Climate Dynamics (2019) 53:3373–3385
https://doi.org/10.1007/s00382-019-04710-7

# Improving the CPC's ENSO Forecasts using Bayesian model averaging

**Hanpei Zhang[1] · Pao-Shin Chu[1] · Luke He[2] · David Unger[2]**

Received: 3 November 2018 / Accepted: 1 March 2019 / Published online: 9 March 2019


**Abstract**
Statistical and dynamical model simulations have been commonly used separately in El Niño–Southern Oscillation (ENSO) prediction. Current models are imperfect representations of ENSO and each of them has strength and weakness for capturing different aspects in ENSO prediction. Thus, it is important to utilize the results from a variety of different models. The Bayesian model averaging (BMA) is an effective tool not only in describing uncertainties associated with each model simulation but also providing the forecast performance of different models. The BMA method was developed to combine the NCEP/CPC three statistical and one dynamical model forecasts of seasonal Ocean Niño Index (ONI) from 1982 to 2010. The BMA weights were derived directly from the predictive performance of the combined models. The highly efficient expectation–maximization (EM) algorithm was used to achieve numerical solutions. We show that the BMA method can be used to assess the performance of the individual models and assign greater weights to better performing models. The continuous ranked probability score is applied to evaluate the BMA probability forecasts. As an elaboration of the reliability diagram, the attributes diagram is used that includes the calibration function, refinement distribution, and reference lines. The combination of statistical and dynamical models is found to provide a more skillful prediction of ENSO than only using a suite of statistical models, a single bias-corrected dynamical model, or the equally weighted average forecasts from all four models. Probability forecasts of El Niño events based only on winter ONI values are reliable and exhibit sharpness. In contrast, an under-forecasting bias and less reliable forecasts are noted for La Niña.

## 1 Introduction

The El Niño–Southern Oscillation (ENSO) is a dominant large-scale coupled ocean–atmosphere phenomenon that strongly influences global climate and weather (e.g. Rasmusson and Wallace 1983; Glantz 2001; McPhaden et al. 2006; Sarachik and Cane 2010). As such, the predictability of climate and weather systems throughout the globe highly depends on the accuracy of ENSO prediction. It is thus of great importance to improve understanding and make more skillful and reliable forecast of ENSO. Significant progress has been achieved in forecasting ENSO in the past (e.g., Chen et al. 1995; Barnston et al. 1999, 2012; Coelho et al. 2004; Kirtman and Min 2009; Zhang et al. 2017). Traditionally, ENSO forecasts are obtained through dynamical or statistical modeling.

Statistical models are produced by the statistical relationships with the historical data. Over the last three decades, the National Centers for Environmental Prediction (NCEP)/Climate Prediction Center (CPC) of NOAA has developed some statistical tools for monthly and seasonal sea surface temperature (SST) forecasts in the tropical Pacific that include: Constructed Analogues (CA) (van den Dool 1994), Canonical Correlation Analysis (CCA) (Barnston et al. 1994; He and Barnston 1996) and Markov Model (MKV) (Xue and Leetmaa 2000). We use the seasonal SST forecasts from these three statistical models. The skill of CA is competitive with other empirical as well as dynamical methods (Barnston et al. 1994). An evaluation of the period 1996–1998 (Barnston et al. 1999; Landsea and Knaff 2000) shows CA and CCA to be the clear frontrunners among the empirical methods. Statistical models are useful for the foreseeable future due to their reduced cost and simplicity to develop as compared to dynamical models, which add in the difficulty of dealing with the complexity of nature.

✉ Pao-Shin Chu
chu@hawaii.edu

[1] Department of Atmospheric Sciences, School of Ocean and Earth Science and Technology, University of Hawaii at Manoa, Honolulu, HI 96822, USA

[2] Climate Prediction Center, NCEP, NOAA, College Park, Maryland, USA

Dynamical models are produced by the physical understanding of the atmosphere, land, ocean, and their interactions. One of the advantages of dynamical models is that they attempt to capture nonlinear interactions of climate systems and are adaptable to shifts in climate regimes. The Climate Forecast System Version 2 (CFSv2) is one of the most widely used fully coupled dynamical models. It was made operational at the NCEP in 2011. The CFSv2 model greatly improves the global SST forecasts and creates a vast array of products for subseasonal and seasonal forecasting with an extensive set of retrospective forecasts (Saha et al. 2014). Previous studies suggest that CFSv2 has a statistically significant improved and visibly better probabilistic reliability of ENSO prediction than CFSv1 model (Barnston and Tippett 2013).

Statistical and dynamical models have strengths and weaknesses in predicting ENSO which indicates that if only one type of model is considered and implemented, it can only provide a limited predictability of climate variables and may ignore and underestimate the uncertainty of the ENSO prediction. The strengths and weaknesses of individual models have led model evaluation studies to conclude that "no single model can be considered 'best' and it is important to utilize the results from a range of coupled models" (McAvaney et al., 2001). The above reasons inspired the current study of combining multiple models in Nino 3.4 SST forecasts. Use of multi-model averaging is a promising approach that takes the advantages of the statistical and dynamical information and therefore produces more skillful predictions than a forecast derived from a single model simulation. Thus, a multi-model prediction system is of great importance to forecast the behavior of ENSO. There is a general consensus in the seasonal forecasting community that probabilistic forecast information should be adopted (e.g., Kirtman and Pirani 2009). More specifically, Tebaldi et al. (2004) suggested that a Bayesian probabilistic approach is a useful platform from which to synthesize the information of simulation.

The Bayesian model averaging (BMA) method was concisely presented by Raftery et al. (2005) as a statistical method for postprocessing the ensembles and producing probabilistic forecasts from ensembles in the form of predictive probability density functions (PDF). BMA weights can be used to estimate the relative importance of each model and hence used as a basis for selecting models (Raftery et al. 2005). In other words, the BMA weights of models can be considered as their relative contribution to predictive skill over the training period. Thus, BMA differs from other model averaging methods in that it not only describes the uncertainty associated with model simulations but also provides the diverse capabilities of different models (e.g. Fang and Li 2016). The BMA method has been widely applied to various scientific areas, including: soil moisture simulation (Tian et al. 2012), economic forecasting (Faust and Wright 2013), meteorology and hydrology problems (Raftery et al. 2005; Gneiting et al. 2005; Min et al.. 2007; Vrugt and Robinson 2007; Bishop and Shanley 2008; Wang et al. 2012).

In this study, we apply the BMA method to multi-model SST forecasts over the Nino 3.4 region. The aim is to use BMA to weight a combination of statistical and dynamical models so that the weighted estimate is a better predictor of ENSO than any single model. The BMA weights are derived directly from the predictive performance of the combined models. The maximum likelihood estimation of model parameters based on Expectation–Maximization (EM) algorithm (e.g. Chu and Zhao 2011) is used in this study. A prior that gives preference toward evenly distributed weights will be applied.

Section 2 describes the relevant information of the models and dataset that has been used in this study. The BMA methodology and verification tools are introduced in Sect. 3. In Sect. 4, the results are presented and analyzed. In Sect. 5, we conclude with a discussion of the strengths and weaknesses of BMA and what we consider promising directions of extending this work.

## 2 Data and description of four operational forecast models

The data used in this study are the seasonal sea surface temperature (SST) forecasts in the Niño 3.4 region (5°N–5°S) (170°W–120°W) from 1982 to 2010 using four climate models managed by the Climate Prediction Center (CPC). The three statistical models are CA, CCA, and MKV, whereas the dynamical model is the CFS v2. The CA model produces a statistical forecast that is a linear combination of past observed anomaly patterns in the predictor fields such that the combination is as close as desired to the initial state (or 'base'). This can be expressed as $A\alpha = b$ where $\alpha$ is determined by minimizing the distance between A (predictand) and b (predictor) (Van Den Dool 1994). This can be achieved by a least square fit using standard matrix inversion as multiple linear regression. The CCA model is a multivariate regression technique that relates patterns in the predictor fields to patterns in the predictand field (e.g., Yu et al. 1997). The Markov Model is a statistical model built in a reduced multivariate empirical orthogonal function (MEOF) space and represents the sea surface temperature anomaly, sea level and wind stress anomaly fields. In the MEOF calculations, the anomalous SST and sea level fields are normalized by the square root of the total variance and then combined to construct the covariance matrix. The Markov model is defined by $X_{t+1} = AX_t + \varepsilon_t$, where $X_t$ is the principle components of MEOF at the tth month, A is the transition matrix and $\varepsilon_t$ is the residual (Xue and Leetmaa 2000).

The dynamical model used in this study is the CFSv2 model. It was made operational at NCEP in March 2011 and provides retrospective forecasts (also known as hindcasts) from 1982 to 2010 and onward for real time subseasonal and seasonal predictions (Saha et al. 2014). The CFSv2 data have a negligible systematic error in the later years (after 1998), whereas the earlier years have a modest cold bias. This occurs because in later years the models are initialized with much more data. Thus, the dynamical model (CFSv2) retrospective forecasts need to be calibrated. This was done by correcting the modest cold bias before 1998 and using linear regression on the bias corrected data. All the data used in this study are 3-month running averages of SST forecast for Niño 3.4 from 1982 to 2010 (hindcast analysis time period) with lead times of one to 7 months. The observation data are available at the CPC website.

A hindcast is also known as historical re-forecast or retrospective forecast and integrates the model forward in time. The difference between a forecast and hindcast is that the latter performs the forecast again using the information that was not available originally. That new information might be observations (for assimilation or for verification), assimilation system, or forecast model. For example, let $t_0$ be the time instant of interest, $t_{-1}$ be some time before $t_0$, and $t_{+1}$ be some instant in time after $t_0$. Initializing the model at $t_{-1}$ and runs through to $t_{+1}$. If a forecast system can make use of observations at $t_0$, then it would be used in the same way that it would with a forecast. Figure 1 illustrates the lead time structure for a forecast of the winter target season (December-January-February) of 1984/1985. Each of the four rows represents a different lead time. For a 1 month lead time, October is the latest observation available at the issued time for forecast of December-January-February (DJF). This works similarly for the other lead times. For convenience, the words "hindcast" and "forecast" are used interchangeably in this study.

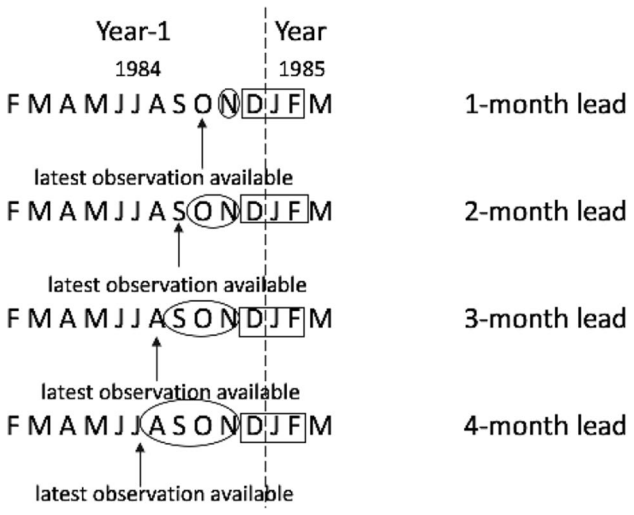


**Fig. 1** An example of the timing of the seasonal forecasts for 4 lead times in months. An arrow denotes the latest observation available for forecast

## 3 Methodology and verification tools

### 3.1 Bayes' Theorem

Bayesian theorem can be expressed as

$$P(\theta|y) = \frac{P(y|\theta)f(\theta)}{\int_{\theta} P(y|\theta)f(\theta)d\theta} \quad (1)$$

where θ represents the parameter(s) of the distribution (for example, the mean and variance for a Gaussian distribution or a Poisson intensity rate), and y is the available data. The prior information regarding θ is quantified by the prior distribution f(θ). The likelihood function p($y$|θ) represents the data-generation process and the quantitative influence of different values of θ. The likelihood function p($y$|θ) also expresses the relative "likelihood" of the data at hand as a function of different possible values for θ. The Bayesian approach combines the likelihood with the prior distribution to obtain the posterior distribution of θ, p(θ|y), which is the probability density function for the parameters θ characterizing the current best information regarding uncertainty about θ.

### 3.2 Bayesian model averaging

Bayesian model averaging can be used for deriving the relative weights and variances of the normal conditional PDFs of the individual model. For different climate models, k = 1, 2, …, K, the joint PDF of y conditional on $y_k$ is given by a weighted average of the individual model predictive density as

$$f(y|y_k, k = 1, ..., K) = \sum_{k=1}^{K} w_k \cdot f_k(y|y_k) \quad (2)$$

where y is the observation (i.e., SSTs), $y_k$ is the corresponding forecast SST value from the model k, $w_k$ is the BMA weight for model k and is a nonnegative value that satisfies $\sum_{k=1}^{K} w_k = 1$. The weights $w_k$ are estimated by maximum likelihood based on the model k's performance in the training data set. The simulation skill of the model during the training period relative to other models can be represented by the weights as well. The conditional PDF for the observation given the corresponding simulated variable for model k is $f_k(y|y_k)$. It is assumed that $f_k(y|y_k) \sim N(\bar{y}_k, \sigma_k)$ is a Gaussian distribution and centered on the model forecast

SST values $\bar{y}_k$, with a standard deviation of $\sigma_k$. Note that $\bar{y}_k$ and $\sigma_k$ vary with different seasons and lead times. The BMA deterministic forecast results are computed as $\sum_{k=1}^{K} w_k y_k$ and this can be compared with each model forecast and the multi model ensemble average method.

The weights $w_k$ and variances $\sigma_k^2$ are estimated by maximum likelihood method (Raftery et al. 2005). The maximum-likelihood method is used to estimate the most probable values for the parameters, given the observed data. Instead of maximizing the likelihood function itself, the logarithm of the likelihood function (or log-likelihood function) was used to get the maximum likelihood because of its simplicity and stability. For the forecast distribution of $f_k$, the log-likelihood function, $\mathcal{L}$ is computed as

$$\mathcal{L}\left(w_1, \ldots, w_k, \sigma_k^2\right) = \sum_{t=1}^{T} log \sum_{k=1}^{K} w_k \cdot f_k(y^t | y_k^t,\ \sigma_k^2) \tag{3}$$

where the summation is over all k = 1… K models and t = 1…T observations of the training data set. In this study, the expectation–maximization (EM) algorithm is used to derive the maximum likelihood estimation for model parameters. The EM algorithm has advantages in this study because it is relatively easy to implement and computationally efficient. As indicated in the following Eqs. (4) and (5), the EM algorithm has two steps: In the expectation step, the values of $z_k^{t,(j+1)}$ are calculated given the current values of the BMA weights and variances. The E step is given by Eq. (4) where the function $f_k$ returns the density of a normal distribution with mean $y^t|y_k^t$ and standard deviation $\sigma_k^{(j)}$, and the superscript j signifies iteration counter. In the maximization step, the values of $w_k$ and $\sigma_k^2$ are updated using the current estimates of $z_k^{t,(j+1)}$:

$$z_k^{t,(j+1)} = \frac{w_k^{(j)} f_k(y^t | y_k^t,\ \sigma_k^{2(j)})}{\sum_{k=1}^{K} w_k^{(j)} f_k(y^t | y_k^t,\ \sigma_k^{2(j)})} \tag{4}$$

$$w_k^{(j+1)} = \frac{1}{T} \sum_{t=1}^{T} z_k^{t,(j+1)}$$

$$\sigma_k^{2(j+1)} = \frac{\sum_{t=1}^{T} z_k^{t,(j+1)} \left(y^t - y_k^t\right)^2}{T \sum_{t=1}^{T} z_k^{t,(j+1)}} \tag{5}$$

The maximization step starts with an initial guess for the weights. In this study, a uniform distribution is given to all weights in iteration "0" so that the initial weight for each of the model is $K^{-1}$. The EM algorithm alternates between an expectation and a maximization step. The BMA weights are then estimated iteratively with Eqs. (4) and (5) until the algorithm reaches a convergence of the log-likelihood function in Eq. (3). The convergence is defined as the change of $\mathcal{L}$ between two consecutive iterations is no longer greater than a predefined small tolerance ($10^{-8}$). In this study, BMA is developed for each season independently.

## 3.3 Verification tools

### 3.3.1 Verification scores

In this study, the performance of the BMA deterministic forecasts is assessed through root mean square error (RMSE) and skill score. BMA probability forecasts are verified using continuous rank probability score (CRPS) and the attributes diagram. For each of the forecasts, a leave-two-out cross-validation is applied to the models because the autocorrelation of hindcasts remains rather large at the first two lags and decrease abruptly after lag 2. We first introduce the root mean square error, then move on to skill score, CRPS, and the attributes diagram.

Root mean square error (RMSE) is the standard deviation of the prediction errors. It can be expressed as follows.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \sum_{k=1}^{n} \left(y_k - O_k\right)^2} \tag{6}$$

Equation (6) shows the root average squared difference between the forecast ($y_k$) and observation ($o_k$) pairs. The RMSE increases from zero (perfect forecasts) to larger positive values as the discrepancies between forecasts and observations become increasingly large.

Forecast skill is usually presented as a skill score, which can be interpreted as a percentage improvement over the reference forecasts. The generic form of the skill score is shown as follows.

$$\mathrm{SSref} = \frac{A - A_{ref}}{A_{pref} - A_{ref}} \times 100\% \tag{7}$$

The skill score is represented as a particular measure of accuracy $A$ with respect to the accuracy $A_{ref}$ of a set of reference forecasts. The value of the accuracy measure that would be achieved by perfect forecasts is characterized by $A_{perf}$. Equation (7) can be constructed by using the mean absolute error (MAE), mean square error (MSE) or RMSE as the underlying accuracy measure. In our study, MSE values are used as the accuracy statistic. The skill score can be expressed as

$$\text{Skill Score} = \frac{MSE - MSE_{clim}}{0 - MSE_{clim}} = 1 - \frac{MSE}{MSE_{clim}} \tag{8}$$

$$MSE_{clim} = \frac{1}{n} \sum_{k=1}^{n} (\bar{o} - o_k)^2 \tag{9}$$

where perfect forecasts have $MSE = 0$ and $MSE_{clim}$ is the climatological mean square error values of the predictand. If $MSE = MSE_{clim}$, the skill score attains 0%, indicating no improvement over the reference forecasts.

The Continuous Ranked Probability Score (CRPS) is used to evaluate the difference between the BMA probability forecast and the observed values. It differs from the RMSE and skill score in that it focuses on the probability distribution of the forecast. It is defined as

$$CRPS = \frac{1}{T}\sum_{t=1}^{T}\int \left[F\left(x^t\right) - O\left(x^t\right)\right]^2 dx \tag{10}$$

$$O\left(x^t\right) = \begin{cases} 0, & x < x_0 \\ 1, & x \geq x_0 \end{cases} \quad x_0 \text{ is observation} \tag{11}$$

where $F(x^t)$ is the forecast probability cumulative distribution function (CDF) for the tth forecast case and $O(x^t)$ is the observation CDF. The CRPS has a negative orientation, so the smaller the value the better the forecast.

#### 3.3.2 Attributes diagram

The simple forecast performance measure such as the RMSE is a convenient and quick view but a comprehensive understanding of forecast quality can be achieved through a graphical format such as reliability diagram or attributes diagram. The attributes diagram is an elaboration of the reliability diagram that includes the calibration function, refinement distribution, and reference lines related to the algebraic decomposition of the Brier score and the Brier skill score (Wilks 2011). Specifically, the attributes diagram provides a geometrical framework to compare the empirical curve with lines constituting sets of reference points with respect to specific attributes (Hsu and Murphy 1986) and is adopted in this study. The attributes contain the reliability, sharpness, resolution and the uncertainty of a probability forecast with respect to the observation. Reliability is measured by how closely the forecast probabilities correspond to the conditional frequency of event occurrence. A perfect reliable forecast would be indicated by a reliability line plotted along the 1:1 line between the forecast probability and the observed relative frequency. For example, when a probability forecast of 0.10 is issued, we would expect the event to occur 10% of the time.

Note that a forecasting system that simply forecasts the climatological probabilities of events may be reliable, but is not useful. Thus, it is useful to acquire the sharpness of the forecast as well. Sharpness is the tendency to forecast extreme values (probabilities near 0 or 100%) rather than values clustered around the mean, e.g., a forecast of climatology has no sharpness. It also indicates the variability in the forecasts. Forecast systems that are capable of predicting events with probabilities different from the observed event frequency are said to be 'sharp'. A forecasting system that has sharpness but not reliable is indicative of an unrealistic confidence. Resolution indicates the ability of the forecast to distinguish situations with distinctly different frequencies of occurrence. In other words, it measures how well the observations are "sorted" among the different forecasts.

## 4 Results

In this section we present results of the application of BMA to SST forecasts for the Nino 3.4 region using CA, CCA, MKV, and CFS forecast tools from 1982 to 2010. We will first distinguish the difference between BMA deterministic and probabilistic forecasts. Changes of the weights for individual models in different lead times and target seasons will then be discussed, followed by the RMSE, skill score, CRPS and the attributes diagram based on the cross-validation results of BMA and individual models for various lead times and target seasons. In particular, the reliability of seasonal ONI forecasts will be assessed.

### 4.1 Example of BMA probability forecast and deterministic forecast

Sometimes BMA probability and deterministic forecasts can be confusing. Here, we will give an example to distinguish the difference between these two kinds of forecasts. Figure 2 shows the BMA predictive PDF, weighted PDFs for each of four models and BMA deterministic forecast for one-month lead of August-October (ASO) 2004. The BMA predictive PDF (solid curve) is the weighted sum of four individual PDFs (broken curve). In Fig. 2, the observation is 27.53 °C (solid vertical line) and BMA deterministic forecast is 27.06 °C (a red cross). The observation in this case falls into the 90% BMA prediction interval (dashed vertical lines), although the BMA deterministic forecast is slightly different from the observation. It is obvious that the estimated BMA PDF provides a more reliable description of the probability forecast than any of the individual models. This indicates that BMA can be used to describe the uncertainties associated with each model simulation and provides a probability forecast. Probability forecast verification also can be applied to evaluate whether the observations are within the spread of expected outcomes.

### 4.2 BMA weights, lead times and target seasons

Figure 3 shows the BMA weights in lead one month for different target seasons [DJF, March-April-May (MAM), June–July–August (JJA) and September-October-November

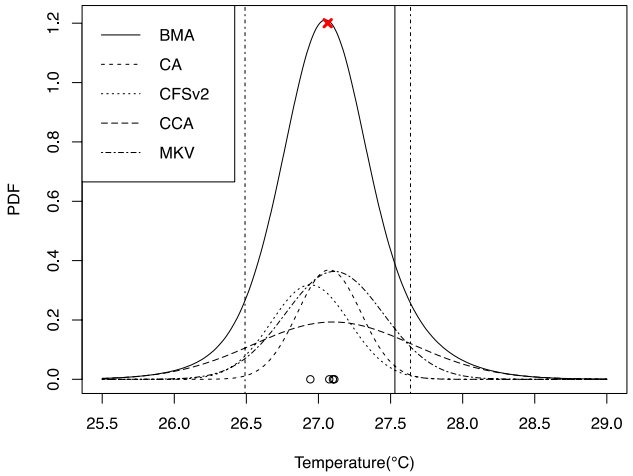


**Fig. 2** An example of the BMA probability forecast (solid pdf), weighted individual model probability forecasts (CA, CFSv2, CCA, MKV), deterministic forecast of each model (circle) and BMA deterministic forecast (red cross) for one-month lead of August-September-October 2004. The vertical solid line represents the observed value. BMA probability forecast is a weighted sum of each model PDF. Dashed vertical lines are BMA 90% prediction interval

(SON)] in the form of box plots. The distribution of the weights among the models varies in different target seasons. The weights represent each individual forecast tools' performance relative to other tools. In DJF, CA and MKV exhibit relatively high weights while CFSv2 does not receive any noticeable weights (almost zero). The location of the median is in the upper end of the box for CA and MKV, suggesting a tendency towards negative skewness. In MAM, CFSv2 is the model with the highest weight and the weights for CA are close to zero. The weights are distributed mainly on CA and the minimum weights are associated with MKV in JJA. In SON, CFSv2 and MKV received higher weights than CA and CCA. Throughout these four seasons, there is not a single model that consistently has the best weight.

From Fig. 3, the weight of each model seems to change from one running season to another. There are several reasons for changes in the weight of each individual model for different target seasons. First, Bayesian model averaging is developed for each season independently so there is no persistence carried through from one season to another. Second, there is collinearity among forecasts from different models. Indeed, some of the model forecasts are highly correlated and not independent. Peng et al. (2002) discussed the unstable weights from the multiple linear regression technique resulting from the collinearity among the predictors. Third, the sample size used is relatively small (n = 38) so the weights may change for different seasons. However, the weight of each model does not vary dramatically with leads. To illustrate this point, each of the four models was selected for the target season when their weights are relatively high. For example, for the target season of AMJ (Fig. 4), the weight of the CFSv2 (~ 0.7 to 0.8) is very high from lead one to seven (months). For CA, its weight remains high and stable from lead one to five for the target season DJF. This is also the case for the CCA for the JFM target season. For MKV, the weight drops gradually from lead one to four, probably because of the nature of the imposed multivariate first-order autoregressive process (i.e., a simple Markov chain).

To understand the relationship between model performance and their weights, we display the RMSE and skill score of each individual model for lead one month in different overlapping target seasons. Figure 5 shows that CFSv2 performed the poorest compared to other models in DJF with the highest RMSE and lowest skill score. However, from February-March-April (FMA) to May–June–July (MJJ), CFSv2 performed the best with the lowest RMSE and the highest skill score relative to all three statistical models. The results of the forecast performance and the weights (Fig. 3) are obviously in agreement, where higher weights correspond to better forecast performance. These results also show that the BMA method can take advantage of the diverse capabilities of the different models.

### 4.3 Combination of statistical-dynamical BMA models and comparison with other models

In the following, BMA is applied to two groups of models: one group consists of purely statistical models (CA, CCA and MKV) and another group is the combination of all statistical and dynamical models (CA, CCA, MKV and CFSv2). Figure 6a shows the summary of RMSE values for statistical-dynamical BMA models in different target seasons and lead times. The RMSE has a negative orientation, implying smaller values correspond to a better forecast. In general, RMSE values are larger for longer lead times. A typical issue of many of the models in ENSO predictions is the poor performance when forecasts go through boreal spring (e.g., Kirtman and Min 2009). This so-called spring predictability barrier is also reflected in Fig. 6a. For example, the forecast for July-August-September (JAS) made in January (lead 5 month) has a RMSE value of 0.6–0.7. This RMSE is relatively high compared to other seasons (e.g., for the target season of JFM also for 5 months' lead). The boreal spring is a transition season when the tropical ocean–atmosphere interaction in the Pacific is usually the weakest and the ENSO signal is relatively not well defined.

We display the RMSE values of statistical-dynamical BMA minus the RMSE values of statistical BMA model in Fig. 6b. For most of the target seasons and lead times, the values are negative which indicates that the overall forecast performance of statistical-dynamical BMA model is better than purely statistical BMA models. Therefore, the

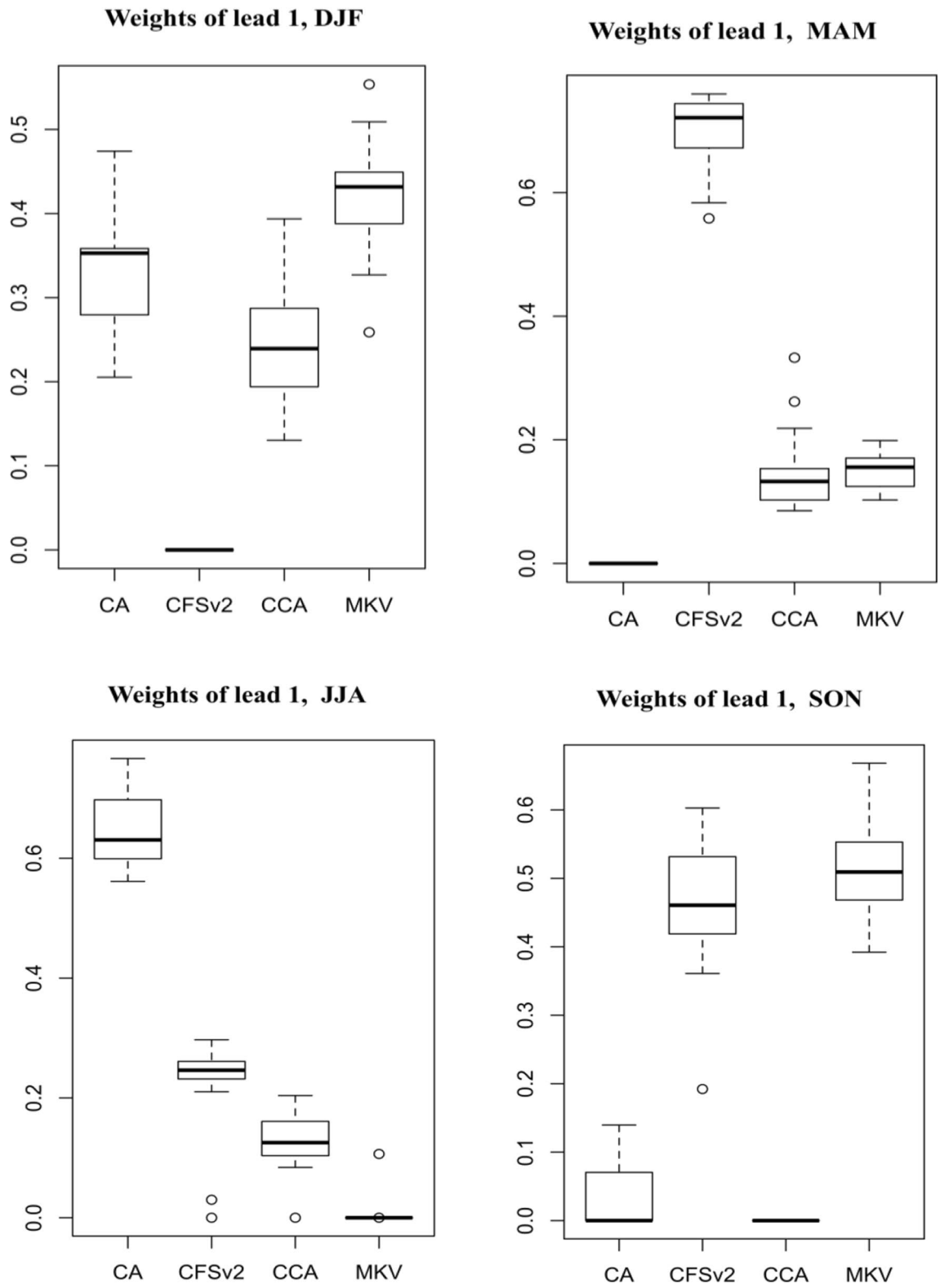


**Fig. 3** Box-plots of BMA weights in cross validation for lead 1 month and seasons DJF, MAM, JJA and SON. The box represents the interquartile range of the weights and the thick line is the median. The outliers are characterized as the open circles

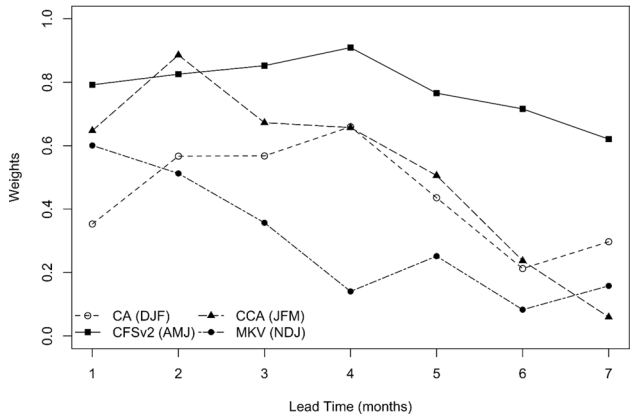


**Fig. 4** BMA weights as a function of leads (months) for four selected models with different target seasons

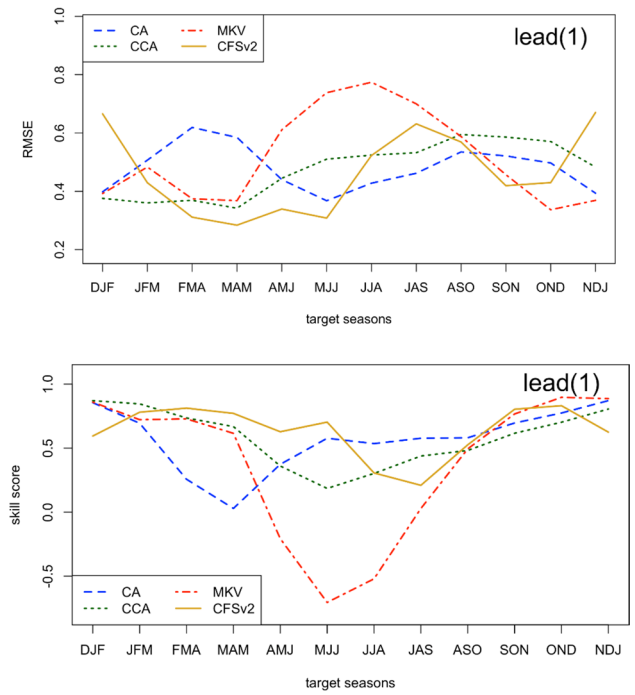


**Fig. 5** RMSE and skill score of each model at lead one month for different target seasons

statistical-dynamical BMA models outperform statistical BMA models in predicting the Nino 3.4 SST, especially through the spring predictability barrier due to the better performance of the dynamical model. Statistical models are developed on monthly or seasonally averaged data and receive inputs that are relatively older. In contrast, dynamical models run more frequently using the most recent observed data as input. Moreover, dynamical models ingest much more observations (such as the subsurface ocean conditions) using complex data assimilation schemes (Barnston et al. 2012). These advantages allow dynamical models to forecast

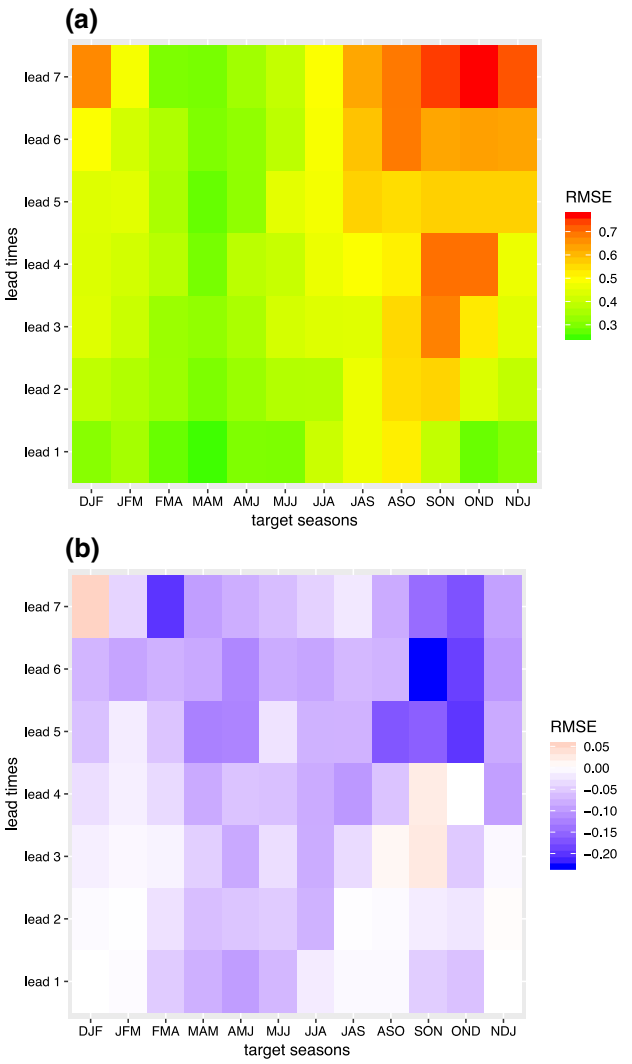


**Fig. 6** **a** RMSE of statistical-dynamical BMA models, **b** the difference of RMSE between the statistical-dynamical models and statistical BMA models for various leads (months) and target seasons

better and predict the important changes more accurately. Figure 6b also shows that the difference between statistical-dynamical BMA models and statistical BMA models has become larger at longer lead times. Similar to Fig. 6b, a comparison is also made between the statistical-dynamical BMA model and the single dynamical model. Again, the optimal combination of a suite of models has a smaller RMSE than the CFSv2 for most leads and target seasons (not shown).

Figure 7 is the same as Fig. 6 but for skill score. The skill score has a positive orientation, so the larger the values the better the forecast. The spring predictability barrier of ENSO and the dependence of skill score on lead times are also indicated. In Fig. 7b, the differences are positive for most of the target seasons and lead months which indicate that statistical-dynamical BMA models outperform the statistical

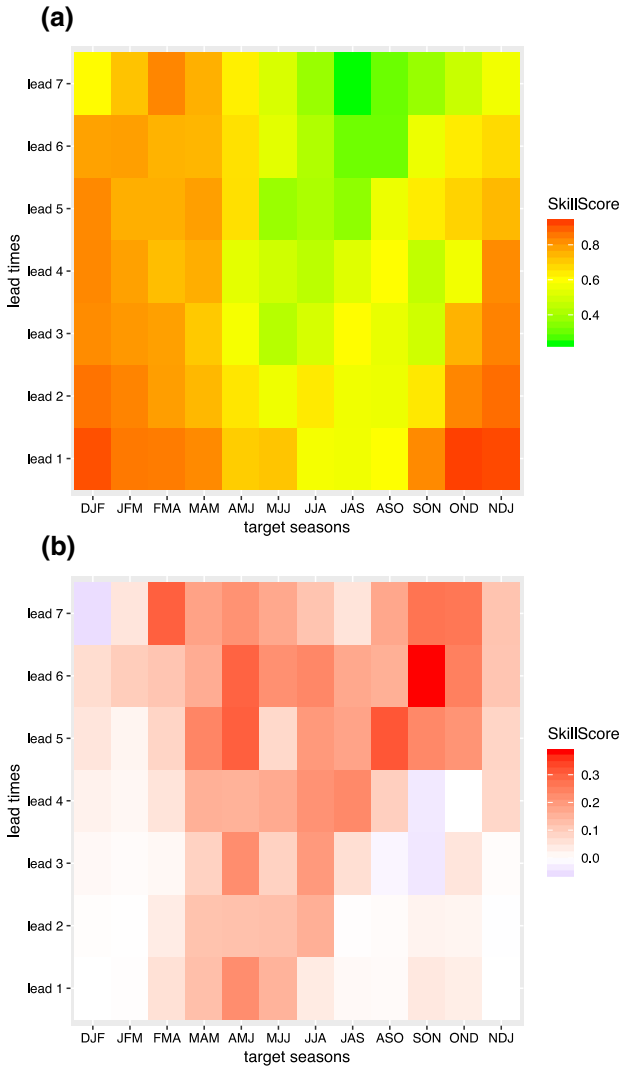


**Fig. 7** Same as Fig. 6 but for skill score

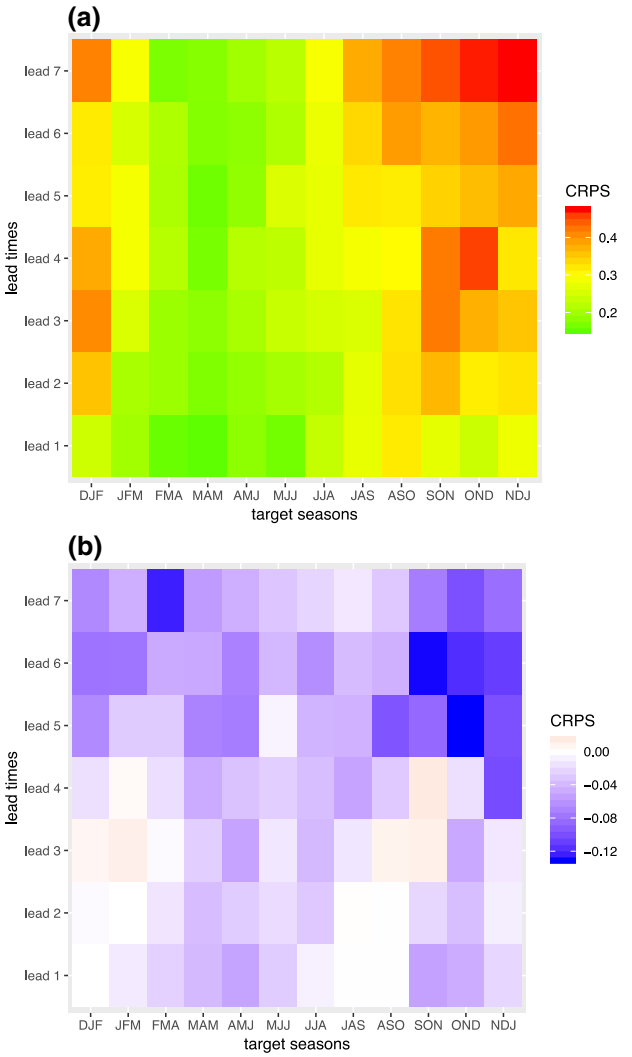


**Fig. 8** Same as Fig. 6 but for CRPS

BMA models. The skill score results are consistent with the RMSE results. The CRPS results are also displayed to validate BMA probability forecasts. Figure 8 shows similar results as the RMSE. Overall, the forecast skills have been improved by taking the advantage of the relative strengths of statistical and dynamical models through BMA, particularly during the ENSO spring predictability barrier.

### 4.4 A comparison of the BMA and multi model ensemble average (MMEA) methods

It is also of interest to compare the BMA forecasts with that of the multi model ensemble average (MMEA) forecasts, while the latter is a common approach for averaging multi model outputs by assuming an equal weight for each model conditional on the fact that the sum of model weights equal to one. Results of the difference between BMA and MMEA method in terms of RMSE values are shown in Fig. 9. This is calculated with the RMSE values of BMA deterministic forecast (mean of the mixture density) minus the RMSE values of MMEA forecast (simply average the four model's deterministic forecasts). In Fig. 9, the majority of RMSE values are negative (about 67% of all the lead months and target seasons), indicating that the BMA forecasts outperform MMEA forecasts for most of the target seasons and lead months, especially during spring and summer seasons. For some seasons such as DJF to FMA, MMEA is better than BMA but the differences of RMSE values between these two methods are very small (~0.00–0.05). Note that because hindcasts are used and the CFSv2 model has been calibrated to remove cold bias, the difference between hindcasts from each model and the corresponding observation is small. Therefore, the difference in RMSE between the BMA and MMEA is also small, although a majority of the

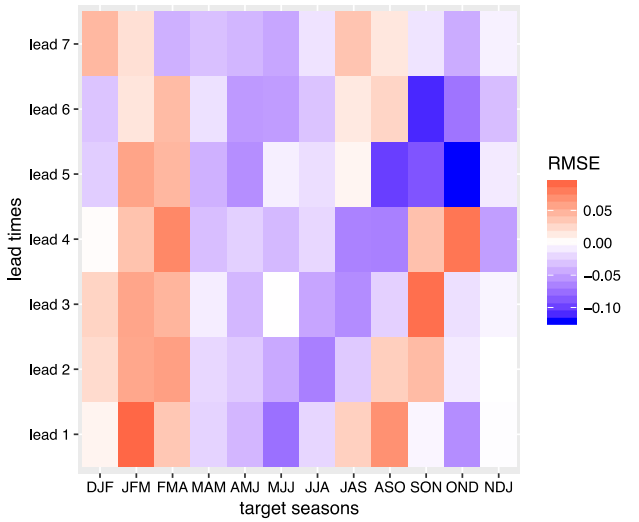


**Fig. 9** Difference of RMSE between the BMA and multi model ensemble average (MMEA)

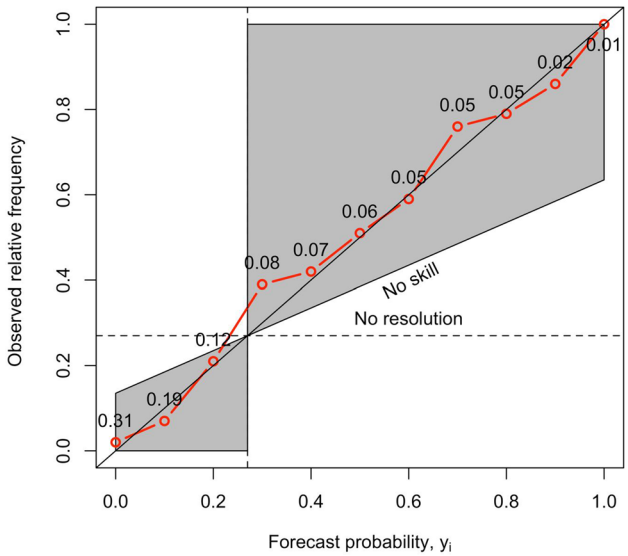


**Fig. 10** Attributes diagram for forecasts probability of seasonal ONI larger than 0.5. This was made by pooling all target seasons, lead months and years together with 0.1 forecast bin width. The relative frequency of each of the forecast values is shown in the numbers. The 45° solid line indicates perfect reliability. The no-resolution line is plotted at the same level of the sample climatological probability. The no-skill line is halfway between the perfect reliability and no-resolution lines

BMA hindcasts has a smaller RMSE than the MMEA hindcasts. Moreover, BMA also considers the uncertainty of each model's forecasts and uses this uncertainty to construct a predictive distribution instead of only a deterministic forecast. In other words, the main advantage of the BMA method is that it provides a forecast distribution that can be used for probabilistic analysis and prediction.

### 4.5 Reliability in probability forecasts of large seasonal ONI values

To assess the overall statistical-dynamical BMA forecast reliability of ONI, the attributes diagram (Hsu and Murphy 1986) for forecast probabilities of seasonal ONI larger than 0.5 (Fig. 10) are displayed. The CPC uses the ONI values in excess of +0.5 °C (less than −0.5 °C) for at least five overlapping seasons (3-month average) as criteria of El Niño (La Niña) episodes. Here, the attributes diagrams are made by pooling all the target seasons, lead times and years together (more than 2000 events). This approach is also used in Schepen et al. (2014) for verifying seasonal forecasts of Australian rainfall.

In Fig. 10, the points show observed relative frequency of ONI > 0.5, conditional on each of the i = 11 (i = 0.1, 0.2, 0.3…1) possible forecast bins. The forecast probability bin width is 0.1. It is noticeable that the points are very close to the perfect forecast line (45° line) which indicates that the overall statistical-dynamical BMA probability forecasts are consistent with the observed ONI > 0.5 frequency and are reliable. No resolution means that a forecast of climatology does not discriminate between events and non-events at all. In this case, no points fall on the no-resolution line, and the forecast exhibit a substantial degree of resolution. The numbers next to the points (sharpness) express the relative frequency with which the event has been predicted (over the reference period and at all events) with different levels of probability.

Note that forecast systems that are capable of predicting extreme events (e.g., probability near 0 or 100%) are said to have "sharpness," meaning it measures the specificity of the probabilistic forecast. Given two reliable forecast systems, the one producing the sharper forecasts is preferable. In Fig. 10, the majority of forecasts predict low probabilities. In other words, there is a tendency for forecast probabilities to be near zero (0.31) or at low forecast probabilities (e.g., 0.19). Therefore, the forecasts in this case exhibit sharpness. The forecast system is also capable of predicting relatively high probabilities but such forecasts are less common (e.g., 0.01). Points in the shaded area bounded by the lines of "no skill" and the overall sample climatology (i.e., the vertical solid line) indicate the positive contribution to forecast skill. In this case, all the points are located inside the shaded area.

The attributes diagram for the probability of ONI < − 0.5 is shown in Fig. 11. In this diagram, the reliability curve lies mainly above the 45° line especially for higher forecast probabilities. This indicates that the statistical-dynamical BMA model slightly under-forecasts the probability of ONI < − 0.5 for higher forecast probabilities (forecast probabilities too low). Nevertheless, the overall probability forecast still contribute positively to the prediction skill. The

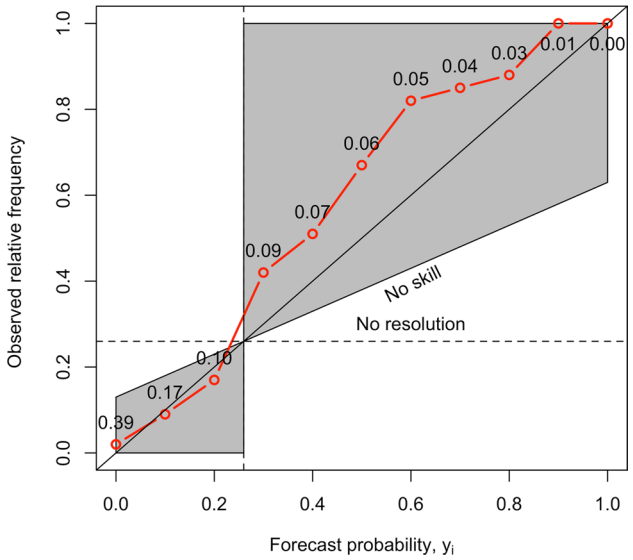


**Fig. 11** Same as Fig. 10 but for the forecasts probability of seasonal ONI smaller than −0.5

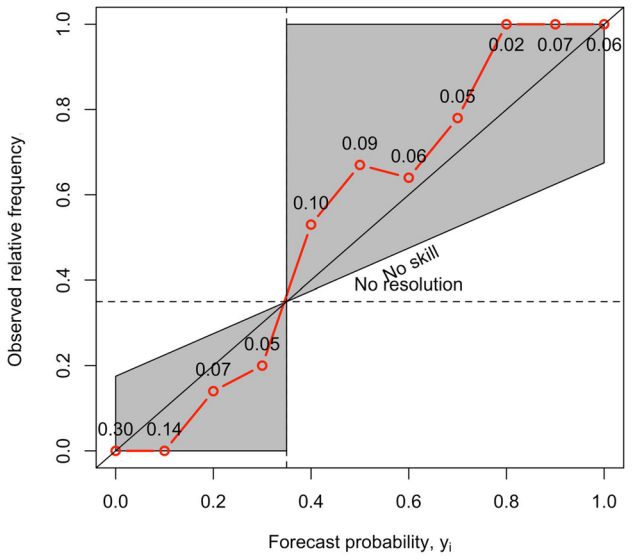


**Fig. 12** Attributes diagram of the forecasts probability for December–January–February ONI equal to or larger than 0.5. This was made by pooling all lead months and years together with 0.1 forecast bin width. The relative frequency of each of the forecast values is shown in the numbers. The 45° solid line indicates perfect reliability. The no-resolution line is plotted at the same level of the sample climatological probability. The no-skill line is halfway between the perfect reliability and no-resolution lines

probability forecast of ONI $< -0.5$ also exhibit sharpness. As a comparison, the reliability of forecasting ONI $> 0.5$ is higher than forecasting ONI $< -0.5$ using the BMA model (Figs. 10 and 11). Overall, statistical-dynamical BMA presents a reliable probability forecasts of ONI. Therefore, BMA can be applied not only to improve the forecast skill, but also provide a reliable probability forecast for ONI.

### 4.6 Reliability in probability forecast conditional on DJF ONI

Now we consider only the DJF ONI equals to or greater than (smaller than) 0.5 °C (− 0.5 °C) while ignoring ONI values in other seasons. This is motivated by the fact that ENSO usually reaches its peak phase in boreal winter. By doing so, our sample size is drastically reduced. Note that reliability or attributes diagrams are very sensitive to small sample sizes so the results presented here should be exercised in caution (e.g., Kirtman and Min 2009). Figure 12 shows that the probability forecasts of large and positive winter ONI values are reliable and all the points are inside the shaded area, indicating that the forecasts contribute positively to the prediction skill. The forecasts also exhibit sharpness. However, compared to Fig. 10, the probability forecasts of large DJF ONI values have less reliability than that based on all the seasonal ONI. For example, when the forecast probability is equal to 0.5, the actual chance of observing the December-January-February ONI equals to or larger than 0.5 is closer to 0.7. This is probably due to the smaller sample size used in forecasting winter ONI.

Figure 13 is the same as Fig. 12 but for the forecast probability of DJF ONI being smaller than −0.5. This forecast is less reliable, for example, the forecast probability is equal to 0.3 but the actual observed frequency is close to 0.8. This is an indication of a substantial underforecasting bias. However, the majority of the points are inside the shaded area. It is interesting to note that the probability forecasts of large

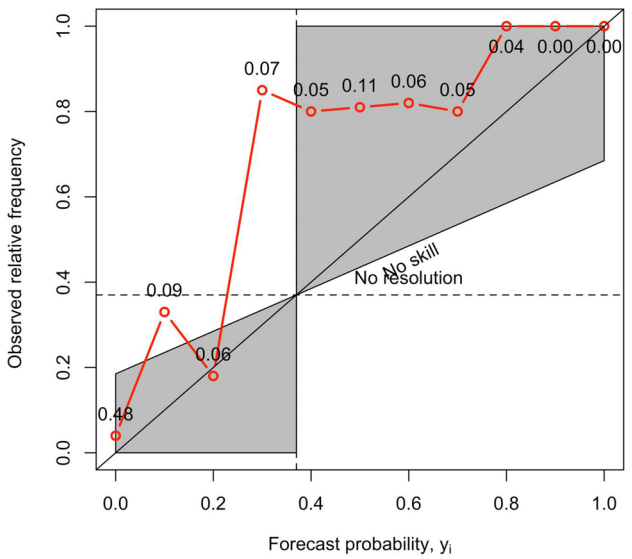


**Fig. 13** Same as Fig. 12 but for the forecasts probability for December-January-February. ONI equal to or smaller than −0.5

and positive DJF ONI values are more reliable than forecasts of large and negative DJF values.

## 5 Summary and discussion

This study aims for applying the BMA approach to dynamical and statistical models in forecasting SST over Nino 3.4 region. The forecast validation was assessed by using leave-two-out cross-validation which provides an estimate of skill for forecasting independent events. Our results clearly indicated a couple of important findings. One of them is that BMA, a statistical postprocessing method, can provide deterministic and probability results of ENSO prediction by assigning weights to different models. The BMA estimated PDF was better adjusted than any of the individual models and the 90% prediction interval of the BMA PDF contains almost the entire observations. This occurs for up to 95% of the cases in the study (not shown).

Perhaps the most interesting result is that the BMA method can be used to assess the relative performance of the statistical and dynamical models in different target seasons and lead times and take the advantage of the strength of the models by assigning the weights. In other words, it gives greater weights to better performing models. Our results also indicate that the BMA weights change with target seasons for each individual model which implies that model performance varies with target seasons.

BMA can be used to combine statistical and dynamical models, with the weighted estimates shown to have a lower RMSE, lower CRPS, and higher skill score than only using a combination of three statistical models, a single bias-corrected dynamical model, or a simple multimodel ensemble average. The overall reliability and sharpness of the ONI forecasts from the statistical and dynamical models are assessed using the attributes diagram. The forecast probabilities are rather consistent with the observed relative probabilities, suggesting that the forecast probabilities are reliable and they also exhibit sharpness. We also restrict the analysis to only the large positive and negative winter ONI values. For large and positive ONI, there is a small degree of overforecasting at low forecast probabilities and some underforecasting bias at high forecast probabilities. For large and negative DJF ONI forecasts, a substantial underforecasting occurs across a broad range of forecast probabilities. If a DJF ONI attains 0.5 or beyond can be simply regarded as an El Niño event and a value smaller than −0.5 as a La Niña event, El Niño forecasts seem to be more reliable than La Niña.

The bias-corrected CFSv2 model outperformed the statistical models with a higher skill score and lower RMSE during the spring predictability barrier due to the incorporation of the most recent changes in the observational evolution. Thus, combining with the CFSv2 model improves the forecast skill of ENSO, especially during the spring predictability barrier. CFSv2 also makes more skillful predictions for longer lead times than the statistical models. Indeed, at longer leads (4–7 months), the weight of the CFSv2 is higher relative to three statistical models and this occurs nine to ten times out of the twelve running seasons (not shown). Dynamical models are capable of capturing nonlinear compounding effects of anomaly growth in the climate system due to their time-marching design using small time steps, enabling faster evolution than statistical models. Trenberth (1998) also suggested that dynamical model forecasts show more skillful and reliable ENSO forecast in longer lead times.

Previous studies have shown that the model averaging method can be used to improve the skill of multi-model ensembles (Madigan and Raftery 1994). In this study, we use the BMA method to combine the statistical and dynamical model and yield a more reliable and accurate forecast of ONI. This is due to the fact that BMA takes the strength of individual models to optimize a combined weighted-average forecast and to make more skillful predictions. Moreover, it allows us to determine which models are the most important in various target seasons and lead times. This method can be applied in other climate variables and models as well. The original BMA approach introduced by Raftery et al. (2005) assumes that the conditional probability density function of each individual model is adequately described by Gaussian or Gamma statistical distribution. However, recent work (Rings et al. 2012) presented a variant of BMA with a flexible representation of the conditional forecast distribution. The BMA method developed here, based on previous work (Sloughter et al. 2010; Wang et al. 2012), is thus applicable to a large fraction of research in atmospheric sciences and highly suitable for weather forecasting.

**Acknowledgements** This study was funded by the Hawaii State Climate Office through SOEST at the University of Hawaii at Manoa and the Climate Prediction Center of NCEP. We thank May Izumi for her editing service.

## References

Barnston AG, van den Dool HM, Zebiak SE, Barnett TP, Ji M, Rodenhuis DR, Cane MA, Leetmaa A, Graham NE, Ropelewski CF, Kousky VE, O'Lenic EA, Livezey RE (1994) Long-lead seasonal forecasts—where do we stand? Bull. Am Meteorol Soc 75:2097–2114

Barnston AG, Glantz MH, He Y (1999) Predictive skill of statistical and dynamical climate models in SST forecasts during the 1997/98 El Nino episode and the 1998 La Nina onset. Bull Am Meteorol Soc 80:217–243

Barnston AG, Tippett MK, L'Heureux ML, Li S, DeWitt DG (2012) Skill of real-time seasonal ENSO model predictions during 2002–11: is our capability increasing? Bull Am Meteorol Soc 93:631–651

Barnston AG, Tippett MK (2013) Predictions of Nino3. 4 SST in CFSv1 and CFSv2: a diagnostic comparison. Clim Dyn 41:1615–1633

Bishop CH, Shanley KT (2008) Bayesian model averaging's problematic treatment of extreme weather and a paradigm shift that fixes it. Mon Wea Rev 136:4641–4652

Chen D, Zebiak SE, Busalacchi AJ, Cane MA (1995) An improved procedure for EI Niño forecasting: implications for predictability. Science 269:1699–1702

Chu P-S, Zhao X (2011) Bayesian analysis for extreme climatic events: a review. Atmos Res 102:243–262

Coelho CAS, Pezzulli S, Balmaseda M, Doblas-Reyes FJ, Stephenson DB (2004) Forecast calibration and combination: a simple Bayesian approach for ENSO. J Climate 17:1504–1516

Fang M, Li X (2016) Application of Bayesian model averaging in the reconstruction of past climate change using PMIP3/CMIP5 multimodel ensemble simulations. J Clim 29:175–189

Faust J, Wright JH (2013) Forecasting inflation. In Handbook of economic forecasting (Vol. 2, pp. 2–56). Elsevier

Glantz MH (2001) Currents of change: impacts of El Niño and La Niña on climate and society. Cambridge Univ. Press, Cambridge

Gneiting T, Raftery AE, Westveld AH III, Goldman T, T (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Mon Wea Rev 133:1098–1118

He Y, Barnston AG (1996) Long-lead forecasts of seasonal precipitation in the tropical Pacific Islands Using CCA. J Climate 9:2020–2035

Hsu W-R, Murphy AH (1986) The attributes diagram A geometrical framework for assessing the quality of probability forecasts. Int J Forecast 2:285–293. https://doi.org/10.1016/0169-2070(86)90048-8

Kirtman BP, Min D (2009) Multimodel ensemble ENSO prediction with CCSM and CFS. Mon Wea Rev 137:2908–2930

Kirtman BP, Pirani A (2009) The state of the art of seasonal prediction: outcomes and recommendations from the First World Climate Research Program Workshop on seasonal prediction. Bull Am Meteorol Soc 90:455–458

Landsea CW, Knaff JA (2000) How much skill was there in forecasting the very strong 1997–98 El Niño? Bull Am Meteorol Soc 81:2107–2120

Madigan D, Raftery AE (1994) Model selection and accounting for model uncertainty in graphical models using OCCAM's window. J Am Stat Assoc 89:1535–1546

McAvaney BJ, Coauthors (2001) Model evaluation. climate change 2001: the scientific basis. In: Houghton JT (ed) Cambridge Univ. Press, Cambridge, pp 471–523

McPhaden MJ, Glantz MH (2006) ENSO as an integrating concept in earth science. Science 314:1740–1745

Min SK, Simonis D, Hense A (2007) Probabilistic climate change predictions applying Bayesian model averaging. Philos Trans R Soc Lond: Math Phys Eng Sci 365:2103–2116

Peng P, Kumar A, van den Dool H, Barnston AG (2002) An analysis of multimodel ensemble predictions for seasonal climate anomalies. J Geophys Res 107:D23, 4710. https://doi.org/10.1029/2002JD002712

Raftery AE, Gneiting T, Balabdaoui TF, Polakowski M (2005) Using Bayesian model averaging to calibrate forecast ensembles. Mon Wea Rev 133:1155–1174

Rasmusson EM, Wallace JM (1983) Meteorological aspects of the El Niño/Southern oscillation. Science 222:1195–1202. https://doi.org/10.1126/science.222.4629.1195

Rings J, Vrugt JA, Schoups G, Huisman JA, Vereecken H (2012) Bayesian model averaging using particle filtering and Gaussian mixture modeling: theory, concepts, and simulation experiments. Water Resour Res 48(5)

Saha S et al. (2014) The NCEP climate forecast system version 2. J Clim 27(6):2185–2208

Sarachik ES, Cane MA (2010) The El Niño-Southern oscillation phenomenon. Cambridge Univ. Press, p 384

Schepen A, Wang QJ, Robertson DE (2014) Combining the strengths of statistical and dynamical modeling approaches for forecasting Australian seasonal rainfall. J Geophys Res (Atmospheres), 117(D20)

Sloughter JM, Gneiting T, Raftery AE (2010) Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. J Amer Stat Assoc 105:25–35

Tebaldi C, Smith RL, Nychka D, Mearns LO (2005) Quantifying uncertainty in projections of regional climate change: a Bayesian approach to the analysis of multimodel ensembles. J Climate 18:1524–1540

Tian X, Xie Z, Wang A, Yang X (2012) A new approach for Bayesian model averaging. Sci China Earth Sciences 55:1336–1344

Trenberth KE (1998) Development and forecasts of the 1997/98 El Niño: CLIVAR scientific issues. CLIVAR Exchanges 3:4–14

Van den Dool HM (1994) Searching for analogues, how long must we wait? Tellus, 46A, 314–324

Vrugt JA, Robinson BA (2007) Treatment of uncertainty using ensemble methods: comparison of sequential data assimilation and Bayesian model averaging. Water Resour Res 43:W01411. https://doi.org/10.1029/2005WR004838

Wang QJ, Schepen A, Robertson DE (2012) Merging seasonal rainfall forecasts from multiple statistical models through Bayesian model averaging. J Climate 25:5524–5537

Wilks DS (2011) Statistical methods in the atmospheric sciences. Academic Press, 676 pp 334–340

Xue Y, Leetmaa A (2000) Forecasts of tropical Pacific SST and sea level using a Markov model. Geophys Res Lett 27:2701–2704

Yu Z-P, Chu P-S, Schroeder T (1997) Predictive skills of seasonal to annual rainfall variations in the U.S. affiliated Pacific islands: Canonical correlation analysis and multivariate principal component regression approaches. J Clim 10:2586–2509

Zhang W, Villarini G, Slater L, Vecchi GA, Bradley AA (2017) Improved ENSO forecasting using bayesian updating and the North American multimodel ensemble (NMME). J Climate 30:9007–9025