Box-and-whisker plots of temperature ME by cluster for (a) SREF1 raw, (b) SREF1 CBC, (c) SREF2 raw, and (d) SREF2 CBC (d). The red line denotes the median, the box edges denote the 25th/75th percentiles, and the tails denote all points not considered outliers. Numbers at the bottom of each subplot are the sample size of each bin.

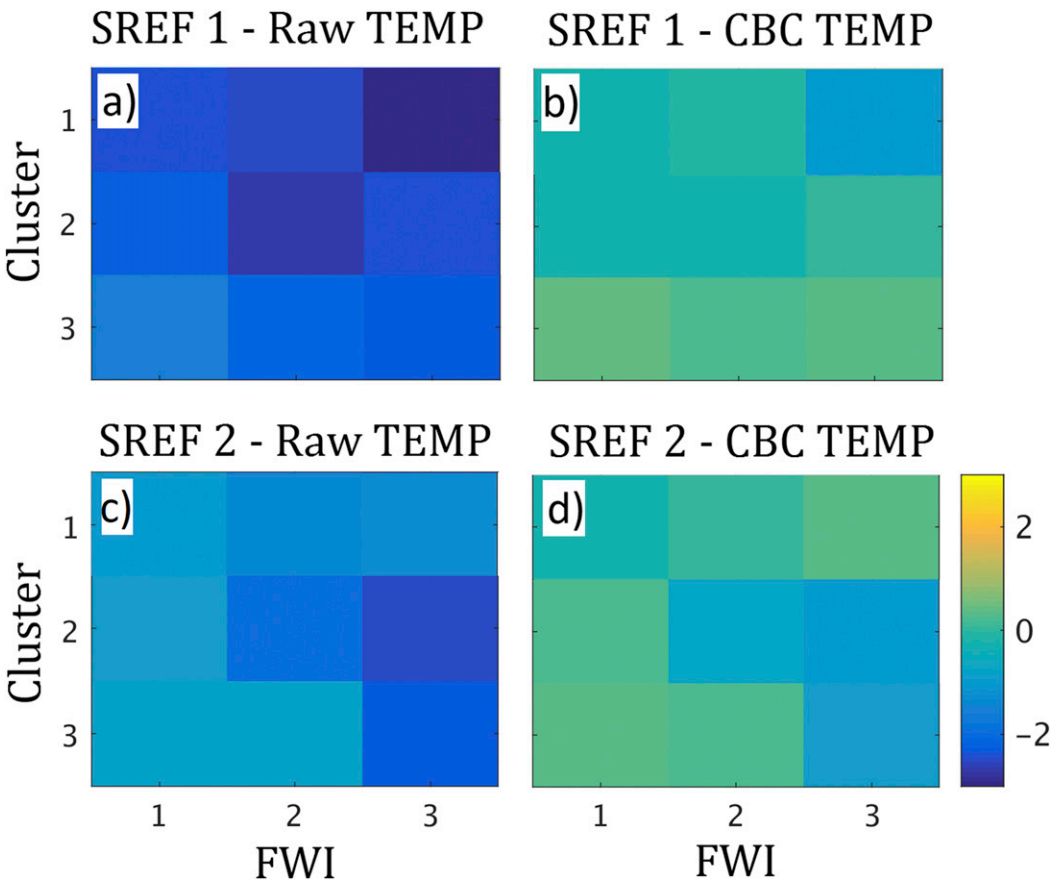

FIG. 17. Temperature ME by cluster and FWI category for (a) SREF1 raw, (b) SREF1 CBC, (c) SREF2 raw, and (d) SREF2 CBC.

relative humidity reach a critical threshold), while the second step would assign the upcoming FWD into the proper weather regime using cluster analysis or a combination of clustering techniques. Additional work would be needed to independently verify the effectiveness of an operational analog technique for FWDs.

Finally, the statistical techniques employed in this study can be used to explore and reduce structural model error within the parameterized model physics. It is possible that some of the cool temperature and high specific humidity mean error on FWDs is the result of a lack of evapotranspiration in the observed atmosphere before and during the spring bloom. However, monthly biases are higher on FWDs for all seasons (not shown). Furthermore, the greater mean error with higher FWI value suggests an amplification of model mean error occurring on the most intense FWDs. The source of this bias is likely related to the fluxes of heat and moisture from the ground to the bottom layer of the atmosphere. Unfortunately, this could be caused by many factors, including the land surface model, planetary boundary layer scheme, soil moisture biases, spurious cloud formation, and possibly nonlinear combinations of one or more of these factors. Future work should focus on optimizing the bottom boundary of the model to isolate sources of potential biases. Data assimilation could prove useful in exploring how rapidly model biases grow on the most severe FWDs.

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). We thank the three anonymous reviewers for their constructive comments to help improve this manuscript.

REFERENCES

Benjamin, S. G., and Coauthors, 2004: An hourly assimilation–forecast cycle: The RUC. *Mon. Wea. Rev.*, **132**, 495–518, https://doi.org/10.1175/1520-0493(2004)132<0495:AHACTR>2.0.CO;2.

——, and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, https://doi.org/10.1175/MWR-D-15-0242.1.

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, https://doi.org/10.1175/MWR3341.1.

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, https://doi.org/10.1071/WF9980159.

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, https://doi.org/10.1071/WF08191.

——, and ——, 2013: The diagnosis of mixed-layer depth above an eastern U.S wildfire using a mesoscale numerical weather prediction model. *Fourth Fire Behavior and Fuels Conf.*, Raleigh, NC, International Association of Wildfire Fire, 24, http://www.iawfonline.org/2013FuelsConference/Final%20Program%20Booklet.pdf.

Deeming, J. E., J. W. Lancaster, M. A. Fosberg, R. W. Furman, and M. J. Schroeder, 1972: The National Fire-Danger Rating System. USDA Forest Service Rocky Mountain Forest and Range Experiment Station Research Paper RM-84, 165 pp.

Du, J., and Coauthors, 2012: New 16km NCEP Short-Range Ensemble Forecast (SREF) system: What we have and what we need? National Centers for Environmental Prediction, 25 pp., http://www.dtcenter.org/events/workshops12/nuopc_2012/Presentations/SREF_2012ensembleworkshop_JDu.pdf.

Erickson, M. J., B. A. Colle, and J. 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, https://doi.org/10.1175/WAF-D-11-00149.1.

——, J. J. Charney, and B. A. Colle, 2016: Development of a fire weather index using meteorological observations within the northeast United States. *J. Appl. Meteor. Climatol.*, **55**, 389–402, https://doi.org/10.1175/JAMC-D-15-0046.1.

Hamill, T. M., G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau Jr., Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565, https://doi.org/10.1175/BAMS-D-12-00014.1.

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. Climatol.*, **43**, 1333–1347, https://doi.org/10.1175/JAM2146.1.

——, 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, https://doi.org/10.1071/WF05015.

Huth, R., C. Beck, A. Phillip, M. Demuzere, Z. Ustrnul, M. Cahynova, K. Kysely, and O. E. Tveito, 2008: Classifications of atmospheric circulation patterns: Recent advances and applications. *Ann. N. Y. Acad. Sci.*, **1146**, 105–152, https://doi.org/10.1196/annals.1446.019.

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, https://doi.org/10.1007/s00703-008-0004-5.

Lloyd, S. P., 1982: Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, **28** (2), 129–137, https://doi.org/10.1109/TIT.1982.1056489.

McFadden, R. D., 1995: Woodland fire finally tamed on Long Island. *New York Times*, accessed 20 March 2017, http://www.nytimes.com/1995/08/27/nyregion/fire-on-long-island-the-overview-woodland-fire-finally-tamed-on-long-island.html.

Mesinger, F., and Coauthors, 2006: North American Regional Reanalysis. *Bull. Amer. Meteor. Soc.*, **87**, 343–360, https://doi.org/10.1175/BAMS-87-3-343.

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, https://doi.org/10.1175/2008WAF2007062.1.

National Snow and Ice Data Center, 2008: IMS daily Northern Hemisphere snow and ice analysis at 4 km resolution. NSIDC, accessed 20 March 2017, http://dx.doi.org/10.7265/N52R3PMC.

Pollina, J. B., B. A. Colle, and J. J. Charney, 2013: Climatology and meteorological evolution of major wildfire events over the northeast United States. *Wea. Forecasting*, **28**, 175–193, https://doi.org/10.1175/WAF-D-12-00009.1.

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, https://doi.org/10.1175/MWR2906.1.

Rousseeuw, P. J., and A. M. Leroy, 1987: *Robust Regression and Outlier Detection.* Wiley Series in Applied Probability and Statistics, John Wiley & Sons, 329 pp., https://doi.org/10.1002/0471725382.

Simpson, C. C., H. G. Pearce, A. P. Sturman, and P. Zawar-Reza, 2014a: Verification of WRF modelled fire weather in the 2009–10 New Zealand fire season. *Int. J. Wildland Fire*, **23**, 34–45, https://doi.org/10.1071/WF12152.

——, ——, ——, and ——, 2014b: Behaviour of fire weather indices in the 2009–10 New Zealand wildland fire season. *Int. J. Wildland Fire*, **23**, 1147–1164, https://doi.org/10.1071/WF12169.

Van Wagner, C. E., 1987: Development and structure of the Canadian Forest Fire Weather Index System. Canadian Forestry Service Tech. Rep. 35, 37 pp., http://cfs.nrcan.gc.ca/pubwarehouse/pdfs/19927.pdf.

Weygandt, S. S., and Coauthors, 2013: Hourly updated models: Rapid Refresh/HRRR review. NCEP Production Suite Overview, https://rapidrefresh.noaa.gov/internal/pdfs/NCEP_PSR_2013_RAP_FINAL_v5.pdf.

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

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, https://doi.org/10.1175/MWR3347.1.

Yan, E., and Coauthors, 2012: Long Island firefighters tackle huge brush fire. *Newsday*, accessed 20 March 2017, http://www.newsday.com/long-island/suffolk/breaking/long-island-firefighters-tackle-huge-brush-fire-1.3650232.