Application of a Hybrid Statistical–Dynamical System to Seasonal Prediction of North American Temperature and Precipitation

SARAH STRAZZO

NOAA/NWS/NCEP/Climate Prediction Center, College Park, and Innovim, LLC, Greenbelt, Maryland

DAN C. COLLINS

NOAA/NWS/NCEP/Climate Prediction Center, College Park, Maryland

ANDREW SCHEPEN

CSIRO Land and Water, Dutton Park, Queensland, Australia

Q. J. WANG

The University of Melbourne, Parkville, Victoria, Australia

EMILY BECKER AND LIWEI JIA

NOAA/NWS/NCEP/Climate Prediction Center, College Park, and Innovim, LLC, Greenbelt, Maryland

(Manuscript received 2 May 2018, in final form 15 November 2018)

ABSTRACT

Recent research demonstrates that dynamical models sometimes fail to represent observed teleconnection patterns associated with predictable modes of climate variability. As a result, model forecast skill may be reduced. We address this gap in skill through the application of a Bayesian postprocessing technique—the calibration, bridging, and merging (CBaM) method—which previously has been shown to improve probabilistic seasonal forecast skill over Australia. Calibration models developed from dynamical model reforecasts and observations are employed to statistically correct dynamical model forecasts. Bridging models use dynamical model forecasts of relevant climate modes (e.g., ENSO) as predictors of remote temperature and precipitation. Bridging and calibration models are first developed separately using Bayesian joint probability modeling and then merged using Bayesian model averaging to yield an optimal forecast. We apply CBaM to seasonal forecasts of North American 2-m temperature and precipitation from the North American Multimodel Ensemble (NMME) hindcast. Bridging is done using the model-predicted Niño-3.4 index. Overall, the fully merged CBaM forecasts achieve higher Brier skill scores and better reliability compared to raw NMME forecasts. Bridging enhances forecast skill for individual NMME member model forecasts of temperature, but does not result in significant improvements in precipitation forecast skill, possibly because the models of the NMME better represent the ENSO–precipitation teleconnection pattern compared to the ENSO–temperature pattern. These results demonstrate the potential utility of the CBaM method to improve seasonal forecast skill over North America.

1. Introduction

Seasonal climate forecasts provide valuable information for a number of climate-sensitive societal sectors, including agriculture, energy, and public health (e.g., Challinor et al. 2005; Hawkins et al. 2013; Shukla et al. 2014; Tompkins and Di Giuseppe 2015; Torralba et al. 2017). Over time, seasonal climate prediction has evolved from an endeavor relying primarily on statistical modeling (e.g., van den Dool 2007) to one that increasingly utilizes dynamical climate models (e.g., Saha et al. 2014; Jia et al. 2015; MacLachlan et al. 2015, and others). In particular, multimodel ensembles such as the North American Multimodel Ensemble (NMME; Kirtman et al. 2014) tend to produce more skillful and statistically reliable forecasts compared to individual

Corresponding author: Sarah Strazzo, sarah.strazzo@noaa.gov

DOI: 10.1175/MWR-D-18-0156.1

model ensemble systems, likely a result of the cancellation of uncorrelated model errors (Hagedorn et al. 2005). Although dynamical models, particularly multimodel ensembles, yield skillful predictions of tropical climate, forecasts of extratropical climate remain relatively less skillful (Doblas-Reyes et al. 2013). For example, while Becker and van Den Dool (2016) found Brier skill scores from probabilistic NMME forecasts of SST in the Niño-3.4 region to be as high as 0.68, skill scores from probabilistic forecasts of midlatitude Northern Hemisphere 2-m temperature did not exceed 0.14. In light of these deficiencies, a number of statistical postprocessing methods have emerged to improve the skill and reliability of ensemble and multimodel ensemble forecasting systems (e.g., Unger et al. 2009; Schepen et al. 2014; Zhang et al. 2017; Narapusetty et al. 2018). Here we apply one such method—the calibration, bridging, and merging (CBaM) method (Schepen et al. 2014, 2016)—in an effort to improve the seasonal forecast skill and reliability of the NMME.

The CBaM methodology relies on Bayesian statistical modeling to postprocess dynamical model forecasts with an ultimate goal of generating hybrid statistical–dynamical forecasts that are free of bias and reliable in conveying forecast uncertainty. The calibration component of CBaM consists of a statistical model relating dynamical model forecasts of temperature (precipitation) to observed temperature (precipitation). Once established, a calibration model can be used to correct new dynamical model forecasts. Schepen et al. (2014) demonstrated that this calibration approach improves the accuracy and reliability of dynamical model forecasts that already exhibit underlying skill.

As Schepen et al. (2016) note, dynamical models sometimes fail to accurately represent teleconnection patterns associated with critical drivers of climate variability (e.g., ENSO). In such instances, we need an alternative postprocessing method to correct for model teleconnection errors. Using seasonal forecasts from the Australian POAMA model, Schepen et al. (2014, 2016) found that bridging—the second component of CBaM—improved forecast skill beyond what was achieved through calibration for some seasons and regions. Bridging models are established very similarly to calibration models, but instead relate dynamical model forecasts of relevant remote drivers of climate variability to observed temperature and precipitation. For example, Schepen et al. (2016) developed bridging models using POAMA forecasts of several ENSO indices and the Indian Ocean dipole mode index as predictors of minimum and maximum temperature over Australia. They then applied Bayesian model averaging to merge—the final component to CBaM—the calibrated and bridged forecasts. The resulting fully merged CBaM forecasts achieved higher skill scores and better statistical reliability than raw mean-corrected forecasts. Additionally, Peng et al. (2014) applied the CBaM method to postprocess ECMWF System 4 precipitation forecasts over China and similarly found that the fully merged forecasts produced higher skill scores than calibrated forecasts. While published research has explored the application of CBaM to individual model ensembles, the method has not yet been applied to postprocess multimodel ensemble systems such as the NMME.

Importantly, recent research reveals that at short lead times the NMME sometimes fails to represent the ENSO–temperature teleconnection pattern over North America (Chen et al. 2017). Because ENSO is the dominant source of seasonal climate predictability over North America and is often used as a benchmark for judging dynamical models (e.g., van Oldenborgh et al. 2005; Xue et al. 2013), this particular model shortcoming has the potential to degrade forecast skill over North America. In light of this research and given the demonstrated skill of the CBaM methodology using the POAMA model, here we apply CBaM to postprocess the NMME hindcast dataset. We specifically seek to understand 1) whether statistical–dynamical bridging enhances forecast skill in the NMME beyond what is achieved through calibration, 2) whether CBaM improves the statistical reliability of the NMME, and 3) how CBaM forecasts from individual models compare to CBaM forecasts from a multimodel ensemble. Because 1) ENSO is the most predictable mode of climate variability influencing U.S. temperature and precipitation on seasonal time scales (Goddard et al. 2001), and 2) previous research suggests that the NMME does not capture the observed ENSO influence on North American climate, we limit this initial study to focus on bridging using an ENSO index. Subsequent research will expand this work to examine additional bridging predictors for global seasonal climate prediction.

Finally, we also are interested in developing tools to improve seasonal temperature and precipitation prediction for the National Oceanic and Atmospheric Administration's Climate Prediction Center (CPC). Given this interest, we seek to develop and apply the CBaM method in a manner that is consistent with current operational practices at CPC. Although the research presented here is not intended to serve as a comprehensive survey of the wide array of statistical postprocessing tools available to forecasters, we do include some results comparing CBaM to another calibration method currently in use at CPC.

The remainder of the paper is organized as follows. In section 2, we first provide information about the NMME

and observed data and then describe the CBaM method and verification metrics used to assess forecast skill. Section 3 briefly compares observed versus NMME ENSO teleconnection patterns. Sections 4 and 5 present CBaM results for NMME temperature and precipitation reforecasts, respectively. Finally, section 6 provides a summary and discussion of the results.

2. Data and methods

a. NMME and observed data

We use monthly mean model reforecast precipitation rate, 2-m temperature, and sea surface temperature (SST) data from Phase I of the NMME hindcast, which covers the period 1982–2010 (Kirtman et al. 2014; NOAA/NSF/NASA/DOE 2014). In all, seven models are included: the NCEP Climate Forecast System, version 2 (CFSv2; Saha et al. 2014), the Canadian Centre for Climate Modelling and Analysis Third and Fourth Generation Canadian Coupled Global Climate Model (referred to here as CMC1 and CMC2; Merryfield et al. 2013), version 2.2 of the Geophysical Fluid Dynamics Laboratory climate model (GFDL; Zhang et al. 2007), the Forecast-Oriented Low Ocean Resolution version of GFDL climate model 2.5 (GFDL-FLOR; Vecchi et al. 2014; Jia et al. 2015), the NASA Goddard Earth Observing System model, version 5 (NASA; Vernieres et al. 2012), and the Community Climate System Model, version 4 (NCAR-CCSM4; Gent et al. 2011). Details pertaining to Phase I NMME data can be found in Kirtman et al. (2014). (The data are available for download at http://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME/.) As an initial test, we focus on 1-month lead forecasts of 2-m temperature and precipitation rate for the 12 overlapping 3-month seasons. We define a 1-month lead forecast as a forecast target period beginning one month after the model initial date. For example, for an NMME forecast initialized in early November, the 1-month lead seasonal forecast period would be December–January–February (DJF).

Development of the statistical–dynamical models and verification of the resulting forecasts relies on observed 2-m temperature data from the Global Historical Climatology Network and Climate Anomaly Monitoring System (GHCN-CAMS; NOAA/OAR/ESRL/PSD 2008; Fan and van den Dool 2008), observed SST data from the NOAA Optimum Interpolation SST dataset (NOAA/OAR/ESRL/PSD 2002; Banzon et al. 2016), and observed precipitation rate data from the CPC Merged Analysis of Precipitation dataset (CMAP; Xie and Arkin 1997). CMAP (precipitation) and GHCN-CAMS (temperature) data are used for verification of seasonal forecast tools at CPC. Although the observed data are available on 1/4° (SST), 1/2° (2-m temperature), and 2.5° (precipitation) grids, we linearly interpolate these to match the 1° grid spacing of the NMME data.

When verifying forecasts, we subset on grid points over land surfaces and compare results for the entire North American continent. Observed and forecast Niño-3.4 index values are calculated as the three-month seasonal mean anomalies over the Niño-3.4 region (5°N–5°S, 170°–120°W). We obtain model mean forecast Niño-3.4 anomalies for each of the seven dynamical models by subtracting the lead-dependent model mean climatology (1982–2010) from the reforecast SSTs over the Niño-3.4 region. Note that for the CFSv2, we follow Xue et al. (2013) and remove the 1982–98 and 1999–2010 climatologies separately to account for a discontinuity in the hindcast.

b. Calibration, bridging, and merging method

The CBaM method relies on Bayesian statistical modeling to postprocess and merge dynamical model forecasts. We apply Bayesian joint probability modeling (BJP; Wang et al. 2009) to generate calibrated and bridged forecasts, and then merge the calibrated and bridged forecasts using Bayesian model averaging (BMA; Raftery et al. 1997; Hoeting et al. 1999; Wang et al. 2012a). We provide a brief description of the BJP and BMA methods below and refer readers to Schepen et al. (2014, 2016) for a more detailed mathematical derivation of the posterior and predictive distributions. Additionally, Table 1 provides a concise summary of the postprocessing steps applied here.

1) Bayesian Joint Probability Modeling

Both bridging and calibration models are Bayesian joint probability models that relate dynamical model output to observed climate variables (temperature or precipitation). A successful calibration model corrects both model bias and ensemble spread in raw model forecasts and returns the forecasts to climatology in the absence of a correlation between the forecasts and observations. Calibration models are developed using raw dynamical model reforecasts of 2-m temperature (precipitation) over North America as the predictor of observed 2-m temperature (precipitation) over North America. In contrast, bridging models use dynamical model reforecasts of Niño-3.4 anomalies as the predictor of 2-m temperature or precipitation over North America, although any relevant climate index may be employed as the predictor. The BJP method used to establish bridging and calibration models involves modeling the predictor and predictand using a continuous bivariate normal distribution.

TABLE 1. Summary of the calibration, bridging, and merging frameworks.

	Calibration	Bridging	Merging
Method	Bayesian joint probability modeling	Bayesian joint probability modeling	Bayesian model averaging
Predictor	Temperature or precipitation	Niño-3.4 index	—
Predictand	Temperature or precipitation	Temperature or precipitation	—

To make this possible, the predictor and predictand are first transformed using the Yeo–Johnson transformation for temperature and climate index data, or the log-sinh transformation for precipitation data (Yeo and Johnson 2000; Wang et al. 2012b). A simple K-S test shows that transformation of temperature data may not be necessary for the majority of grid points over North America, with perhaps a small number of high-latitude exceptions. Indeed, we find that application of the Yeo–Johnson transformation to temperature data does not significantly affect our results. We apply the transformation to temperature data out of an abundance of caution. In contrast, transformation is necessary when working with precipitation data, which are often better described by a gamma or similar distribution. Other methods at CPC (e.g., ensemble regression) typically apply a third or fourth root power transform to make the non-Gaussian precipitation data appear more Gaussian. However, the log-sinh transformation previously has been shown to satisfactorily transform precipitation data for use in BJP, which is why we select this method for the present study. We refer readers to Schepen et al. (2014) for a detailed description of the Yeo–Johnson and log-sinh transformations and their parameters. Here, we simply denote a generic normalizing transformation function ψ with parameters $\boldsymbol{\Delta}$. The transformed predictor is $x = \psi_{\boldsymbol{\Delta}}(x_o)$ and the transformed predictand is $y = \psi_{\boldsymbol{\Delta}}(y_o)$, where x_o and y_o are the original predictor and predictand variables, respectively. We assume that the joint distribution of the transformed variables is bivariate normal, that is,

$$x, y \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{1}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the means and covariance matrix parameters from the bivariate normal distribution, respectively. The covariance matrix embeds ρ, the correlation between x and y.

Model parameter inference proceeds in two phases. In the first phase, the transformation parameters for x_o and y_o are inferred using data $\mathbf{X}_o = (x_{o,1}, \ldots, x_{o,n})$ and $\mathbf{Y}_o = (y_{o,1}, \ldots, y_{o,n})$, respectively. We estimate a single "best" set of transformation parameters using a Bayesian maximum a posteriori (MAP) solution.

In the second phase, the BJP bivariate normal distribution parameters are estimated from $\mathbf{D} = [(x_1, y_1), \ldots, (x_n, y_n)]$, a sequence of transformed predictor–predictand data pairs, for $n = 29$ years. In contrast to the inference of the transformation parameters, the inference of the bivariate normal parameters allows for parameter uncertainty. Let $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. A Gibbs sampler is used to obtain m samples from the posterior distribution:

$$p(\boldsymbol{\theta}|\mathbf{D}) = p(\boldsymbol{\theta})p(\mathbf{D}|\boldsymbol{\theta}), \tag{2}$$

where $p(\boldsymbol{\theta})$ is the prior distribution of the parameters and $p(\mathbf{D}|\boldsymbol{\theta})$ is the likelihood. A noninformative prior is specified. The collection of sampled parameter sets is $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m)$. Here we obtain a sample of $m = 1000$ parameter sets.

Once the model parameter sets have been sampled, BJP can be used in predictive mode. For each of the m bivariate normal parameter sets, a Gibbs sampler is used to obtain a single sample of $y|x^*, \boldsymbol{\theta}_i$, where x^* is a new transformed predictor value. All of these samples of y are collected in $\mathbf{Y}^* = (y_1^*, \ldots, y_m^*)$, which is a numerical sample representative of the posterior predictive distribution:

$$f(y) = \int p(y|x, \boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{D})\, d\boldsymbol{\theta}\,. \tag{3}$$

Moreover, $\mathbf{Y}^*$ is a forecast from a BJP model, albeit normally distributed. The forecasts are backtransformed to the original space using the appropriate inverse transformation $\psi_{\boldsymbol{\Delta}}^{-1}(y)$ to obtain $\mathbf{Y}_O^*$.

We develop separate bridging and calibration BJP models for each grid point, initial time, and lead for each of the seven NMME member models. Note that for each of the seven models, we do not develop separate BJP models for the individual dynamical model members but rather use the model ensemble mean. For example, consider a 1-month lead CFSv2 forecast for a single grid point initialized in November. Rather than develop 24 bridging and 24 calibration models for each of the 24 CFSv2 individual members, we develop one bridging and one calibration model using the CFSv2 model mean. Therefore, this method does not directly incorporate dynamical model spread information. However, when we compare BJP calibration with the ensemble regression calibration method (Unger et al. 2009), which does incorporate dynamical model spread into calibrated forecast probabilities, we find that the two methods yield similarly skillful and reliable forecasts (not shown).

BJP models for temperature and precipitation generally follow the same development work flow, with some exceptions. First, as noted previously, temperature data are transformed using the Yeo–Johnson transformation while precipitation data are transformed using the log-sinh transformation. Additionally, BJP models used to postprocess precipitation forecasts must account for zero values, which occasionally occur for grid points in the southwestern United States. We follow Wang and Robertson (2011) and treat zero values as censored data with unknown values below or near zero. This allows us to work within the framework of a continuous bivariate normal distribution. When forecasting, predicted values below zero are converted to zero.

Finally, we apply leave-one-year-out cross validation to test the method over the NMME hindcast period. For a given year, BJP bridging and calibration models are developed from the rescaled and transformed data after removing the predictor–predictand pair for that year, such that the data rescaling and transformation and BJP model development are all done within each cross-validation fold. The resulting BJP models are then used to generate a calibrated and bridged forecast for the removed year using the forecast predictor value from that year. While concern exists that the relatively high autocorrelation of the Niño-3.4 index could introduce artificial skill when leaving out only one year, Schepen et al. (2014) found previously that the use of a more stringent leave-three-years-out cross validation did not affect the results. An alternative method of cross validation might instead develop the BJP models using data from the NMME hindcast period and then apply these BJP models to postprocess NMME data from the real-time period. However, NMME real-time data are only consistently available beginning in 2012, leaving only 6 years of data with which to test BJP. While we have tested this method and find generally positive skill, we choose not to use it here given the lack of a sufficiently large real-time sample for calculation of verification statistics.

2) BAYESIAN MODEL AVERAGING

Once bridged and calibrated forecasts have been generated for a given grid point, we calculate a weighted average of the two forecasts using Bayesian model averaging (Raftery et al. 1997; Wang et al. 2012a). We refer to this step as ''merging.'' For each dynamical model, we create a merged forecast using the calibrated and bridged forecasts for that model. Additionally, we calculate an NMME merged forecast by merging the calibrated and bridged forecasts from all seven dynamical models. Similarly, NMME bridged (calibrated) forecasts are calculated by merging all bridged (calibrated) forecasts from the seven dynamical models. The BMA density forecast, $f_{\mathrm{BMA}}(y)$, is expressed as

$$f_{\mathrm{BMA}}(y) = \sum_{k=1}^{K} w_k f_k(y), \tag{4}$$

where $f_k(y)$ is the density forecast for model k, and w_k is the weight for model k. When merging calibrated and bridged forecasts for individual models (e.g., CFSv2), we use $k = 2$, one bridged forecast and one calibrated forecast. In contrast, to create the NMME bridged (calibrated) forecast, we merge the $k = 7$ member model bridged (calibrated) forecasts. Finally, to create the NMME merged forecast, we merge all member model bridged and calibrated forecasts ($k = 14$). We calculate the MAP estimate of the weights by maximizing the posterior distribution of the weights. As in Schepen et al. (2016), we use a Dirichlet prior distribution, $p(\boldsymbol{\pi})$,

$$p(\boldsymbol{\pi}) \propto \prod_{k=1}^{K} (w_k)^{a-1}, \tag{5}$$

where a is the concentration parameter, $a = 1 + a_0/K$, and a_0 is a free parameter. We set a_0 equal to 1.0 to force more even weights among the models.

Once we have calculated estimates of the BMA weights, we obtain the merged forecast by taking a random sample from each BJP forecast ensemble we seek to merge. The size of the random sample from each forecast ensemble is determined by the weights. The resulting ensemble of forecasts represents $f_{\mathrm{BMA}}(y)$. We again apply leave-one-year-out cross validation to calculate the model weights. We note that BJP forecasts and their associated weights are not computed within the same cross-validation fold. Schepen et al. (2014) examined the impact of this data leakage by estimating weights using forecast–observation pairs for the same events used to fit the BJP model. They found that this method did not significantly influence the results and in fact tended to worsen overfitting. Therefore, in this study we permit some minor data leakage in an effort to minimize model overfitting.

c. *Probabilistic forecast verification metrics*

We generate calibrated, bridged, and merged probabilistic forecasts of above and below normal 2-m temperature (precipitation), where ''normal'' is defined as the middle tercile of the observed 2-m temperature (precipitation) distribution over the hindcast period, 1982–2010. We assess forecast skill using Brier skill score (BSS; Brier 1950; Wilks 2011):

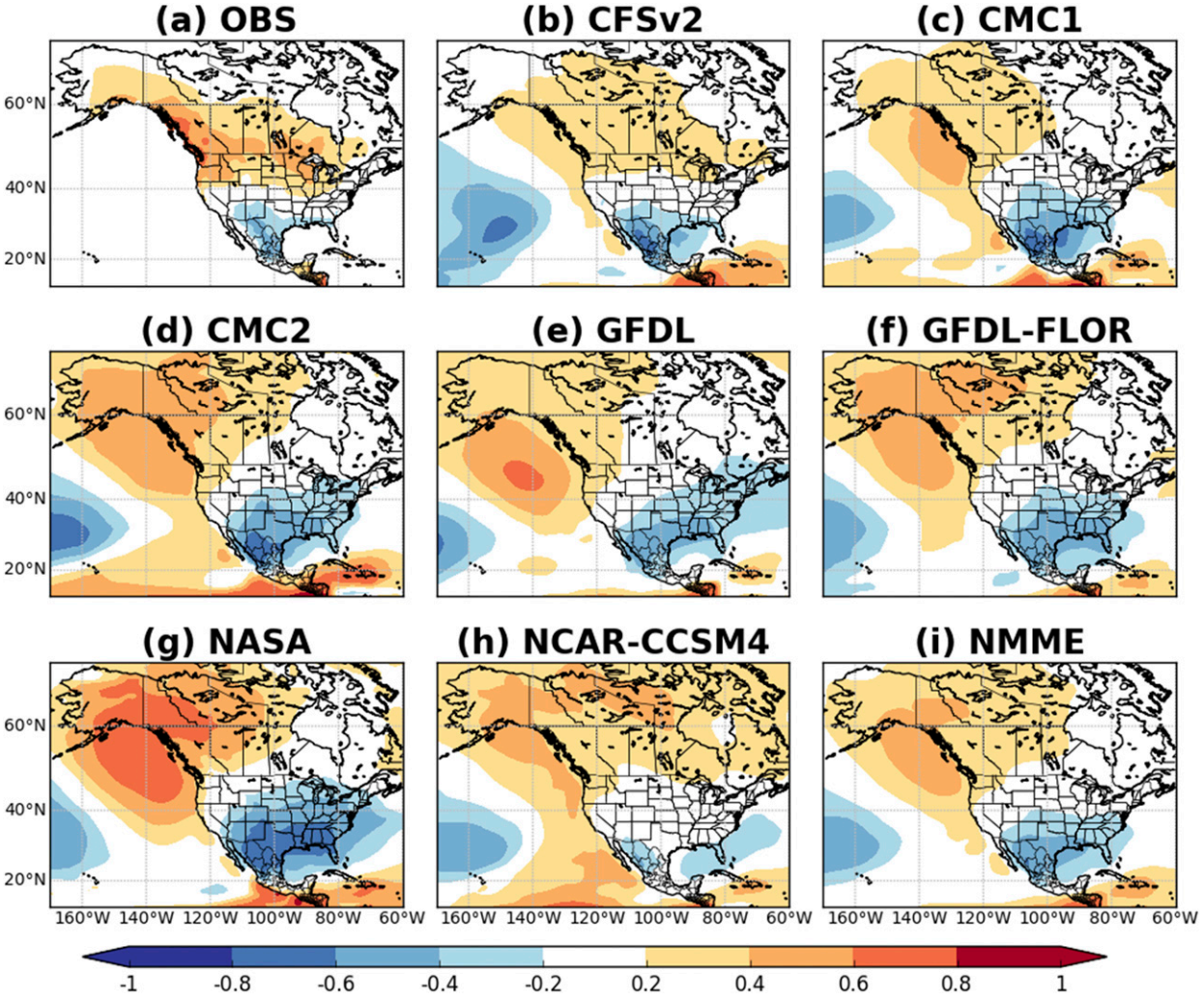

FIG. 1. (a) The correlation between the observed Niño-3.4 index and observed 2-m temperature over North America during DJF over the 1982–2010 period compared with the (b)–(i) correlation between the 1-month lead forecast Niño-3.4 index and 1-month lead forecast 2-m temperature during DJF for each of the 7 NMME member models and the multimodel mean. For each model, the correlations were obtained by averaging the correlations of each individual member.

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}}. \tag{6}$$

The Brier score (BS) is calculated as

$$\mathrm{BS} = \frac{1}{n}\sum_{k=1}^{n}(p_k - o_k)^2, \tag{7}$$

where for a given forecast–event pair, p_k is the forecast probability and o_k is 1 if the event occurred and 0 if it did not. Here we consider two events: 1) below-normal temperature (precipitation) occurs, and 2) above-normal temperature (precipitation) occurs. The term $\mathrm{BS}_{\mathrm{ref}}$ refers to the Brier score of a reference forecast. We use a climatological reference forecast of $p_k = 0.33$. The BSS is positively oriented such that BSS = 1.0 represents a perfect forecast.

We additionally assess the reliability of model forecasts using reliability diagrams (Wilks 2011; Hartmann et al. 2002). Predicted probabilities are binned into 10 separate probability groups ranging from 0–0.1 to 0.9–1.0 and are compared to the observed relative frequency.

3. NMME representation of ENSO and ENSO teleconnection patterns

Because statistical–dynamical bridging uses model forecasts of the Niño-3.4 index to predict temperatures over North America, dynamical models must skillfully

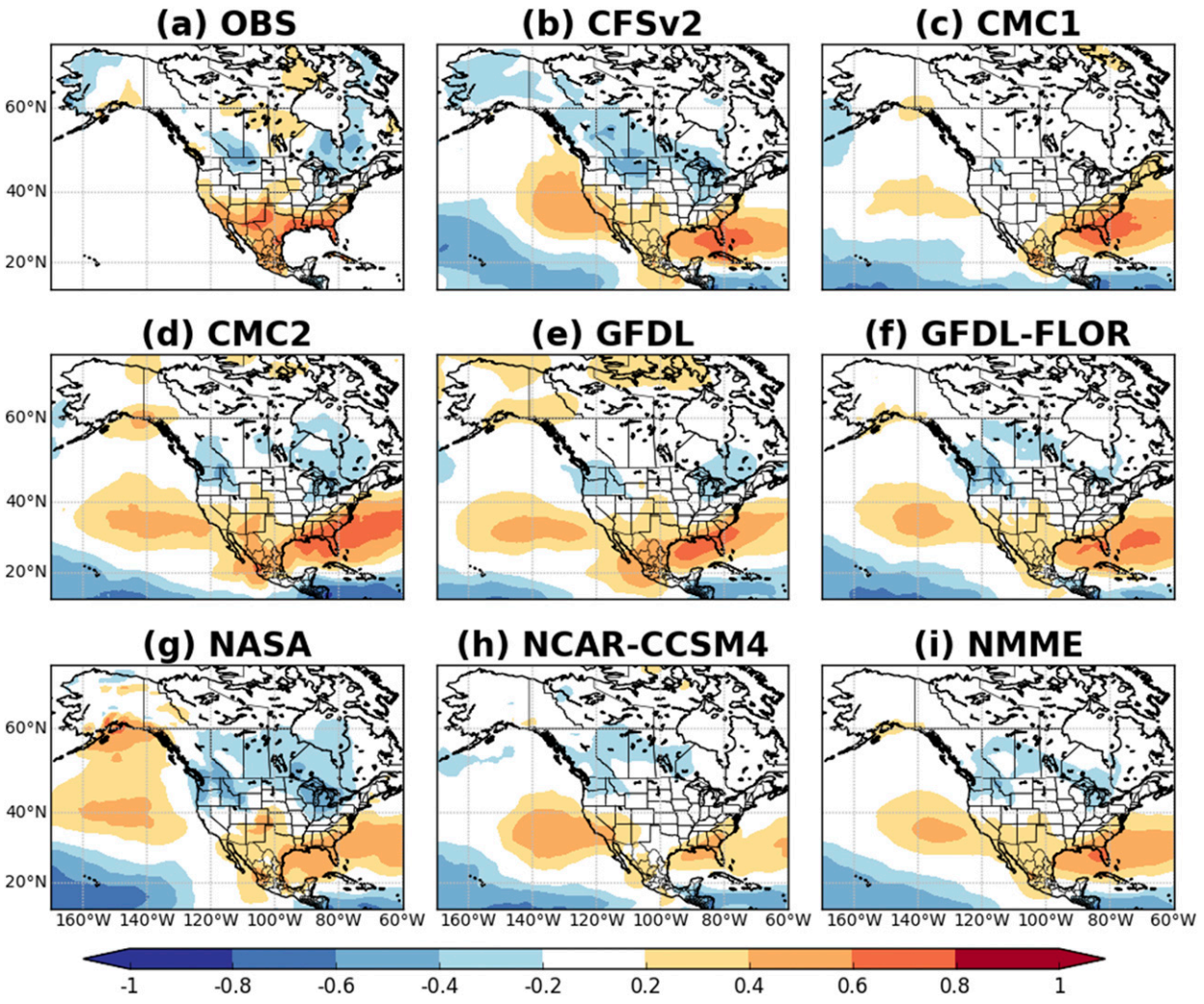

FIG. 2. (a) The correlation between the observed Niño-3.4 index and observed precipitation rate over North America during DJF over the 1982–2010 period compared with (b)–(i) the correlation between the 1-month lead forecast Niño-3.4 index and 1-month lead forecast precipitation rate during DJF for each of the 7 NMME member models and the multimodel mean. For each model, the correlations were obtained by averaging the correlations of each individual member.

predict the Niño-3.4 index for bridging to be successful. Fortunately, previous research demonstrates that the NMME ENSO forecast skill is very high (Barnston et al. 2018). The correlation between forecast and observed Niño-3.4 anomalies exceeds 0.8 for all 12 overlapping seasons for forecasts made with 1-month lead. As expected, forecast skill decreases as lead time increases, although the correlation between the observed and multimodel mean forecast Niño-3.4 index never falls below 0.6. Skill is sufficiently high to attempt statistical–dynamical bridging for forecasts made 1–6 months in advance of the target season.

Bridging enhances forecast skill in instances when models fail to represent the observed teleconnection patterns between ENSO and climate conditions over North America. Chen et al. (2017) documented discrepancies between observed and NMME composite temperature patterns over North America during cold and warm ENSO events. Likewise, we find that several models in the NMME (e.g., NASA and GFDL-FLOR) produce different representations of the ENSO–temperature teleconnection pattern over North America when compared against the observed pattern (Fig. 1). Focusing on DJF—when the ENSO influence on North American climate is strongest—we find that the largest differences occur over the northern and midwestern United States, where the observed correlation between Niño-3.4 anomalies and North American 2-m temperature is positive ($r = 0.4$–0.6) while the correlation between forecast Niño-3.4 anomalies and forecast 2-m temperature is

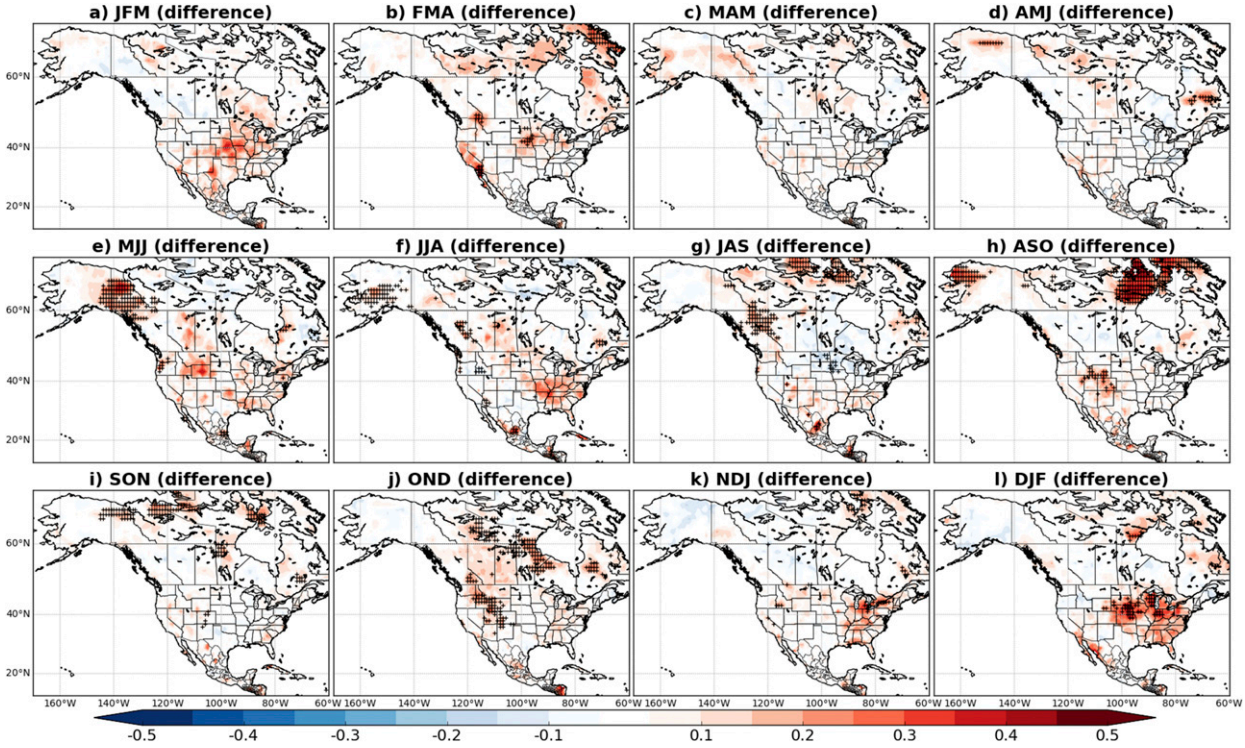

FIG. 3. Shading indicates Brier skill score differences between 1-month lead calibrated and 1-month lead raw forecasts of below-normal 2-m temperature for the NMME for each of the 12 overlapping 3-month seasons. Red shading indicates that calibrated forecast mean Brier skill scores exceeded raw forecast mean Brier skill scores over the 1982–2010 hindcast period. Hatching indicates significance at the 95% confidence level, as determined via a Wilcoxon rank sum test without accounting for field significance.

less than or near zero ($r = -0.6$–0) for the CMC1, CMC2, GFDL, GFDL-FLOR, and NASA models. Taking the multimodel mean (Fig. 1i) results in some improvements in the ENSO–temperature teleconnection pattern; however, discrepancies remain over parts of the northern United States. Although the relatively short hindcast period makes it difficult to determine whether the models truly misrepresent the ENSO teleconnection, any biases in forecast teleconnection patterns have the potential to reduce forecast skill. Statistical–dynamical bridging provides alternative forecasts based on the historical representation of ENSO teleconnection patterns and therefore may improve forecast skill for some of the NMME models.

In contrast to temperature, Chen et al. (2017) found that the ENSO–precipitation teleconnection pattern is generally well represented by the NMME. Examining the correlation between 1-month lead forecast Niño-3.4 anomalies and North American precipitation rate (Figs. 2b–i), we similarly find that overall the models reproduce the observed pattern (Fig. 2a) relatively well, although there are exceptions. The magnitude of the model–forecast relationship tends to be smaller than the magnitude of the observed relationship for some of the models (e.g., CMC1, NASA, NCAR-CCSM4), but most models capture the general spatial pattern. Because of this, we do not expect bridging with the Niño-3.4 index to enhance seasonal precipitation forecast skill. We intend to examine additional bridging predictors beyond ENSO in future work.

4. CBaM forecast skill: 2-m temperature

We first examine 2-m temperature forecast skill by calculating the BSS associated with calibrated and bridged 1-month lead seasonal NMME forecasts for each of the 12 overlapping 3-month seasons. NMME bridged (calibrated) forecasts are obtained by merging the bridged (calibrated) forecasts from the 7 NMME member models. We only show results for probabilistic forecasts of below normal temperature for brevity, although we note that the results for probabilistic forecasts of above normal temperature are very similar. We compare bridged, calibrated, and merged forecasts to raw NMME forecasts, where the raw forecasts are calculated as ensemble frequencies relative to model mean terciles. Overall, 1-month lead calibrated forecasts of temperature (Fig. 3) outperform bridged forecasts

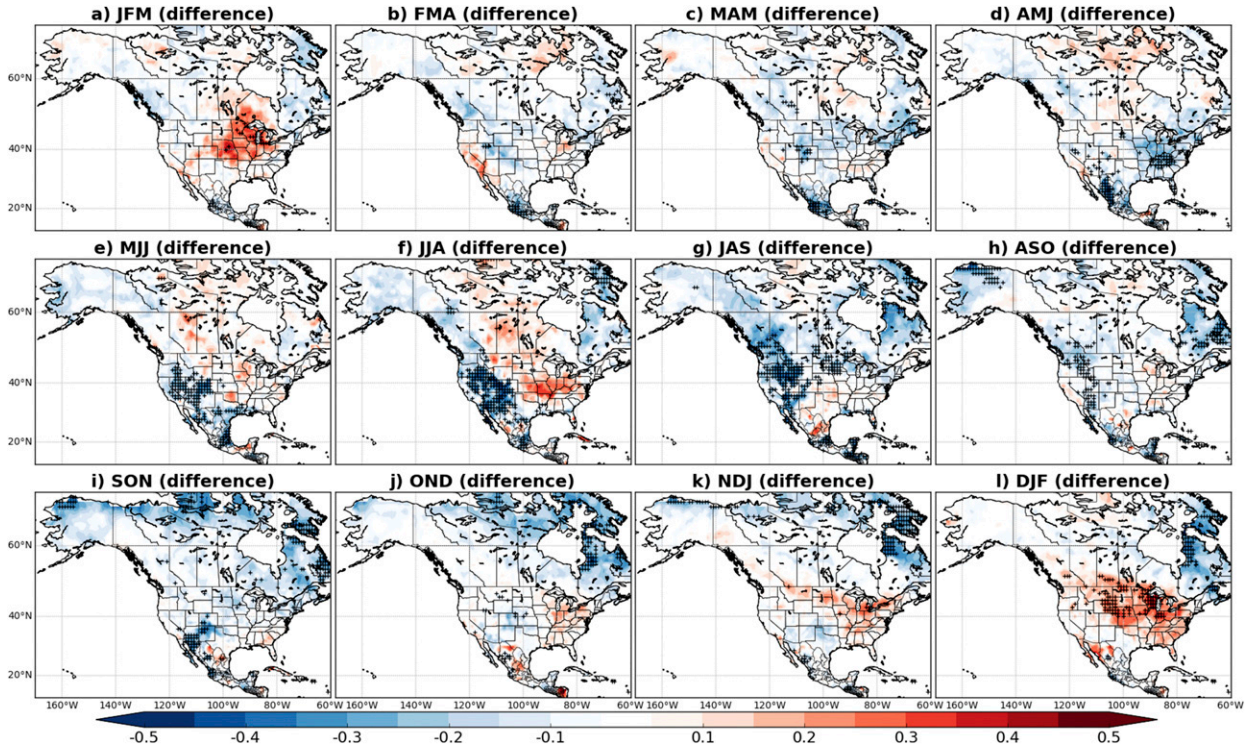

FIG. 4. Shading indicates Brier skill score differences between 1-month lead bridged and 1-month lead raw forecasts of below-normal 2-m temperature for the NMME for each of the 12 overlapping 3-month seasons. Red shading indicates that bridged forecast mean Brier skill scores exceeded raw forecast mean Brier skill scores over the 1982–2010 hindcast period. Hatching indicates significance at the 95% confidence level, as determined via a Wilcoxon rank sum test without accounting for field significance.

(Fig. 4), where bridging is done using model forecasts of the Niño-3.4 index. Calibration yields the largest improvement over raw forecasts in the fall and winter, as indicated by the red shading in Fig. 3. It should be noted that although the hatching in Figs. 3–5 indicates significance at the 95% confidence level, as determined by a Wilcoxon rank-sum test, the significance does not generally hold up when stricter field significance tests are applied. We test field significance using both the Walker coefficient and the false discovery rate (Wilks 2006) and find that very few grids meet this stricter standard. Therefore, these differences should be interpreted cautiously.

In contrast to calibrated forecasts, bridged forecasts more often yield lower mean Brier skill scores than raw forecasts, particularly in the spring and summer months (Fig. 4). However, bridged winter temperature forecast skill exceeds raw forecast skill across portions of the northern United States and southern Canada, including some areas where calibration does not result in improved skill. This is not particularly surprising when we consider that the ENSO influence on North American climate tends to be greatest during the winter months (Ropelewski and Halpert 1986). Note that when we instead develop the NMME bridged (calibrated) forecasts from the multimodel mean forecast Niño-3.4 index (2-m temperature), the results are statistically indistinguishable from the BSS maps presented in Figs. 3 and 4, which were obtained by merging all calibrated (Fig. 3) or bridged (Fig. 4) member model forecasts.

Figure 5 suggests that merging the bridged and calibrated forecasts results in marginally higher spatial coverage of positive skill, which agrees well with the findings of Schepen et al. (2016). Again, merging is done by taking a weighted (BMA) average of all of the bridged and calibrated forecasts from the NMME member models. Therefore, the merged forecasts used to create Fig. 5 result from merging a total of 14 forecasts—1 calibrated and 1 bridged forecast from each of the 7 models. Merged forecasts generally outperform raw NMME probabilistic forecasts of 2-m temperature, although there are some exceptions [e.g., parts of the northern United States during July–September (JAS)] for which the CBaM method does not outperform the raw forecasts. Some of the improvement occurs over regions for which the raw forecast skill is negative (e.g., over the eastern United States during DJF). This occurs because the CBaM method yields a climatology forecast when no evidence of positive forecast skill exists.

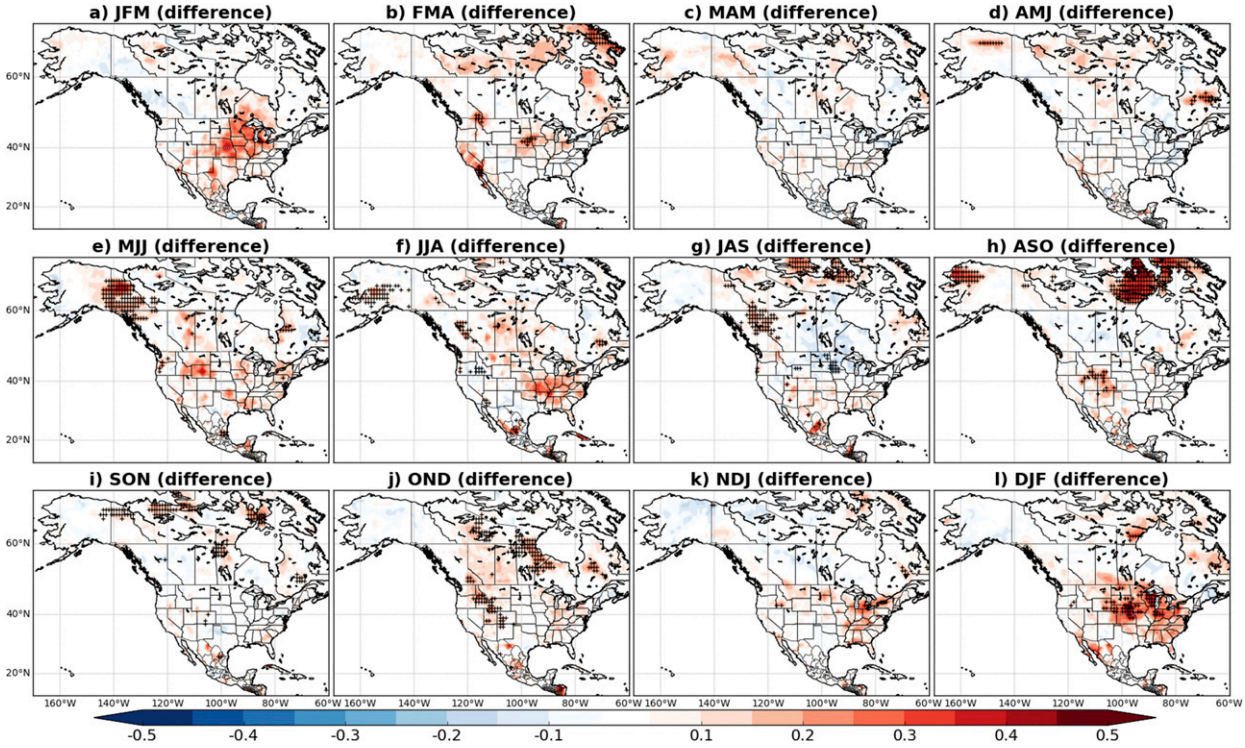

FIG. 5. Shading indicates Brier skill score differences between 1-month lead merged and 1-month lead raw forecasts of below-normal 2-m temperature for the NMME for each of the 12 overlapping 3-month seasons. Red shading indicates that merged forecast mean Brier skill scores exceeded raw forecast mean Brier skill scores over the 1982–2010 hindcast period. Hatching indicates significance at the 95% confidence level, as determined via a Wilcoxon rank sum test without accounting for field significance.

Because DJF appears to be one of the few seasons for which bridging enhances forecast skill, we focus on this season and investigate the merged results by NMME member model (Fig. 6). We apply a bootstrap method similar to that applied in Schepen et al. (2016) to assess whether bridging statistically significantly improves forecast skill beyond what is achieved through calibration. We first use resampling to generate a large

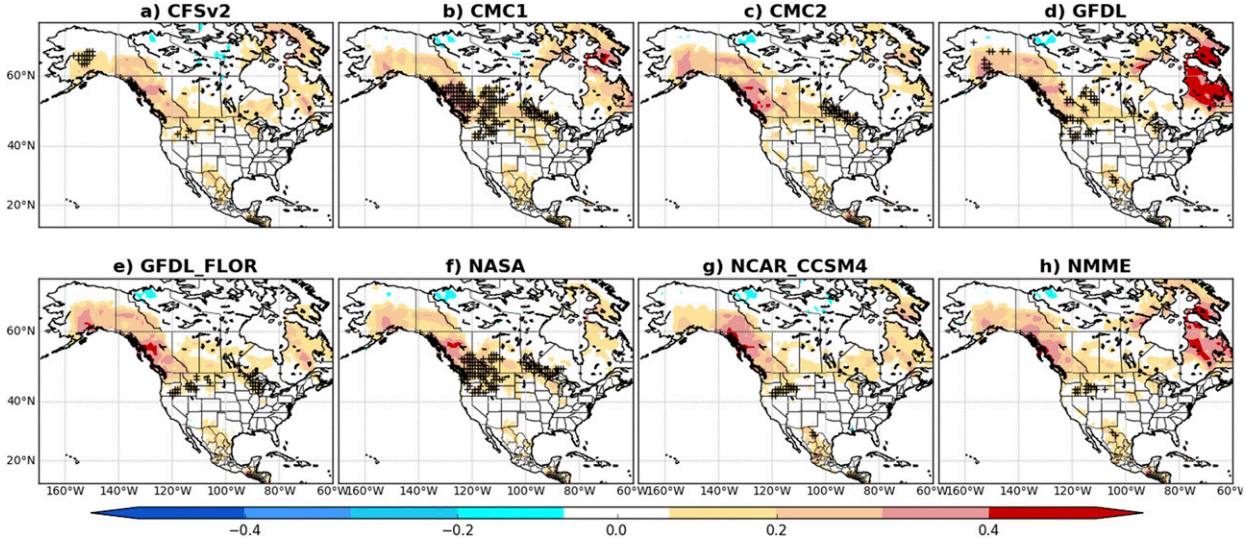

FIG. 6. Shading indicates Brier skill scores associated with 1-month lead merged forecasts of below-normal DJF 2-m temperature for each of the NMME member models and the multimodel mean. Hatching denotes grid cells for which bridging statistically significantly improves forecast skill, where statistical significance is determined using a bootstrap method without accounting for field significance.

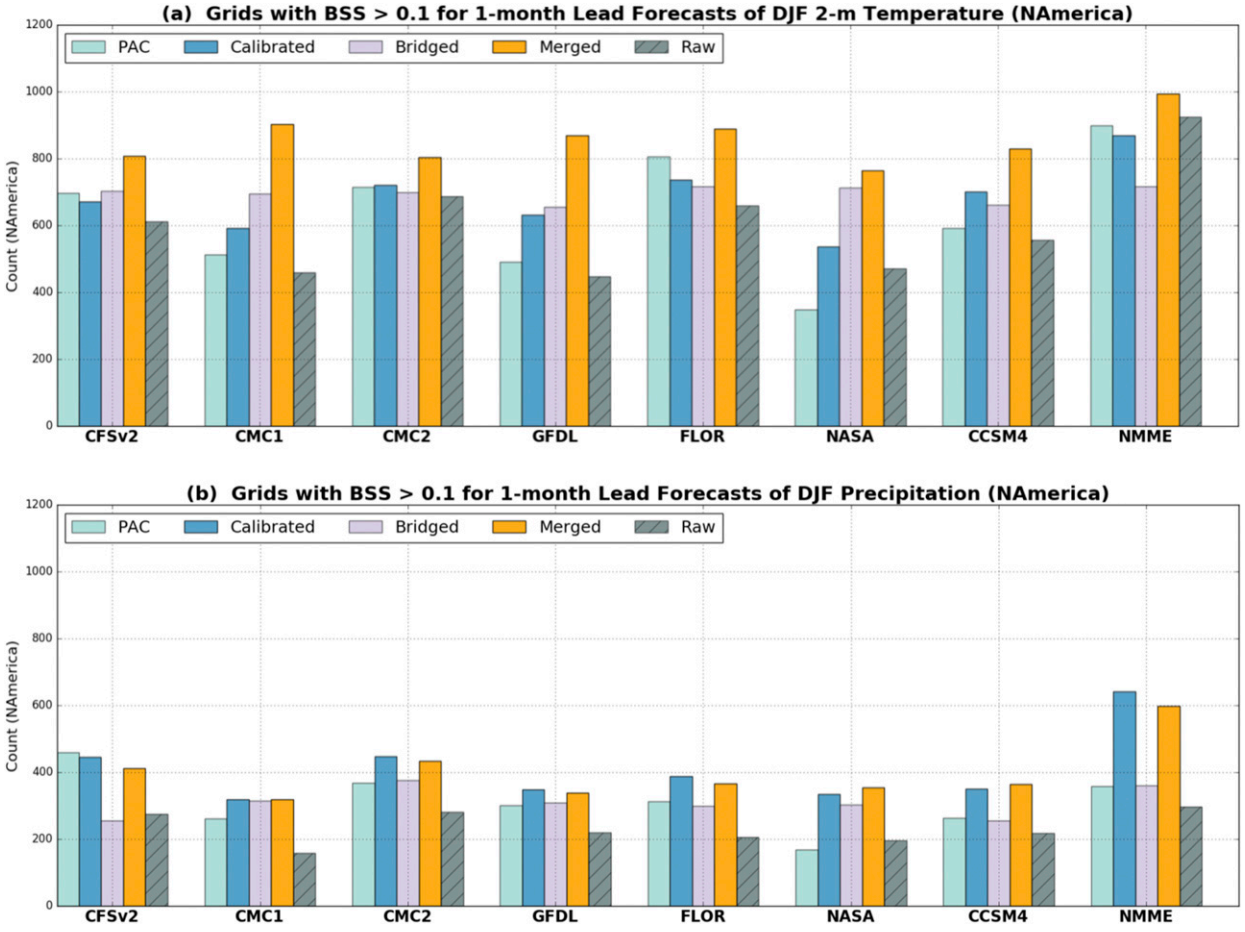

FIG. 7. The number of grid cells over North America with BSS > 0.1 for PAC-calibrated (light blue), BJP-calibrated (dark blue), bridged (purple), merged (orange), and raw (gray) 1-month lead forecasts of below-normal DJF (a) temperature and (b) precipitation from each of the NMME member models and the multimodel mean.

($n = 1000$) sample of calibrated BSS estimates for each grid point. If the merged BSS value exceeds the 95th percentile value from the resampled calibrated BSS distribution, we conclude that bridging enhances forecast skill at that grid point. As the hatching indicates in Fig. 6, the degree to which bridging helps varies by model, although we again note that significance does not hold up to the stricter field significance standard. Bridging does little to improve forecast skill for the CFSv2 and NCAR-CCSM4. Recall that both of these models already represent the ENSO–temperature teleconnection pattern relatively well (Fig. 1). Given this, it seems reasonable to expect that bridging would do little to improve forecast skill for these models. In contrast, bridging improves skill over portions of the northern United States and southwestern Canada for the remaining five models. The area of improved skill coincides with the region for which these models fail to reproduce the observed ENSO–temperature teleconnection. Throughout the year, bridging statistically significantly enhances forecast skill for less than 1%–6.4% of grid cells over North America, depending on the model and season. Interestingly, we find that bridging improves multimodel mean forecast skill for less than 1% of grid cells, which is also true for the NCAR-CCSM4 and CFSv2 models. We similarly apply the bootstrap method described above to determine whether bridging significantly *reduces* forecast skill in the final merged product. For all models and seasons, we find that bridging degrades forecast skill for <1% of grid cells (not shown).

These results support the notion that multimodel ensembles on average yield more skillful forecasts than individual model ensembles (e.g., Palmer et al. 2004; Kirtman et al. 2014). When we compare the number of grid cells with BSS > 0.1 (Fig. 7a), we find that 1) merged forecasts produce the most grid cells with BSS > 0.1

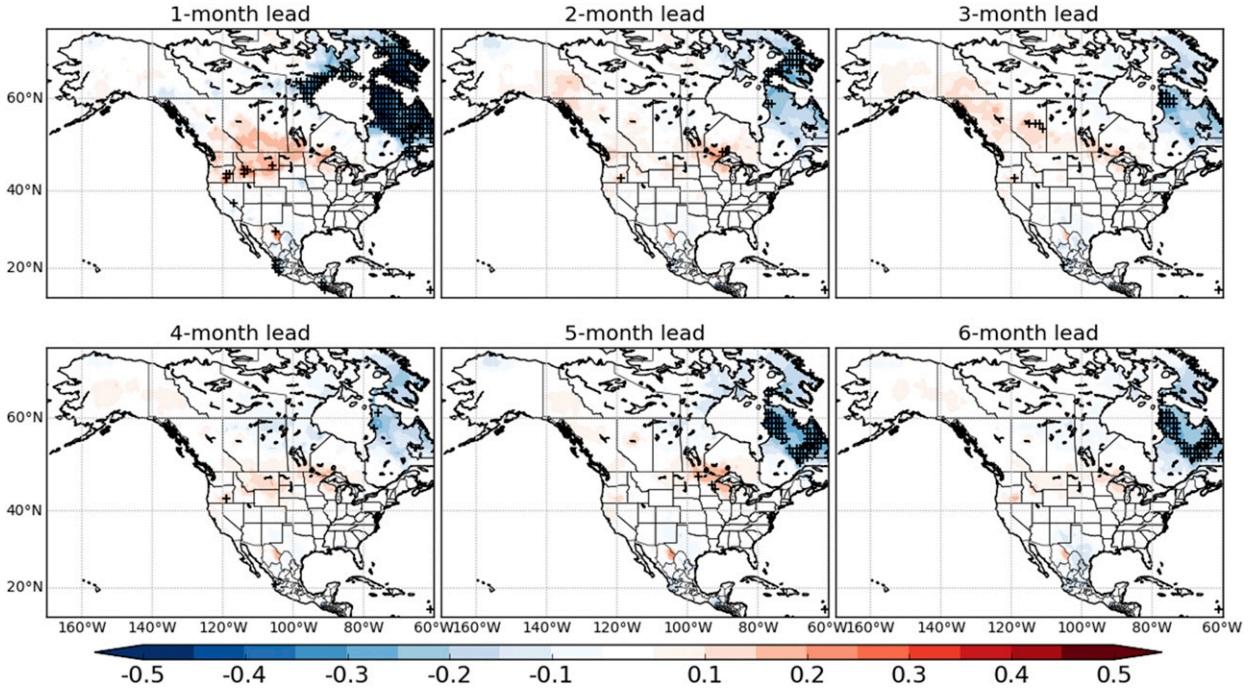

FIG. 8. Shading indicates Brier skill score differences between bridged and calibrated forecasts of below-normal DJF 2-m temperature for the NMME for 1–6-month lead forecasts. Red shading indicates that bridged forecast mean Brier skill scores exceeded calibrated forecast mean Brier skill scores over the 1982–2010 hindcast period. Hatching indicates significance at the 95% confidence level, as determined via a Wilcoxon rank sum test without accounting for field significance.

compared to calibration and bridging for each of the NMME member models, and 2) merged *NMME* forecasts produce the most grid cells with BSS > 0.1 compared to any individual member model. As expected, merged forecasts from the individual models tend to be less skillful than merged NMME forecasts, just as raw forecasts from the individual models are, on average, less skillful than raw NMME forecasts. However, the amount of improvement achieved by taking a multimodel mean tends to be smaller for the merged forecasts compared to the raw forecasts. For example, if we compare the number of merged forecast grids with BSS > 0.1 (the orange bars in Fig. 7a), we find that the merged NMME forecast improves upon the merged individual model forecasts by an average of 15.7%, with a range of 9.1% to 23.1%. In contrast, the raw NMME forecast improves upon the raw individual model forecasts by an average of 39.8%, with a range of 25.7%–51.6%.

Figure 7a also includes a comparison to the probability anomaly correlation (PAC) calibration method, which is based on an ordinary regression of predicted versus observed probabilities (van den Dool et al. 2017). The PAC method is applied at CPC to calibrate the real-time NMME forecasts produced on a monthly basis and used by CPC operational forecasters as a guidance tool. Overall, the PAC method for calibrating forecasts performs comparably to BJP calibration. We note that the NASA model used here for PAC calibration represents a different version than that used in the BJP analysis and is therefore not directly comparable. As with BJP calibration, PAC calibration does not yield as many grids with BSS > 0.1 as we find with the fully merged NMME forecast. This result supports the hypothesis that the statistical–dynamical bridging component to CBaM contributes skill and is a potentially useful addition to the current postprocessing being applied operationally at CPC. While the fully merged CBaM forecast achieves broader spatial coverage of positive skill, PAC calibration, BJP calibration, and merging result in similar mean Brier skill scores of 9.03%, 8.25%, and 9.06%, respectively, for the NMME. These mean BSS values all exceeded the 4.42% raw NMME mean BSS. In general, the improvement we see from application of CBaM occurs through increased spatial coverage of positive skill. Skill associated with fully merged CBaM forecasts may be slightly lower than the skill of the ''best forecast'' at any given grid point.

While we focus on 1-month lead forecasts, we note that the difference between bridged and calibrated DJF temperature forecasts tends to decrease as lead time

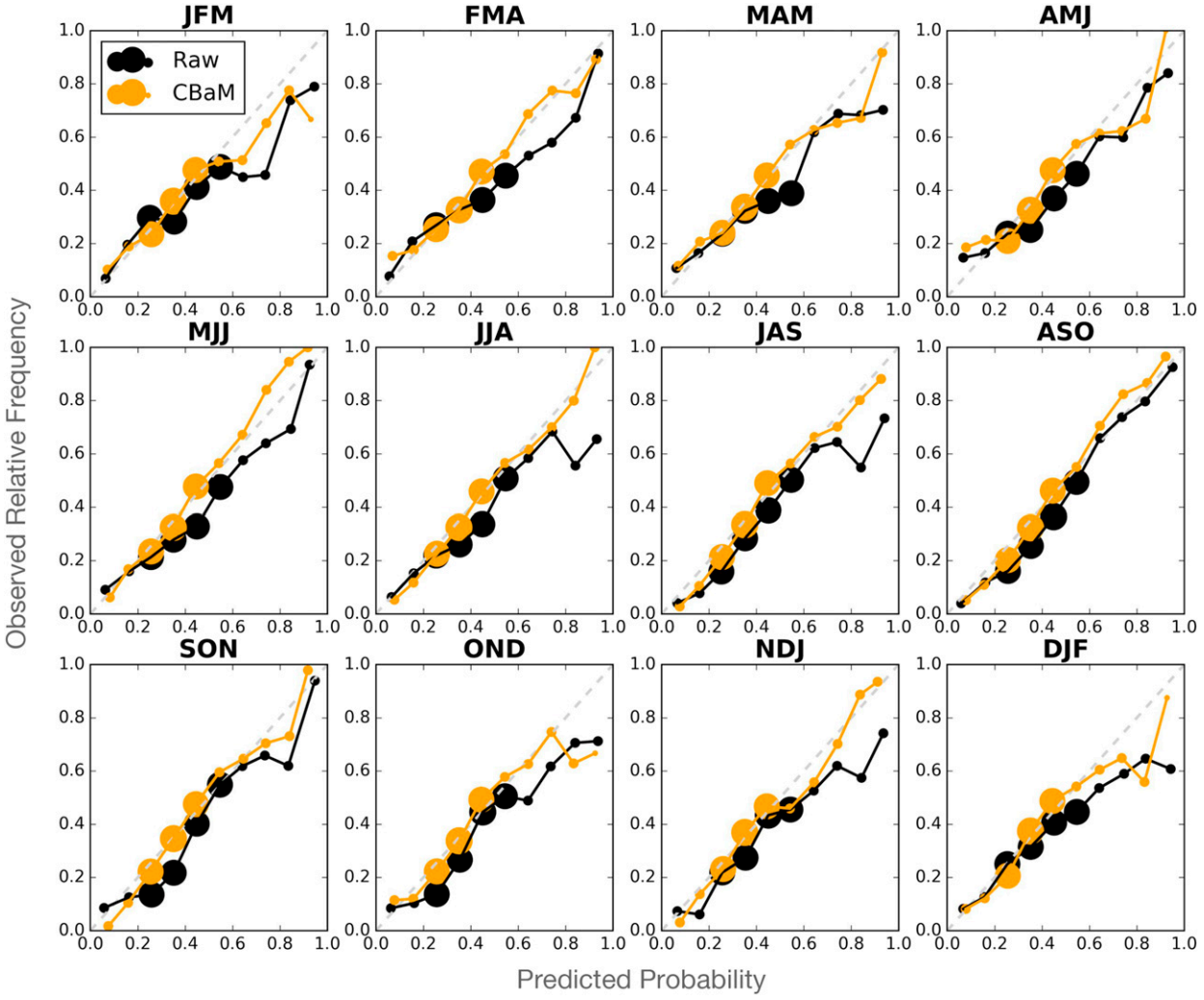

FIG. 9. Reliability diagrams comparing the statistical reliability of 1-month lead raw (black) vs CBaM (orange) forecasts of 2-m temperature. The CBaM results refer to the fully merged multimodel (NMME) forecast. The raw forecast refers to the multimodel mean forecast, without any bias correction applied. The horizontal axis denotes the predicted probability while the vertical axis denotes the observed relative frequency. The light gray dashed line corresponds to a perfectly reliable forecast. The size of the plotted circles is proportional to the number of forecast probabilities that fall into a given predicted probability bin.

increases (Fig. 8). As lead time increases and forecast skill decreases, BJP forecasts tend to be closer to climatology forecasts for more grids. Although both calibrated and bridged forecast skill decrease as lead time increases, mean Brier skill scores remain at or above climatology skill (BSS remains greater than or equal to zero) through 6-month lead, the longest lead time we are able to assess using NMME data.

Importantly, we find that the CBaM method produces statistically reliable forecasts of 2-m temperature (Fig. 9). To make the reliability diagrams, we bin the probabilistic NMME CBaM forecasts of below-normal temperature into 10 ''predicted probability'' bins (0–0.1, 0.1–0.2, … , 0.9–1.0), which are shown on the x axis. The y axis corresponds to the observed relative frequency of an event, where the event of interest is the occurrence of below-normal seasonal temperatures. In a well-calibrated forecast system the predicted probability matches the observed relative frequency. For example, a 40% forecast probability of below-normal temperature should verify as below normal 40% of the time. We find that the fully merged CBaM forecasts on average are more reliable than raw NMME forecasts for all 12 of the 3-month overlapping seasons. The improvement is particularly evident for higher predicted probabilities, although we note that the sample sizes are much smaller for predicted probabilities exceeding 60%.

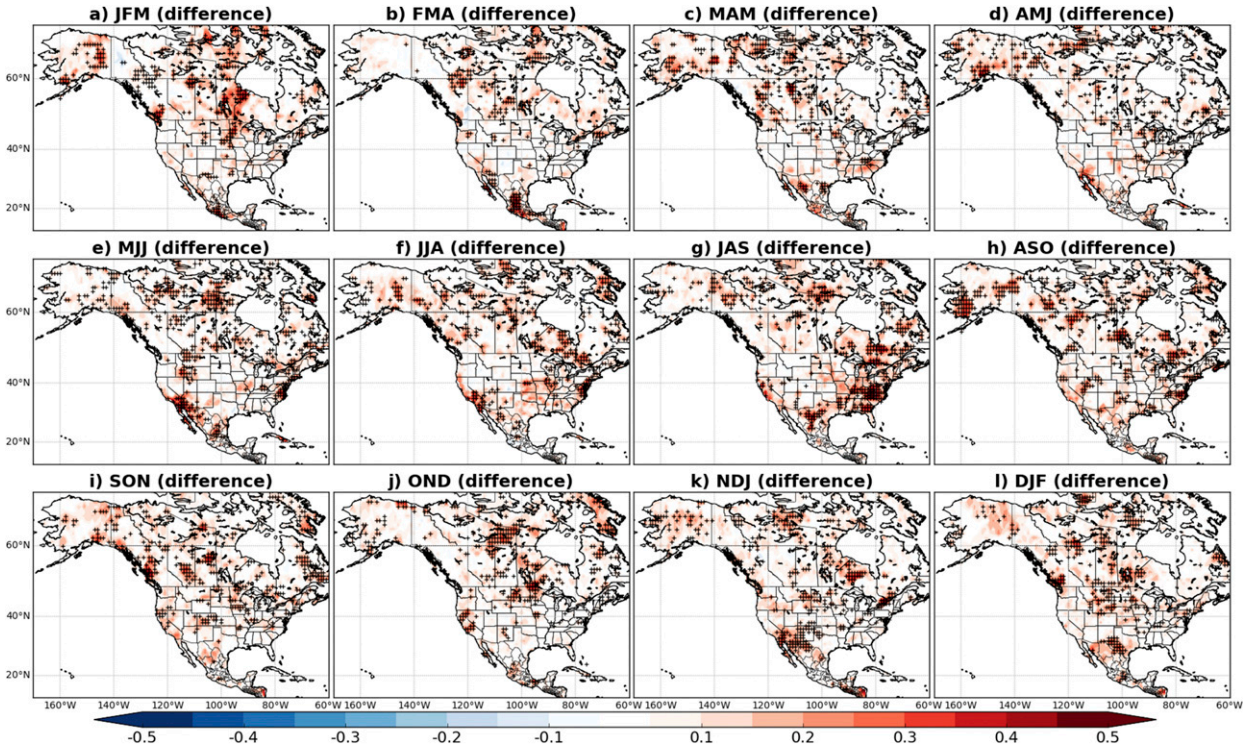

FIG. 10. Shading indicates Brier skill score differences between 1-month lead calibrated and 1-month lead raw forecasts of below-normal precipitation rate for the NMME for each of the 12 overlapping 3-month seasons. Red shading indicates that calibrated forecast mean Brier skill scores exceeded raw forecast mean Brier skill scores over the 1982–2010 hindcast period. Hatching indicates significance at the 95% confidence level, as determined via a Wilcoxon rank sum test without accounting for field significance.

5. CBaM forecast skill: Precipitation rate

When we repeat the above analysis using 1-month lead forecasts of precipitation, we find that bridging and calibration tend to improve upon raw Brier skill scores in the same areas (Figs. 10 and 11). In contrast to 2-m temperature forecasts, calibrated precipitation rate forecasts tend to achieve higher skill during the winter and very little skill throughout the remainder of the year. A qualitative assessment of Figs. 10 and 11 suggests significant overlap between grid cells with improved skill from calibration and grid cells with improved skill from bridging. This is also supported by Fig. 7b where we see that for some models (e.g., CMC1), approximately 300 grid cells achieve a BSS > 0.1 for calibrated, bridged, and merged forecasts. In general, if calibration and bridging yielded skillful forecasts over different areas, we would expect the number of grid cells with positive skill to be highest for the merged forecasts. This result supports the idea that we can attribute much of the wintertime precipitation forecast skill to ENSO. In contrast to temperature, we find that taking a multimodel mean results in large improvements in skill relative to individual models for both merged forecasts (comparing the orange bars in Fig. 7b) and raw forecasts (comparing the gray bars in Fig. 7b). The merged NMME forecast improves upon merged individual model forecasts by an average of 38.3%, and the raw NMME forecast improves upon raw individual model forecasts by an average of 25.2%. Calibrated and merged forecasts in particular result in higher coverage of positive skill compared to raw NMME forecasts, with calibration marginally outperforming merging.

Figure 7b also includes a comparison with PAC-calibrated precipitation forecasts. PAC calibration tends to yield fewer grid cells with BSS > 0.1 than we find with CBaM-calibrated or merged forecasts, with the exception of the CFSv2. It is not clear what contributes to this difference in performance. One possible explanation is that BJP applies a more robust data transformation method and includes censoring as a means of handling grid points with zero values. Of course, other differences in the method may contribute to the difference in skill. The CBaM probabilities are calculated from an ensemble of 1000 members, while the probabilities used in the PAC analysis are calculated from model ensemble frequencies. Additionally, PAC probabilities are damped to climatology in areas where

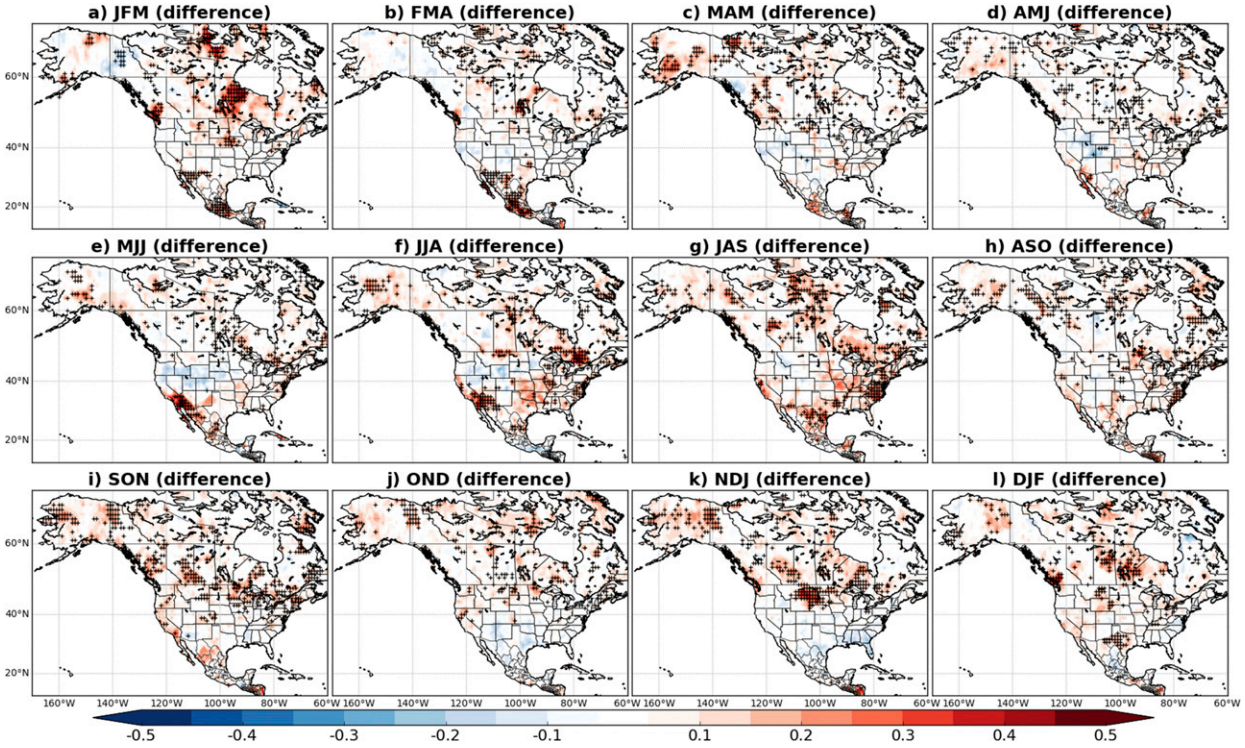

FIG. 11. Shading indicates Brier skill score differences between 1-month lead bridged and 1-month lead raw forecasts of below-normal precipitation rate for the NMME for each of the 12 overlapping 3-month seasons. Red shading indicates that calibrated forecast mean Brier skill scores exceeded raw forecast mean Brier skill scores over the 1982–2010 hindcast period. Hatching indicates significance at the 95% confidence level, as determined via a Wilcoxon rank sum test without accounting for field significance.

historical raw forecast skill is near zero or significantly negative. BJP probabilities are damped to climatology in areas where raw forecast skill is near zero, but not in areas where raw forecast skill is significantly negative. The significant negative correlation informs the BJP calibration, yielding a ''flipped'' calibrated forecast relative to the raw forecast. There is some debate regarding whether this method makes sense in an operational setting given that the resulting probability forecasts may be less physically justifiable in cases when raw forecast skill is significantly negative. As with temperature, we find that PAC calibration, BJP calibration, and merging (Fig. 12) all result in higher mean BSS relative to the raw forecast (PAC = 4.9%, BJP-cal = 5.25%, merged = 4.75%, raw = −2.56%).

The CBaM precipitation results also support our hypothesis that because the ENSO–precipitation teleconnection pattern is well represented by the NMME member models, bridging with the forecast Niño-3.4 index will not contribute significant additional skill. In fact, when we apply the bootstrap significance test, we find that bridging enhances skill for less than 1% of grid cells for most model-season combinations, with the most improvement occurring for CMC1 forecasts of JFM precipitation (3.5%). It seems reasonable to expect some improvement with bridging for the CMC1 given that, as is evident in Fig. 2, the CMC1 representation of the ENSO–precipitation teleconnection pattern least resembles the observed pattern when compared against the other NMME member models.

Although bridging does not significantly increase precipitation forecast skill beyond what is achieved by calibration, application of the CBaM method does improve forecast reliability (Fig. 13). Even so, precipitation forecasts remain less reliable than temperature forecasts overall, particularly at the higher probability bins where the sample size is smaller. Although precipitation forecast skill remains modest, the results presented in Figs. 9 and 13 demonstrate the utility of the CBaM method to improve the representation of forecast uncertainty relative to raw NMME forecasts.

Finally, we note that because the analysis is applied at each grid point, CBaM forecasts, and precipitation forecasts in particular, tend to be noisy. We do not apply any smoothing to the final CBaM forecasts, although future work will likely include some type of smoothing. Precipitation forecasts are particularly noisy in the summertime when the ENSO signal is weaker over

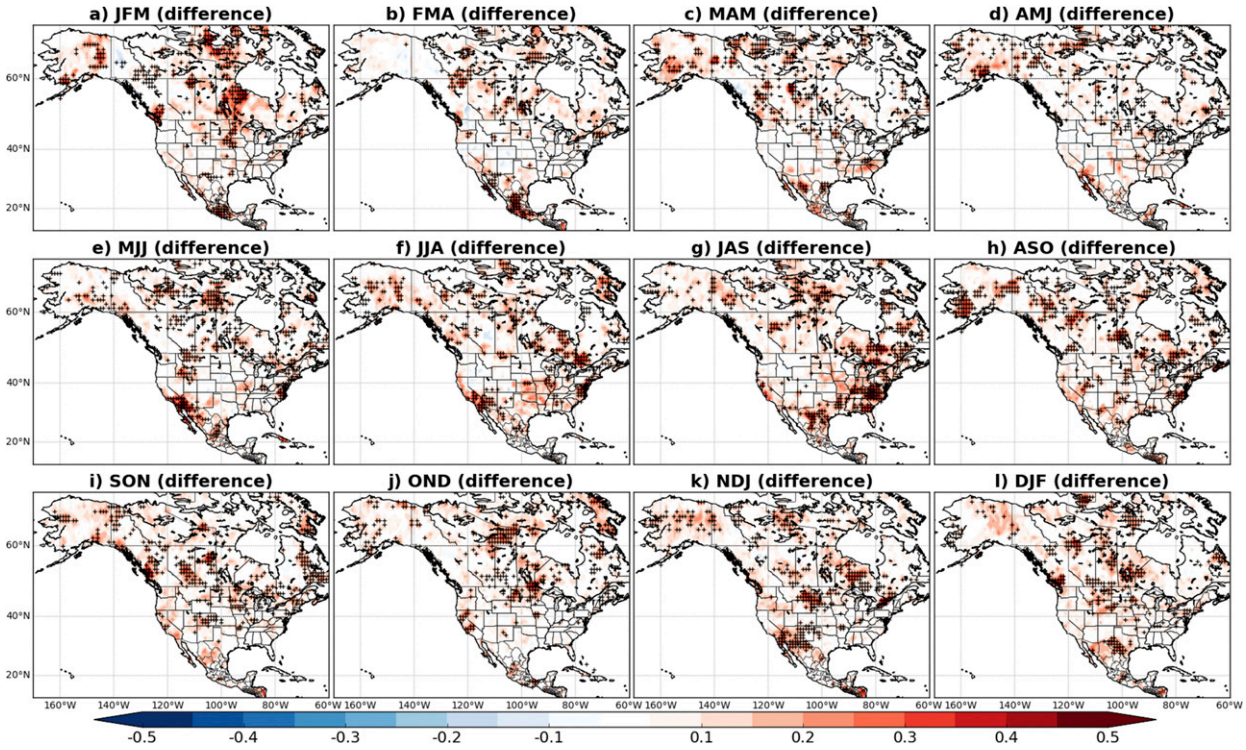

FIG. 12. Shading indicates Brier skill score differences between 1-month lead merged and 1-month lead raw forecasts of below-normal precipitation rate for the NMME for each of the 12 overlapping 3-month seasons. Red shading indicates that calibrated forecast mean Brier skill scores exceeded raw forecast mean Brier skill scores over the 1982–2010 hindcast period. Hatching indicates significance at the 95% confidence level, as determined via a Wilcoxon rank sum test without accounting for field significance.

North America and when much of the precipitation occurring over a large portion of the United States is convective in nature.

6. Discussion

These results suggest that, particularly for individual model ensembles, the CBaM method improves forecast skill and statistical reliability over North America, both through calibration to correct for model biases in regions and seasons with underlying model skill, and through bridging to correct for model misrepresentation of teleconnection patterns. In particular, bridging using the forecast Niño-3.4 index statistically significantly enhances 2-m temperature forecast skill for several of the individual NMME member models, primarily over regions where the model and observed ENSO teleconnection patterns differ. Improvements through bridging largely are confined to the winter season (DJF). While bridging enhances forecast skill for the individual models that make up the NMME, when we calculate the multimodel mean results, we find that bridging improves skill for less than 1% of grid cells. Similarly, for models that better represent the ENSO–temperature teleconnection (e.g., CFSv2, NCAR-CCSM4), bridging significantly improves forecast skill for less than 1% of grid cells over North America. This suggests that the multimodel mean at least partially improves the representation of the teleconnection. Additionally, because the NMME-calibrated forecast was obtained by applying BMA to merge the individual member model-calibrated forecasts, greater weight was given to models that performed better (e.g., CFSv2, NCAR-CCSM4), resulting in a better calibrated forecast and less need for improvement via bridging.

In contrast to temperature, bridging does not improve skill for 1-month lead forecasts of precipitation. This result supports our initial hypothesis that bridging should lead to fewer improvements in precipitation forecast skill because the ENSO–precipitation teleconnection is better represented by the NMME member models. Future work will explore other possible methods for improving precipitation forecast skill, for example, by using calibrated NMME temperature as a predictor of precipitation (e.g., Narapusetty et al. 2018).

Finally, we find that application of CBaM narrows the average gap between the skill of an individual model

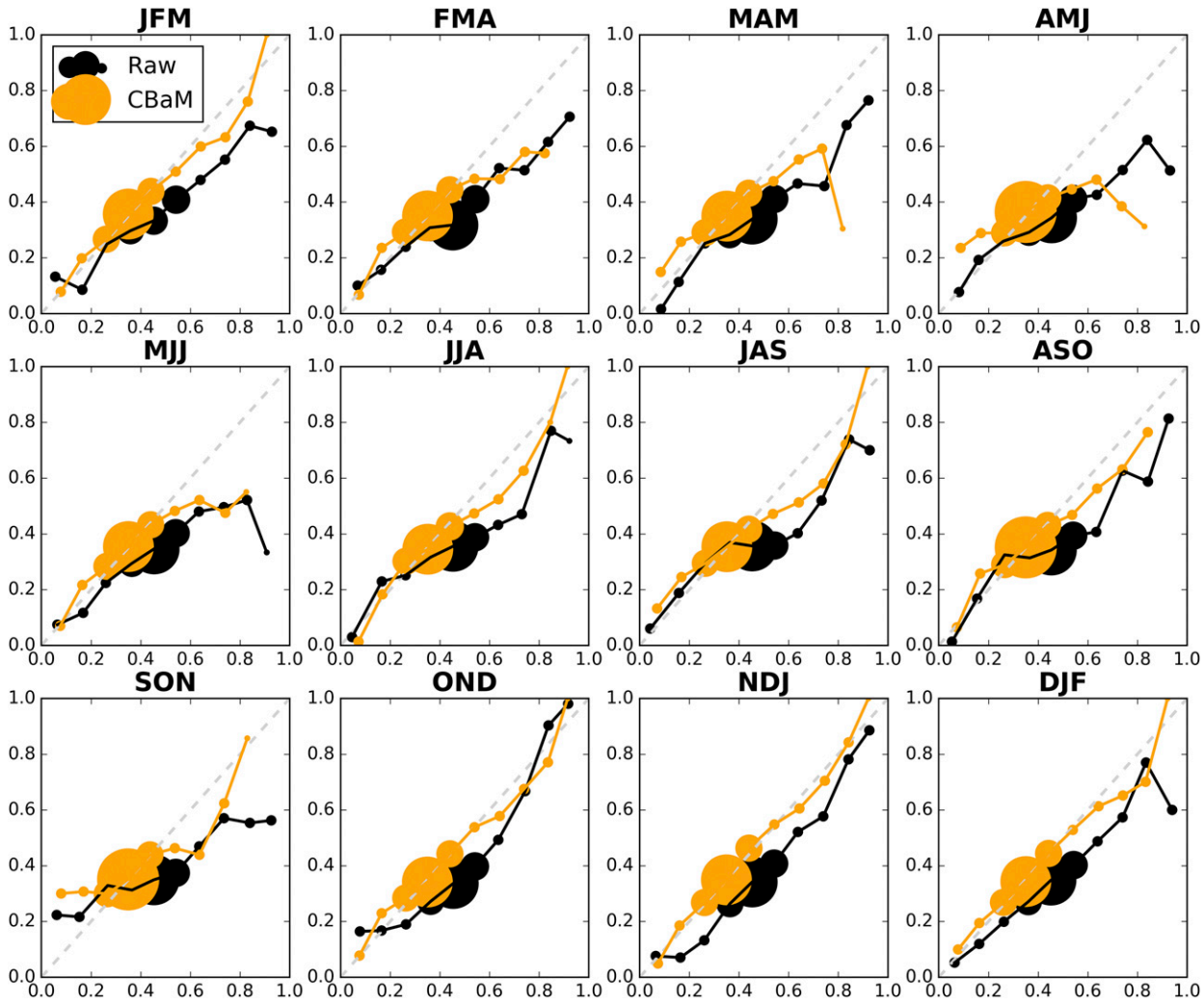

FIG. 13. Reliability diagrams comparing the statistical reliability of 1-month lead raw (black) vs CBaM (orange) forecasts of precipitation rate. The CBaM results refer to the fully merged multimodel (NMME) forecast. The raw forecast refers to the multimodel mean forecast, without any bias correction applied. The horizontal axis denotes the predicted probability while the vertical axis denotes the observed relative frequency. The light gray dashed line corresponds to a perfectly reliable forecast. The size of the plotted circles is proportional to the number of forecast probabilities that fall into a given predicted probability bin.

temperature forecast and the skill of an NMME temperature forecast. Merged temperature forecasts by individual models are on average 15% less skillful than merged NMME temperature forecasts, whereas raw temperature forecasts by individual models are on average 40% less skillful than raw NMME temperature forecasts. It is not immediately clear why this occurs or whether these results would hold true if the method were to be applied to a different multimodel ensemble. Future work will test the method with other multimodel ensembles and may yield additional insight into this result. In contrast to temperature, CBaM does not narrow this skill gap for forecasts of precipitation. NMME precipitation forecasts tend to be substantially better than individual model precipitation forecasts regardless of whether CBaM is applied.

CBaM postprocessing is currently being applied to real-time NMME forecasts. These postprocessed real-time forecasts serve as an experimental tool to aid in the monthly production of operational seasonal forecasts by CPC. The real-time CBaM forecasts, available since October 2018, can be accessed at http://www.cpc.ncep.noaa.gov/products/people/sstrazzo/cbam/. For comparison, PAC-calibrated NMME forecasts can be found at http://www.cpc.ncep.noaa.gov/products/NMME/prob/pac/.

There are several caveats to this study worth mentioning. Owing to the short hindcast period, uncertainty

in observed ENSO teleconnection patterns makes it difficult to reasonably evaluate model teleconnections, although a new approach for doing so was recently introduced (Deser et al. 2017). This problem is further compounded by the variability observed among historical ENSO events (i.e., no two ENSO events are exactly alike). Therefore, the results presented here are heavily influenced by a handful of extreme ENSO events. The short sampling period can also prove problematic when attempting to estimate the BMA weights for model merging. In some cases, the model that performs best is given a very large percentage of the weight, while the remaining models are given weights near zero. Although we find that the BMA weighting method outperforms equal weighting for the hindcast period, differences in hindcast versus real-time individual model configurations (e.g., number of ensemble members) may render the BMA weights less useful in a real-time forecasting setting. We are currently testing alternative weighting approaches to address this issue.

Additionally, we limit bridging to ENSO and therefore do not consider other critical sources of North American climate variability that models may be misrepresenting (e.g., the Arctic Oscillation). Future work also will examine more bridging predictors on a global scale. Finally, CBaM is one among many methods used to postprocess individual and now multimodel ensemble systems. Although here we include some discussion comparing CBaM to methods currently in use operationally (e.g., PAC, ensemble regression), future work should include a more thorough comparison of the full set of methods available to seasonal forecast practitioners.

Acknowledgments. This research was funded by NOAA-OAR-CPO Grant GC16-307. We thank two NOAA internal reviewers for their feedback. We also thank two anonymous reviewers for their constructive suggestions.

REFERENCES

Banzon, V., T. M. Smith, T. M. Chin, C. Liu, and W. Hankins, 2016: A long-term record of blended satellite and in situ sea-surface temperature for climate monitoring, modeling and environmental studies. *Earth Syst. Sci. Data*, **8**, 165–176, https://doi.org/10.5194/essd-8-165-2016.

Barnston, A. G., M. K. Tippett, M. Ranganathan, and M. L. L'Heureux, 2018: Deterministic skill of ENSO predictions from the North American Multimodel Ensemble. *Climate Dyn.*, https://doi.org/10.1007/s00382-017-3603-3, in press.

Becker, E., and H. van Den Dool, 2016: Probabilistic seasonal forecasts in the North American multimodel ensemble: A baseline skill assessment. *J. Climate*, **29**, 3015–3026, https://doi.org/10.1175/JCLI-D-14-00862.1.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Challinor, A., J. Slingo, T. Wheeler, and F. Doblas-Reyes, 2005: Probabilistic simulations of crop yield over western India using the DEMETER seasonal hindcast ensembles. *Tellus*, **57A**, 498–512, https://doi.org/10.3402/tellusa.v57i3.14670.

Chen, L.-C., H. van den Dool, E. Becker, and Q. Zhang, 2017: ENSO precipitation and temperature forecasts in the North American Multimodel Ensemble: Composite analysis and validation. *J. Climate*, **30**, 1103–1125, https://doi.org/10.1175/JCLI-D-15-0903.1.

Deser, C., I. R. Simpson, K. A. McKinnon, and A. S. Phillips, 2017: The Northern Hemisphere extratropical atmospheric circulation response to ENSO: How well do we know it and how do we evaluate models accordingly? *J. Climate*, **30**, 5059–5082, https://doi.org/10.1175/JCLI-D-16-0844.1.

Doblas-Reyes, F. J., J. García-Serrano, F. Lienert, A. P. Biescas, and L. R. Rodrigues, 2013: Seasonal climate predictability and forecasting: status and prospects. *Wiley Interdiscip. Rev.: Climate Change*, **4**, 245–268, https://doi.org/10.1002/wcc.217.

Fan, Y., and H. van den Dool, 2008: A global monthly land surface air temperature analysis for 1948–present. *J. Geophys. Res.*, **113**, D01103, https://doi.org/10.1029/2007JD008470.

Gent, P. R., and Coauthors, 2011: The Community Climate System Model version 4. *J. Climate*, **24**, 4973–4991, https://doi.org/10.1175/2011JCLI4083.1.

Goddard, L., S. J. Mason, S. E. Zebiak, C. F. Ropelewski, R. Basher, and M. A. Cane, 2001: Current approaches to seasonal to interannual climate predictions. *Int. J. Climatol.*, **21**, 1111–1152, https://doi.org/10.1002/joc.636.

Hagedorn, R., F. J. Doblas-Reyes, and T. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. *Tellus*, **57A**, 219–233, https://doi.org/10.1111/j.1600-0870.2005.00103.x.

Hartmann, H. C., T. C. Pagano, S. Sorooshian, and R. Bales, 2002: Confidence builders: Evaluating seasonal climate forecasts from user perspectives. *Bull. Amer. Meteor. Soc.*, **83**, 683–698, https://doi.org/10.1175/1520-0477(2002)083<0683:CBESCF>2.3.CO;2.

Hawkins, E., T. M. Osborne, C. K. Ho, and A. J. Challinor, 2013: Calibration and bias correction of climate projections for crop modelling: An idealised case study over Europe. *Agric. For. Meteor.*, **170**, 19–31, https://doi.org/10.1016/j.agrformet.2012.04.007.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky, 1999: Bayesian model averaging: A tutorial. *Stat. Sci.*, **14**, 382–417, https://doi.org/10.1214/ss/1009212519.

Jia, L., and Coauthors, 2015: Improved seasonal prediction of temperature and precipitation over land in a high-resolution GFDL climate model. *J. Climate*, **28**, 2044–2062, https://doi.org/10.1175/JCLI-D-14-00112.1.

Kirtman, B. P., and Coauthors, 2014: The North American Multimodel Ensemble: Phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bull. Amer. Meteor. Soc.*, **95**, 585–601, https://doi.org/10.1175/BAMS-D-12-00050.1.

MacLachlan, C., and Coauthors, 2015: Global Seasonal forecast system version 5 (GloSea5): A high-resolution seasonal forecast system. *Quart. J. Roy. Meteor. Soc.*, **141**, 1072–1084, https://doi.org/10.1002/qj.2396.

Merryfield, W. J., and Coauthors, 2013: The Canadian seasonal to interannual prediction system. Part I: Models and

initialization. *Mon. Wea. Rev.*, **141**, 2910–2945, https://doi.org/10.1175/MWR-D-12-00216.1.

Narapusetty, B., D. Collins, R. Murtugudde, J. Gottschalck, and C. Peters-Lidard, 2018: Bias correction to improve the skill of summer precipitation forecasts over contiguous United States by the North American Multi-Model Ensemble system. *Atmos. Sci. Lett.*, **19**, e818, https://doi.org/10.1002/asl.818.

NOAA/NSF/NASA/DOE, 2014: The North American multimodel ensemble. NOAA/NSF/NASA/DOE, accessed 3 April 2017, http://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME/.

NOAA/OAR/ESRL/PSD, 2002: NOAA optimum interpolation (OI) sea surface temperature (SST) v2. NOAA/OAR/ESRL/PSD, accessed 12 December 2016, https://www.esrl.noaa.gov/psd/data/gridded/data.noaa.oisst.v2.html.

——, 2008: GHCN-CAMS Gridded 2m Temperature (Land). NOAA/OAR/ESRL/PSD, accessed 12 December 2016, https://www.esrl.noaa.gov/psd/data/gridded/data.ghcncams.html.

Palmer, T., and Coauthors, 2004: Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). *Bull. Amer. Meteor. Soc.*, **85**, 853–872, https://doi.org/10.1175/BAMS-85-6-853.

Peng, Z., Q. Wang, J. C. Bennett, A. Schepen, F. Pappenberger, P. Pokhrel, and Z. Wang, 2014: Statistical calibration and bridging of ECMWF System4 outputs for forecasting seasonal precipitation over China. *J. Geophys. Res. Atmos.*, **119**, 7116–7135, https://doi.org/10.1002/2013JD021162.

Raftery, A. E., D. Madigan, and J. A. Hoeting, 1997: Bayesian model averaging for linear regression models. *J. Amer. Stat. Assoc.*, **92**, 179–191, https://doi.org/10.1080/01621459.1997.10473615.

Ropelewski, C. F., and M. S. Halpert, 1986: North American precipitation and temperature patterns associated with the El Niño/Southern Oscillation (ENSO). *Mon. Wea. Rev.*, **114**, 2352–2362, https://doi.org/10.1175/1520-0493(1986)114<2352:NAPATP>2.0.CO;2.

Saha, S., and Coauthors, 2014: The NCEP Climate Forecast System version 2. *J. Climate*, **27**, 2185–2208, https://doi.org/10.1175/JCLI-D-12-00823.1.

Schepen, A., Q. Wang, and D. E. Robertson, 2014: Seasonal forecasts of Australian rainfall through calibration and bridging of coupled GCM outputs. *Mon. Wea. Rev.*, **142**, 1758–1770, https://doi.org/10.1175/MWR-D-13-00248.1.

——, ——, and Y. Everingham, 2016: Calibration, bridging, and merging to improve GCM seasonal temperature forecasts in Australia. *Mon. Wea. Rev.*, **144**, 2421–2441, https://doi.org/10.1175/MWR-D-15-0384.1.

Shukla, S., A. McNally, G. Husak, and C. Funk, 2014: A seasonal agricultural drought forecast system for food-insecure regions of East Africa. *Hydrol. Earth Syst. Sci.*, **18**, 3907–3921, https://doi.org/10.5194/hess-18-3907-2014.

Tompkins, A. M., and F. Di Giuseppe, 2015: Potential predictability of malaria in Africa using ECMWF monthly and seasonal climate forecasts. *J. Appl. Meteor. Climatol.*, **54**, 521–540, https://doi.org/10.1175/JAMC-D-14-0156.1.

Torralba, V., F. J. Doblas-Reyes, D. MacLeod, I. Christel, and M. Davis, 2017: Seasonal climate prediction: A new source of information for the management of wind energy resources. *J. Appl. Meteor. Climatol.*, **56**, 1231–1247, https://doi.org/10.1175/JAMC-D-16-0204.1.

Unger, D. A., H. van den Dool, E. O'Lenic, and D. Collins, 2009: Ensemble regression. *Mon. Wea. Rev.*, **137**, 2365–2379, https://doi.org/10.1175/2008MWR2605.1.

van den Dool, H., 2007: *Empirical Methods in Short-Term Climate Prediction.* Oxford University Press, 240 pp.

——, E. Becker, L.-C. Chen, and Q. Zhang, 2017: The probability anomaly correlation and calibration of probabilistic forecasts. *Wea. Forecasting*, **32**, 199–206, https://doi.org/10.1175/WAF-D-16-0115.1.

van Oldenborgh, G. J., M. A. Balmaseda, L. Ferranti, T. N. Stockdale, and D. L. Anderson, 2005: Evaluation of atmospheric fields from the ECMWF seasonal forecasts over a 15-year period. *J. Climate*, **18**, 3250–3269, https://doi.org/10.1175/JCLI3421.1.

Vecchi, G. A., and Coauthors, 2014: On the seasonal forecasting of regional tropical cyclone activity. *J. Climate*, **27**, 7994–8016, https://doi.org/10.1175/JCLI-D-14-00158.1.

Vernieres, G., M. M. Rienecker, R. Kovach, and C. L. Keppenne, 2012: The GEOS-iODAS: Description and evaluation. NASA/TM-2012-104606/Vol. 30, NASA, 73 pp.

Wang, Q., and D. Robertson, 2011: Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences. *Water Resour. Res.*, **47**, W02546, https://doi.org/10.1029/2010WR009333.

——, ——, and F. Chiew, 2009: A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites. *Water Resour. Res.*, **45**, W05407, https://doi.org/10.1029/2008WR007355.

——, A. Schepen, and D. E. Robertson, 2012a: Merging seasonal rainfall forecasts from multiple statistical models through Bayesian model averaging. *J. Climate*, **25**, 5524–5537, https://doi.org/10.1175/JCLI-D-11-00386.1.

——, D. L. Shrestha, D. Robertson, and P. Pokhrel, 2012b: A log-sinh transformation for data normalization and variance stabilization. *Water Resour. Res.*, **48**, W05514, https://doi.org/10.1029/2011WR010973.

Wilks, D. S., 2006: On "field significance" and the false discovery rate. *J. Appl. Meteor. Climatol.*, **45**, 1181–1189, https://doi.org/10.1175/JAM2404.1.

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

Xie, P., and P. A. Arkin, 1997: CPC merged analysis of precipitation (CMAP). NOAA/NWS/CPC, accessed 12 December 2016, http://www.cpc.ncep.noaa.gov/products/global_precip/html/wpage.cmap.html.

Xue, Y., M. Chen, A. Kumar, Z.-Z. Hu, and W. Wang, 2013: Prediction skill and bias of tropical Pacific sea surface temperatures in the NCEP Climate Forecast System version 2. *J. Climate*, **26**, 5358–5378, https://doi.org/10.1175/JCLI-D-12-00600.1.

Yeo, I.-K., and R. A. Johnson, 2000: A new family of power transformations to improve normality or symmetry. *Biometrika*, **87**, 954–959, https://doi.org/10.1093/biomet/87.4.954.

Zhang, S., M. Harrison, A. Rosati, and A. Wittenberg, 2007: System design and evaluation of coupled ensemble data assimilation for global oceanic climate studies. *Mon. Wea. Rev.*, **135**, 3541–3564, https://doi.org/10.1175/MWR3466.1.

Zhang, W., G. Villarini, L. Slater, G. A. Vecchi, and A. A. Bradley, 2017: Improved ENSO forecasting using Bayesian updating and the North American Multimodel Ensemble (NMME). *J. Climate*, **30**, 9007–9025, https://doi.org/10.1175/JCLI-D-17-0073.1.