Communications for Statistical Applications and Methods
2017, Vol. 24, No. 1, 67–80

http://dx.doi.org/10.5351/CSAM.2017.24.1.067
Print ISSN 2287-7843 / Online ISSN 2383-4757

Leave-one-out Bayesian model averaging for probabilistic ensemble forecasting

Yongdai Kim[1,a], Woosung Kim[b], Ilsang Ohn[a], Young-Oh Kim[c]

[a]Department of Statistics, Seoul National University, Korea; [b]NAVER Corp., Korea;
[c]Department of Civil & Environmental Engineering, Seoul National University, Korea

Abstract

Over the last few decades, ensemble forecasts based on global climate models have become an important part of climate forecast due to the ability to reduce uncertainty in prediction. Moreover in ensemble forecast, assessing the prediction uncertainty is as important as estimating the optimal weights, and this is achieved through a probabilistic forecast which is based on the predictive distribution of future climate. The Bayesian model averaging has received much attention as a tool of probabilistic forecasting due to its simplicity and superior prediction. In this paper, we propose a new Bayesian model averaging method for probabilistic ensemble forecasting. The proposed method combines a deterministic ensemble forecast based on a multivariate regression approach with Bayesian model averaging. We demonstrate that the proposed method is better in prediction than the standard Bayesian model averaging approach by analyzing monthly average precipitations and temperatures for ten cities in Korea.

Keywords: Bayesian model averaging, climate forecast, ensemble forecast, global climate models, mixture models, multivariate regression

1. Introduction

Global climate models (GCMs) are computer models that generate meteorological variables under various emission scenarios. They have adequately explained past variations of climate and are now used to predict future climate. One of the important problems in using GCMs for the prediction or projection of future climate is that large uncertainties exist in GCMs and climate. For example, climate itself has a large variability that is hardly predictable, and GCMs are sensitive to the change of emission scenarios (Mearns *et al.*, 2001). Over the last few decades, ensemble forecasts based on GCMs have become an important part of climate forecast due to the ability to reduce uncertainty in prediction.

There are various methods to combine different GCMs to forecast future climate. The simplest approach is to assign the equal weights to GCMs and take the simple average (Lambert and Boer, 2001; Sperber *et al.*, 2004). A better approach would be to assign different weights to GCMs based on individual capabilities measured on past and future climate data (Dessai *et al.*, 2005; Giorgi and Mearns, 2002). A more sophisticated approach uses a multivariate linear regression model where projected values simulated by GCMs are treated as covariates and observed data as responses. Regression coefficients obtained from the regression model are used to assign weights to GCMs. In general, ensemble forecasts based on multivariate linear regression approaches often outperform other ensemble

[1] Corresponding author: Department of Statistics, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul 08826, Korea. E-mail: ydkim903@snu.ac.kr

Published 31 January 2017 / journal homepage: http://csam.or.kr

approaches and have been studied extensively (Gneiting *et al.*, 2005; Kharin and Zwiers, 2002; Krishnamurti *et al.*, 1999, 2000; Unger *et al.*, 2009).

In climate forecasting, assessing the prediction uncertainty is as important as estimating the optimal weights, and this is achieved through a probabilistic forecast which is based on the predictive distribution of future climate. A standard approach assumes that the predictive distribution belongs to a given parametric family and estimates parameters accordingly while the mean is fixed at the projected value by a deterministic ensemble forecast (Gneiting *et al.*, 2005). However, this standard approach would be suboptimal when the predictive distribution is not close to a given parametric family. An alternative approach is Bayesian model averaging (BMA) that combines predictive distributions of GCMs instead of combining projected values. Many empirical studies including Raftery *et al.* (2005) and Sloughter *et al.* (2007) have shown that various BMA approaches outperform other competitors in probabilistic forecasting.

BMA approaches work well; however, it is unclear how BMA approaches are related to Bayesian principles. In this paper, we propose a new method to estimate predictive distribution based on the BMA approach, which can be understood as model averaging with respect to approximated posterior distribution. Suppose there are K many GCMs. In the proposed method, we first make the K many sets of GCMs, each of which consists of $(K-1)$ many GCMs obtained by deleting a GCM. For each set of GCMs, we construct a predictive distribution based on a standard multivariate regression approach with the Gaussian assumption. Finally, we combine the K many predictive distributions constructed using the BMA approach. We call the proposed method the *leave-one-out* (LOO) BMA since the K many predictive distributions combined by the BMA are constructed based on K many sets of GCMs obtained by deleting a GCM. By analyzing real data, we demonstrate that the LOO BMA approach outperforms the standard BMA approach in climate forecasting.

The LOO is popularly used to estimate the prediction error (Krzanowski and Hand, 1997). In particular, the leave-one-out cross-validation error is an (nearly) unbiased estimator of the prediction error (Efron, 1983). However, application of the leave-one-out approach for estimation of the predictive distribution is new.

The paper is organized as follows. In Section 2, we review various methods for ensemble forecasting. The estimation of the predictive distribution based on the LOO BMA approach is explained in Section 3. Results from analyzing monthly average precipitations and temperatures collected in 10 cities of Korea are presented in Section 4. In Section 5, another modification of the standard BMA approach, called the perturbed BMA, is proposed and is compared with LOO BMA. Concluding remarks follow in Section 5.

2. Review of ensemble methods

2.1. Deterministic ensemble forecasting

The simplest method for deterministic ensemble forecasting uses the simple average of GCMs that gives equal weights to all GCMs. A better method would be to assign different weights to GCMs based on their individual ability. Reliability ensemble averaging (Giorgi and Mearns, 2002) and regional skill score Dessai *et al.* (2005) are two such methods that both decide the weights based on the two abilities of performance and convergence. The performance of a GCM is a quantity proportional to the difference between the GCM forecasts and observations in past data. Convergence is measured by the difference between the GCM forecasts and the average of the forecasts made by multiple GCMs in future data. In general, GCMs with smaller performance and convergence receive higher weights.

Regression approaches have recently received significant attention. Let $y_1, \ldots, y_T$ be past obser-

vations of a quantity of interest we want to forecast (e.g. precipitation), and let $\mathbf{f}_1, \ldots, \mathbf{f}_T$ be the corresponding projections of K many GCMs, where $\mathbf{f}_t = (f_{t1}, \ldots, f_{tK})'$. Regression approaches assume that

$$y_t = \beta_0 + \sum_{k=1}^{K} \beta_k f_{tk} + \epsilon_t,$$

where ϵ_t are errors with mean 0 and variance σ^2. Here, β_0 is a term for bias correction and $\beta_k, k = 1, \ldots, K$ are considered to be the weights. A standard method to estimate the regression coefficients $\beta_0, \ldots, \beta_K$ is to use the least square estimator that minimizes the sum of squared residuals

$$\sum_{t=1}^{T} \left(y_t - \beta_0 - \sum_{k=1}^{K} \beta_k f_{tk} \right)^2 .$$

See, for example, Kharin and Zwiers (2002); Krishnamurti *et al.* (1999, 2000). Alternative methods are the maximal entropy estimator by Laurent and Cai (2007), minimum Continuous Rank Probability Score (CRPS) estimator by Gneiting *et al.* (2005) and Bayesian methods by Greene *et al.* (2006) and Min and Hense (2006, 2007). Nonlinear models such as neural network (Maqsood *et al.*, 2004) have also been used.

2.2. Probabilistic ensemble forecasting

A standard approach for probabilistic forecasting assumes that

$$y_t = \hat{y}_t + \epsilon_t,$$

where $\hat{y}_t$'s are the projected values from a deterministic ensemble forecast and ϵ_t's are assumed to be independent random variables with mean 0. When $\hat{y}_t$ are obtained by the regression approach, we typically assume that ϵ_t follows a Gaussian distribution with mean 0 and variance σ^2, and estimate σ^2 by the mean squared error. Gneiting *et al.* (2005) considered an additional regression model for σ^2. They assumed that y_t follows a Gaussian distribution with mean $\beta_0 + \sum_{k=1}^{K} \beta_k f_{tk}$ and variance $c + d s_t^2$, where s_t^2 is the variance of $f_{t1}, \ldots, f_{tK}$.

There is a way of using projections of multiple GCMs directly to estimate the predictive distribution without deterministic forecasting. Raftery *et al.* (2005) suggested the BMA in climate forecast, which assumes

$$p(y|\mathbf{f}) = \sum_{k=1}^{K} w_k g_k(y|f_k),$$

where g_k is a Gaussian distribution with the mean, $a_k + b_k f_k$ and variance, σ^2. They used the expectation-maximize (EM) algorithm to estimate the parameters. Duan *et al.* (2007) extended the BMA model by allowing unequal variances by assuming that g_k is a Gaussian distribution with mean μ_k and variance σ_k. Sloughter *et al.* (2007) used gamma distributions for g_k to forecast precipitation. See Section 3.1 for more discussions of the BMA approach.

2.3. Verification methods

In general, there are two kinds of the verification methods for climate forecast. The first one, called the determinant verification, measures the distance between observations and forecasted values. The

second method, called the density verification, assesses the performance of the calibration of the predictive distribution. Here, we say that a forecast method is calibrated if a meteorological event with probability p occur with a proportion p on average.

The determinant verification typically uses the mean absolute error (MAE) and the root mean square error (RMSE). The MAE is defined as

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T}|y_t - \hat{y}_t|,$$

where $\hat{y}_t$ denotes a forecasted value in time t. Similarly, the RMSE is defined as

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}.$$

Two popular measures for the density verification are the CRPS and the ignorance score (IGN). The CRPS is defined as

$$\text{CRPS} = \frac{1}{T}\sum_{t=1}^{T}\int_{-\infty}^{\infty}\{P(y|\mathbf{f}_t) - I(y_t \le y)\}^2\, dy,$$

where $P(y|\mathbf{f}) = \int_{-\infty}^{y} p(u|\mathbf{f})du$ and $I(x < y) = 1$ if $x < y$, and $I(x < y) = 0$ otherwise. The IGN is the negative log-likelihood, given as

$$\text{IGN} = -\sum_{t=1}^{T}\log p(y_t|\mathbf{f}).$$

The smaller value means the better prediction for both the CRPS and IGN, and both scores are proper but the IGN is lack of robustness to outliers (Gneiting and Raftery, 2007).

Along with the CRPS and IGN, the probability integral transformation (PIT) is a graphical tool to check the degree of calibration (Gneiting *et al.*, 2007). The PIT value p_t is defined as $p_t = P(y_t|\mathbf{f}_t)$. When $P(y|\mathbf{f}_t)$ is the true predictive distribution function, it is known that p_t follows the uniform distribution. By comparing the histogram of p_t with the density of the uniform distribution, we can check graphically how well a given predictive distribution is calibrated.

3. Leave-one-out Bayesian model averaging

We begin this section to explain the three possible BMA approaches that motivate the proposed LOO BMA approach.

3.1. Three versions of the Bayesian model averaging

The BMA approach proposed by Raftery *et al.* (2005) assumes that

$$p(y|\mathbf{f},\theta) = \sum_{k=1}^{K} w_k g_k(y|f_k), \tag{3.1}$$

where g_k is a Gaussian distribution with the mean, $a_k + b_k f_k$ and variance, σ_k^2. The parameter vector θ in the model (3.1) consists of (a_k, b_k), σ_k^2 and w_k for $k = 1, \ldots, K$.

According to the way of estimating the parameters, we can think of three versions of the BMA approaches. The first one is to estimate all of the parameters simultaneously by maximizing the likelihood of the BMA model (3.1):

$$L(\theta) = \prod_{t=1}^{T} p\left(y_t|\mathbf{f}_t, \theta\right). \tag{3.2}$$

For the second BMA approach, we first estimate (a_k, b_k) by the least square estimates of the univariate regression model for (y_t, f_{tk}). That is, we estimate (a_k, b_k) by $(\hat{a}_k, \hat{b}_k)$ which minimizes $\sum_{t=1}^{T}(y_t - a_k - b_k f_{tk})^2$. Then, we estimate σ_k and w_k for $k = 1, \ldots, K$ by maximizing the likelihood (3.2) of the BMA model with (a_k, b_k) being fixed at $(\hat{a}_k, \hat{b}_k)$. In fact, this is the one Raftery *et al.* (2005) proposed, and we call it the standard Bayesian model averaging.

By extending the second BMA approach, we can think of the following third BMA approach. We estimate (a_k, b_k) and σ_k^2 by

$$\left(\hat{a}_k, \hat{b}_k\right) = \text{argmin}_{a,b} \sum_{t=1}^{T} (y_t - a - b f_{tk})^2$$

and

$$\hat{\sigma}_k^2 = \sum_{t=1}^{T} \frac{\left(y_t - \hat{a}_k - \hat{b}_k f_{tk}\right)^2}{T-1}.$$

Then, we estimate w_k for $k = 1, \ldots, K$ by maximizing the likelihood (3.2) of the BMA model with (a_k, b_k) and σ_k^2 being fixed at their estimates.

Table 1 compares the predictive performance of the three BMA approaches on the data sets which are explained in Section 4. The second BMA approach outperforms the other two BMA approaches in 9 out of 10 cities. It seems that the first BMA approach overfits predictive distribution while the third BMA approach underfits. These results suggest that the key to success of the BMA approach is to estimate σ_k^2 and w_k using the BMA likelihood (3.2), while the mean functions are of secondary importance. The proposed LOO BMA in this paper is devised to improve the standard BMA approach by estimating the mean functions in the Gaussian distributions differently.

3.2. The proposed Bayesian model averaging

The proposed BMA approach assumes

$$p^{\text{LOO}}\left(y|\mathbf{f}, \theta\right) = \sum_{k=1}^{K} w_k g_k\left(y|\mathbf{f}_{(-k)}\right), \tag{3.3}$$

where $\mathbf{f}_{(-k)} = (f_l, l \neq k)$ and $g_k(y|\mathbf{f}_{(-k)})$ are Gaussian distribution with mean

$$\beta_0^{(k)} + \sum_{l \neq k} \beta_l^{(k)} f_l$$

and σ_k^2. Here, θ consists of $\beta_0^{(k)}, (\beta_l^{(k)}, l \neq k), \sigma_k^2$ and w_k for $k = 1, \ldots, K$. Note that the mean of g_k depends on $\mathbf{f}_{(-k)}$ in the proposed BMA model while it depends on f_k in the standard BMA model. We

Table 1: Comparison of the three BMA approaches

Area	Method	IGN	CRPS	Area	Method	IGN	CRPS
Seoul	BMA-1	185.604	60.801	Imsil	BMA-1	152.443	48.317
	BMA-2	149.540	46.027		BMA-2	140.816	43.169
	BMA-3	153.418	47.330		BMA-3	143.409	44.326
Incheon	BMA-1	190.280	61.478	Jeonju	BMA-1	163.524	52.399
	BMA-2	161.526	52.672		BMA-2	137.820	42.579
	BMA-3	168.517	54.250		BMA-3	140.664	43.291
Daejeon	BMA-1	170.998	53.654	Gwangju	BMA-1	160.633	49.950
	BMA-2	145.053	44.796		BMA-2	138.637	43.211
	BMA-3	143.863	44.473		BMA-3	141.122	44.101
Daegu	BMA-1	185.445	65.578	Chuncheon	BMA-1	189.722	60.283
	BMA-2	176.611	62.869		BMA-2	152.232	47.588
	BMA-3	196.811	61.819		BMA-3	153.045	48.357
Busan	BMA-1	165.041	59.475	Gangneung	BMA-1	178.401	67.991
	BMA-2	170.095	59.629		BMA-2	185.195	67.551
	BMA-3	189.763	59.353		BMA-3	214.174	66.916

IGN = ignorance score; CRPS = continuous ranked probability score; BMA = Bayesian model averaging.

name the proposed model the LOO BMA model since the mean of each component (i.e. g_k) depends on all of the GCM projections except one (i.e. $\mathbf{f}_{(-k)}$).

To estimate the parameters, we use the method similar to the standard BMA approach. The regression coefficients $\beta_0^{(k)}$ and $\beta_l^{(k)}, l \neq k$ are estimated by $\hat{\beta}_0^{(k)}$ and $\hat{\beta}_l^{(k)}, l \neq k$ which minimize

$$\sum_{t=1}^{T} \left(y_t - \beta_0^{(k)} - \sum_{l \neq k} \beta_l^{(k)} f_{tl} \right)^2 .$$

Then, w_k and σ_k^2 for $k = 1, \ldots, K$ are estimated by maximizing the likelihood of the LOO BMA model (3.3): $L(\theta) = \prod_{t=1}^{T} p^{\mathrm{LOO}}(y_t | \mathbf{f}_t, \theta)$. The maximum likelihood estimators of w_k and σ_k^2 are calculated easily by the following EM algorithm.

Suppose δ_t are independent multinomial random vectors with the cell probabilities $w_1, \ldots, w_k$. Then, the complete log-likelihood of the LOO BMA likelihood is given by

$$l_{\mathrm{comp}} = \sum_{t=1}^{T} \left[\sum_{k=1}^{K} I(\delta_t = k) \left\{ -\frac{1}{2} \log \sigma_k^2 - \frac{\left(y_t - \hat{\mu}_k \left(\mathbf{f}_{(-k)} \right) \right)^2}{2\sigma_k^2} + \log w_k \right\} \right],$$

where

$$\hat{\mu}_k \left(\mathbf{f}_{(-k)} \right) = \hat{\beta}_0^{(k)} + \sum_{l \neq k} \hat{\beta}_l^{(k)} f_{tl}.$$

The E-step is to calculate $v_{kt} = \mathrm{E}(I(\delta_t = k) | data, \eta^c)$, where η^c is the current estimate of $\eta = (\sigma_k^2, w_k, k = 1, \ldots, K)$. It turns out that

$$v_{kt} = \frac{w_k^c \phi \left(y_t | \hat{\mu}_k \left(\mathbf{f}_{(-k)} \right), \left(\sigma_k^2 \right)^c \right)}{\sum_{l=1}^{K} w_l^c \phi \left(y_t | \hat{\mu}_k \left(\mathbf{f}_{(-k)} \right), \left(\sigma_l^2 \right)^c \right)},$$

where $\phi(y|\mu, \sigma^2)$ is the density function of the Gaussian distribution with mean μ and variance σ^2. The M-step updates σ_k^2 and w_k by maximizing the expected complete log-likelihood given by

$$E\left(l_{\text{comp}}|\text{data}, \eta^c\right) = \sum_{t=1}^{T}\left[\sum_{k=1}^{K} v_{kt}\left\{-\frac{1}{2}\log\sigma_k^2 - \frac{\left(y_t - \hat{\mu}_k\left(\mathbf{f}_{(-k)}\right)\right)^2}{2\sigma_k^2} + \log w_k\right\}\right]$$

The $\hat{\sigma}_k^2$ and $\hat{w}_k$ maximizing $E(l_{\text{comp}}|\text{data}, \eta^c)$ are

$$\hat{\sigma}_k^2 = \frac{\sum_{t=1}^{T} v_{kt}\left(y_t - \hat{\mu}_k\left(\mathbf{f}_{(-k)}\right)\right)^2}{\sum_{t=1}^{T} v_{kt}}$$

and

$$\hat{w}_k = \frac{\sum_{t=1}^{T} v_{kt}}{\sum_{l=1}^{K}\sum_{t=1}^{T} v_{lt}}.$$

We repeat the E and M steps until convergence.

Remark 1. The predictive distribution (3.3) of the LOO BMA model can be rewritten as

$$p^{LOO}(y|\mathbf{f}, \theta) = \int_{\eta} g(y|\mathbf{f}, \eta)\,\pi(d\eta), \tag{3.4}$$

where $\eta = (\beta_0, \beta, \sigma^2)$ and π has masses $\hat{w}_k$ at $\eta_k = (\hat{\beta}_0^{(k)}, (\hat{\beta}_l^{(k)}, l \neq k), \hat{\sigma}_k^2)$. We can consider $\pi(\eta)$ as a proxi of the posterior distribution of η, which implies that the LOO BMA model can be understood as a proxy of the Bayesian predictive distribution. That is, we approximate the posterior distribution by the LOO distribution (i.e. the jack-knife distribution). Approximating the posterior distribution by the jack-knife or bootstrap are well known (Efron, 2012; Simmons *et al.*, 2004). This is the main motivation of the LOO BMA model. In Section 5, we consider another proxy of the Bayesian predictive distribution using random perturbation and show that the performances of the two proxies are similar, which confirms that our interpretation of the LOO BMA as a Bayesian predictive distribution makes sense.

4. Numerical studies

4.1. Description of data

We analyze the monthly averages of precipitations and temperatures collected at 10 cities in Korea: Seoul, Incheon, Daejeon, Daegu, Busan, Jeonju, Imsil, Gwangju, Chuncheon and Gangneung. Figure 1 presents the locations of the 10 cities on a map of Korea. The data set consists of 444 many monthly average precipitation values collected from Jan 1973 to Dec 2009. The record lengths at all the sites are identical. For each month, there are 5 GCM projections of the monthly average precipitation and temperature. The 5 GCMs used in the analysis are given in Table 2. Data from Jan 1973 to Dec 1999 are used as the training data set and data from Jan 2000 to Dec 2009 as the test data set. For precipitation, we apply the log transformation to the data before estimating the predictive distribution.

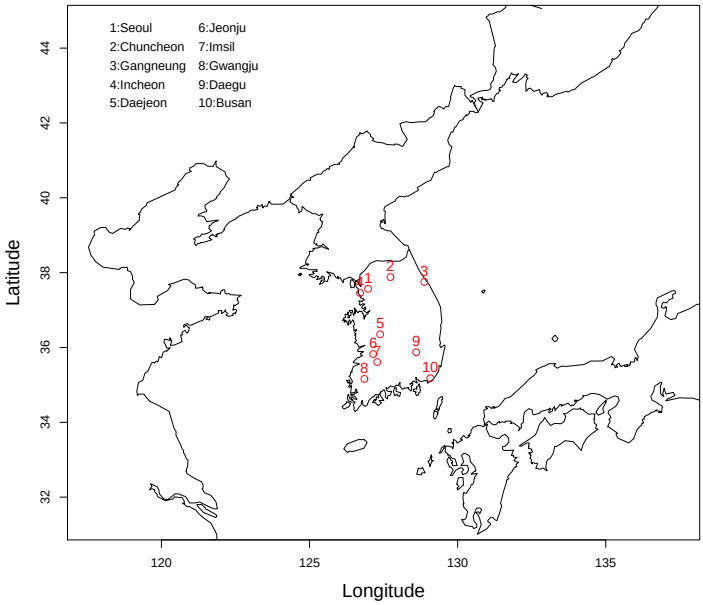

Figure 1: *Map of 10 cities in Korea.*

Table 2: Five global climate models used in the analysis

Abbreviation	Model (agency: version)	Country
CSR	CSIRO: MK3	Australia
GF1	GFDL: CM2_1	USA
MIU	CONS: ECHO-G	Germany/Korea
MRI	MRI: CGCM2_3_2	Japan
UKC	UKMO: HADGEM1	UK

4.2. Results for monthly average precipitations

The performance of the proposed LOO BMA (L-BMA) approach for the density verification is investigated. In particular, we compare the L-BMA approach with the standard multivariate linear regression model with Gaussian error (GA) and the standard BMA (S-BMA) approach.

Table 3 compares the two measures of the density verification for the three models: GA, S-BMA and L-BMA. In general, the GA is the worst and the L-BMA is the best. The L-BMA also outperforms the S-BMA for 8 out of 10 cities.

Figure 2 draws the coverage probabilities and interval lengths of the 50% and 90% predictive intervals of the GA, S-BMA and L-BMA. The L-BMA achieves the nominal levels reasonably and generally gives shorter interval lengths, which means that the L-BMA achieves better sharpness in calibration.

Figure 3 provides the marginal predictive density functions of the three approaches at Seoul. The predictive distributions of the S-BMA and L-BMA, which are very similar, have heavier tails than that

Table 3: Comparison of the predictive performance

Area	Method	IGN	CRPS	Area	Method	IGN	CRPS
Seoul	GA	152.712	47.822	Imsil	GA	141.684	43.876
	S-BMA	149.540	46.027		S-BMA	140.816	43.169
	L-BMA	147.735	46.177		L-BMA	138.391	42.880
Incheon	GA	168.238	54.049	Jeonju	GA	139.383	42.261
	S-BMA	161.526	52.672		S-BMA	137.820	42.579
	L-BMA	164.212	52.935		L-BMA	135.490	41.788
Daejeon	GA	139.027	43.224	Gwangju	GA	146.350	42.617
	S-BMA	145.053	44.796		S-BMA	138.637	43.211
	L-BMA	141.938	43.666		L-BMA	135.850	42.382
Daegu	GA	195.519	61.625	Chuncheon	GA	154.696	47.712
	S-BMA	176.611	62.869		S-BMA	152.232	47.588
	L-BMA	176.501	62.084		L-BMA	148.523	46.960
Busan	GA	192.645	59.618	Gangneung	GA	211.997	66.517
	S-BMA	170.095	59.629		S-BMA	185.195	67.551
	L-BMA	170.535	58.643		L-BMA	181.746	65.784

IGN = ignorance score; CRPS = continuous ranked probability score; GA= Gaussian error; L-BMA = leave-one-out (LOO) BMA; S-BMA = standard BMA.

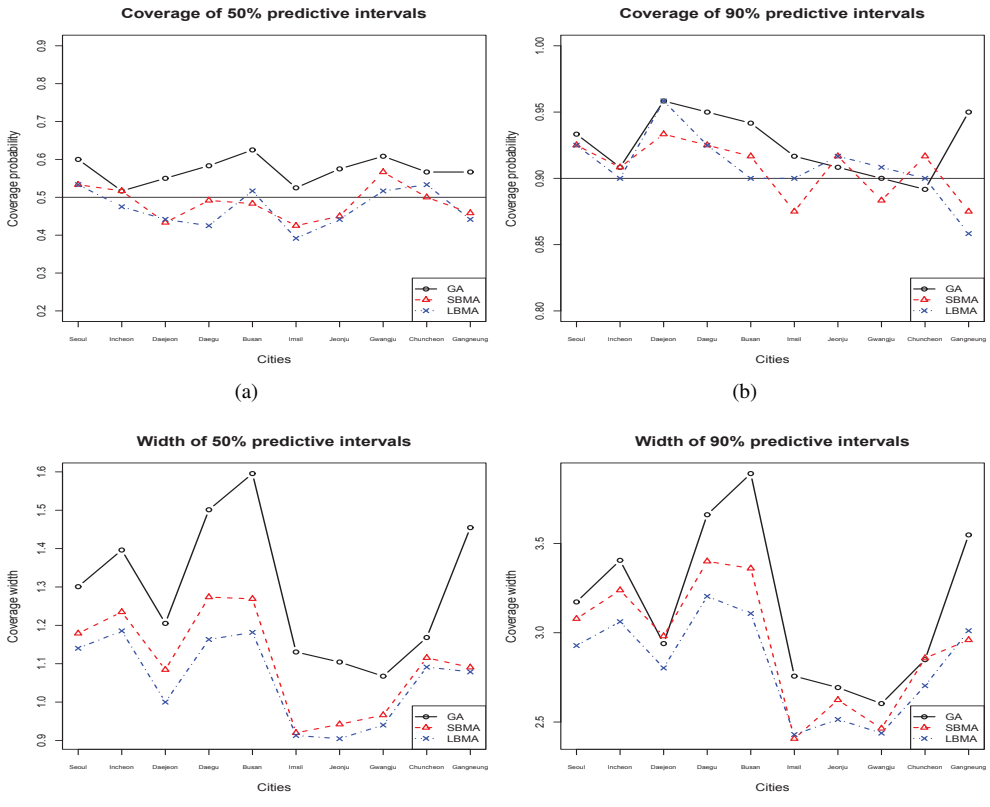

Figure 2: *Coverage probabilities of (a) 50%, (b) 90% predictive intervals and lengths of (c) 50%, (d) 90% predictive intervals.*

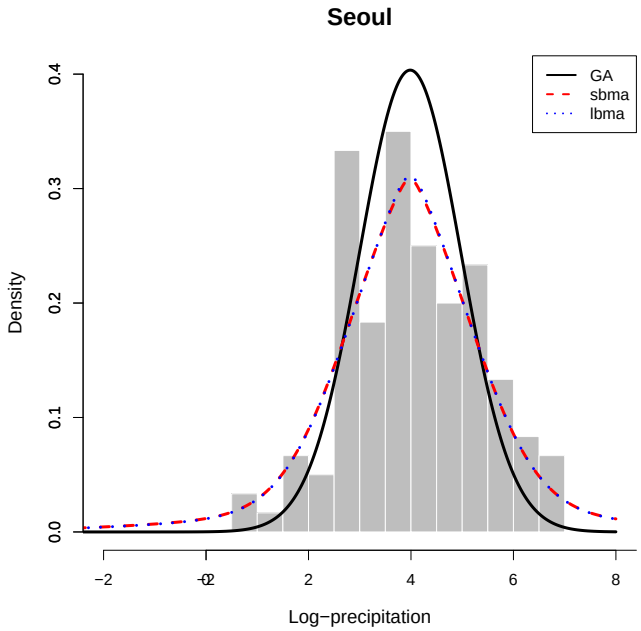

Figure 3: *Histogram of the monthly average precipitation of Seoul with the marginal predictive densities based on the three methods.*

of the GA. The distribution of monthly precipitation has a heavier tails than the Gaussian distribution, and the BMA approaches capture this characteristic successfully. The marginal predictive densities of the other cities are similar.

Figure 4 draws the time series plot of the test data set of Seoul with 90% predictive intervals estimated by the L-BMA approach. The predictive intervals cover future precipitations well.

4.3. Prediction performance for forecasting monthly average temperatures

We applied the L-BMA approach to monthly mean temperatures of the 10 cities in Korea to get Table 4. Similarly to precipitation, the L-BMA outperforms the GA and S-BMA in predicting temperature.

5. Perturbed BMA

The L-BMA consists of the two steps. The first step is to estimate the mean functions of the Gaussian components by the LOO method, and the second step is to estimate the weights and variances by the maximum likelihood estimates. We can modify the L-BMA by estimating mean functions by perturbing data as follows. We assume that

$$p^{\text{pert}}(y|\mathbf{f},\theta) = \sum_{m=1}^{M} w_m g_m(y|\mathbf{f}) \tag{5.1}$$

for some $M > 0$, where $g_m(y|\mathbf{f})$ are Gaussian distributions with mean $\mu_m(\mathbf{f})$ and variance σ_m^2. For each $m = 1,\dots,M$, we generate perturbed outputs $\tilde{y}_t^{(m)}, t = 1,\dots,T$, by $\tilde{y}_t^{(m)} = y_t + \epsilon_t^{(m)}$, where $\epsilon_t^{(m)}$ are

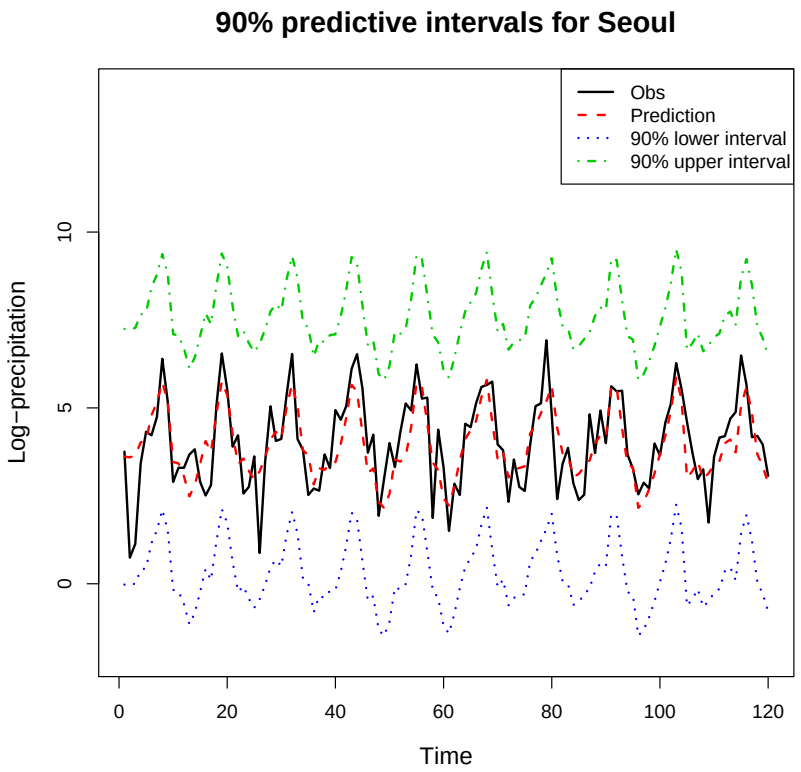

Figure 4: *Time series plot of the monthly average precipitation with the 90% predictive intervals of the LOO BMA approach for Seoul.*

independent Gaussian random variables with mean 0 and variance σ_ϵ^2. We estimate the mean function $\mu_m(\mathbf{f})$ by $\hat{\beta}_0^{(m)} + \sum_{k=1}^{K} \hat{\beta}_k^{(m)} f_k$, where $\hat{\beta}_0^{(m)}$ and $\hat{\beta}_k^{(m)}, k = 1, \ldots, K$ are obtained by minimizing

$$\sum_{t=1}^{T} \left(\tilde{y}_t^{(m)} - \beta_0 - \sum_{k=1}^{K} \beta_k f_{tk} \right)^2 .$$

Finally, the weights w_m and variances σ_m^2 are estimated by maximum likelihood estimates. We call this method the perturbed BMA. In the perturbed BMA, the number of components M does not have to be equal to the number of GCMs K. The perturbed BMA is motivated by the ensemble algorithm called random forest of Breiman (2001) for regression and classification. While random forest makes multiple models by selecting the parameters randomly, the perturbed BMA generates multiple models by injecting randomness into data. However, the estimated parameter obtained with data injected by random noises can be considered as random and the perturbed BMA is therefore similar to random forest.

Table 5 compares the L-BMA and perturbed BMA. For the perturbed BMA, we set $M = 10$ and

Table 4: Comparison of the predictive performance of mean temperatures

Area	Method	IGN	CRPS	Area	Method	IGN	CRPS
Seoul	GA	222.745	42.102	Imsil	GA	216.394	34.811
	S-BMA	210.310	41.847		S-BMA	202.345	36.254
	L-BMA	202.268	41.895		L-BMA	196.960	34.679
Incheon	GA	213.975	38.788	Jeonju	GA	208.064	43.317
	S-BMA	215.759	41.356		S-BMA	199.909	45.669
	L-BMA	202.495	39.579		L-BMA	191.882	43.238
Daejeon	GA	214.478	41.814	Gwangju	GA	201.585	43.134
	S-BMA	204.837	41.325		S-BMA	195.078	45.080
	L-BMA	199.399	41.638		L-BMA	187.871	43.126
Daegu	GA	214.542	51.784	Chuncheon	GA	214.676	29.708
	S-BMA	208.229	51.970		S-BMA	199.277	31.535
	L-BMA	203.435	52.293		L-BMA	191.630	30.161
Busan	GA	210.020	45.179	Gangneung	GA	225.769	49.197
	S-BMA	199.833	47.653		S-BMA	209.332	49.196
	L-BMA	203.348	48.036		L-BMA	206.253	49.085

IGN = ignorance score; CRPS = continuous ranked probability score; GA= Gaussian error; L-BMA = leave-one-out (LOO) BMA; S-BMA = standard BMA.

Table 5: Comparison of the predictive performance of the LOO BMA (L-BMA) and perturbed BMA (P-BMA) on monthly average precipitation

Area	Method	IGN	CRPS	Area	Method	IGN	CRPS
Seoul	L-BMA	147.735	46.177	Imsil	L-BMA	138.391	42.880
	P-BMA	146.748	45.867		P-BMA	142.363	43.847
Incheon	L-BMA	164.212	52.935	Jeonju	L-BMA	135.490	41.788
	P-BMA	161.388	52.530		P-BMA	137.479	42.166
Daejeon	L-BMA	141.938	43.666	Gwangju	L-BMA	135.850	42.382
	P-BMA	141.155	43.329		P-BMA	133.661	41.853
Daegu	L-BMA	176.501	62.084	Chuncheon	L-BMA	148.523	46.960
	P-BMA	172.259	60.620		P-BMA	149.274	47.226
Busan	L-BMA	170.535	58.643	Gangneung	L-BMA	181.746	65.784
	P-BMA	165.380	57.516		P-BMA	185.168	66.474

IGN = ignorance score; CRPS = continuous ranked probability score; L-BMA = leave-one-out (LOO) BMA; P-BMA = perturbed BMA.

$\sigma_\epsilon^2 = 2\text{mse}$, where

$$\text{mse} = \frac{\sum_{t=1}^{T}\left(y_t - \hat{\beta}_0 - \sum_{k=1}^{K}\hat{\beta}_k f_{tk}\right)^2}{T-K-1}$$

and $\hat{\beta}_0$ and $\hat{\beta}_k, k = 1, \ldots, K$ are the least square estimates based on $(y_t, \mathbf{f}_t), t = 1 \ldots, T$. The prediction performance of the perturbed BMA for other values of M and σ_ϵ^2 are similar unless these two values are too large or too small. We can see from Table 5 that the two BMA approaches show similar prediction performance.

6. Concluding remarks

We proposed the two variations of the standard BMA approach: the LOO BMA and perturbed BMA. We showed empirically that the two proposed BMA approaches outperformed the standard BMA.

Even though we considered the Gaussian mixture model, the proposed BMA approaches can be applied to other mixture distributions as long as the mean of each component is assumed to be linear with respect to GCM projections. For example, instead of the mixture of log-normal distributions, we may consider the BMA approach with gamma distributions for monthly average precipitation; therefore, the corresponding LOO and perturbed BMA approaches can be developed without significant difficulty.

In Section 5, we showed that there is an analogy between the BMA approaches and ensemble methods to generate predictive models for regression and classification. Breiman (2001) explained the success of ensemble methods by the trade-off between strength and diversity. We could explain the success of the BMA to estimate the predictive distribution similarly, which we leave for future work.

Acknowledgement

This research was supported by a grant (14AWMP-B082564-01) from the Advanced Water Management Research Program funded by Ministry of Land, Infrastructure and Transport of Korean government.

References

Breiman L (2001). Random forests, *Machine Learning*, **45**, 5–32.

Dessai S, Lu X, and Hulme M (2005). Limited sensitivity analysis of regional climate change probabilities for the 21st century, *Journal of Geophysical Research*, **110**, D19108.

Duan Q, Ajami NK, Gao X, and Sorooshian S (2007). Multi-model ensemble hydrologic prediction using Bayesian model averaging, *Advances in Water Resources*, **30**, 1371–1386.

Efron B(1983). Estimating the error rate of a prediction rule: improvement on cross-validation, *Journal of the American Statistical Association*, **78**, 316–331.

Efron B (2012). Bayesian inference and the parametric bootstrap, *The Annals of Applied Statistics*, **6**, 1971–1997.

Giorgi F and Mearns LO (2002). Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the" reliability ensemble averaging" (REA) method, *Journal of Climate*, **15**, 1141–1158.

Gneiting T, Balabdaoui F, and Raftery AE (2007). Probabilistic forecasts, calibration and sharpness, *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, **69**, 243–268.

Gneiting T and Raftery AE (2007). Strictly proper scoring rules, prediction, and estimation, *Journal of the American Statistical Association*, **102**, 359–378.

Gneiting T, Raftery AE, Westveld III AH, and Goldman T (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, *Monthly Weather Review*, **133**, 1098–1118.

Greene AM, Goddard L, and Lall U (2006). Probabilistic multimodel regional temperature change projections, *Journal of Climate*, **19**, 4326–4343.

Kharin VV and Zwiers FW (2002). Climate predictions with multimodel ensembles, *Journal of Climate*, **15**, 793–799.

Krishnamurti TN, Kishtawal CM, LaRow TE, Bachiochi DR, Zhang Z, Williford CE, Gadgil S, and Surendran S (1999). Improved weather and seasonal climate forecasts from multimodel superensemble, *Science*, **285**, 1548–1550.

Krishnamurti TN, Kishtawal CM, Zhang Z, LaRow T, Bachiochi D, Williford E, Gadgil S, and Suren-

dran S (2000). Multimodel ensemble forecasts for weather and seasonal climate, *Journal of Climate*, **13**, 4196–4216.

Krzanowski WJ and Hand DJ (1997). Assessing error rate estimators: the leave-one-out method reconsidered, *Australian & New Zealand Journal of Statistics*, **39**, 35–46.

Lambert SJ and Boer GJ (2001). CMIP1 evaluation and intercomparison of coupled climate models, *Climate Dynamics*, **17**, 83–106.

Laurent R and Cai X (2007). A maximum entropy method for combining AOGCMs for regional intra-year climate change assessment, *Climatic Change*, **82**, 411–435.

Maqsood I, Khan MR, and Abraham A (2004). An ensemble of neural networks for weather forecasting, *Neural Computing & Applications*, **13**, 112–122.

Mearns LO, Hulme M, Carter TR, Leemans R, Lal M, and Whetton P (2001). Climate scenario development. In Houghton JT et al. (Eds), *Climate Change 2001: The Scientific Basis* (pp. 739–768), Cambridge University Press, Cambridge.

Min SK and Hense A (2006). A Bayesian assessment of climate change using multimodel ensembles. Part I: Global mean surface temperature, *Journal of Climate*, **19**, 3237–3256.

Min SK and Hense A (2007). A Bayesian assessment of climate change using multimodel ensembles. Part II: Regional and seasonal mean surface temperatures, *Journal of Climate*, **20**, 2769–2790.

Raftery AE, Gneiting T, Balabdaoui F, and Polakowski M (2005). Using Bayesian model averaging to calibrate forecast ensembles, *Monthly Weather Review*, **133**, 1155–1174.

Simmons MP, Pickett KM, and Miya M (2004). How meaningful are Bayesian support values?, *Molecular Biology and Evolution*, **21**, 188–199.

Sloughter JM, Raftery AE, Gneiting T, and Fraley C (2007). Probabilistic quantitative precipitation forecasting using Bayesian model averaging, *Monthly Weather Review*, **135**, 3209–3220.

Sperber K, Gleckler P, Covey C, Taylor K, Bader D, Phillips T, Fiorino M, and AchutaRao K (2004). *An Appraisal of Coupled Climate Model Simulations*, Lawrence Livermore National Laboratory, Livermore, CA.

Unger DA, Van Den Dool H, O'Lenic E, and Collins D (2009). Ensemble regression, *Monthly Weather Review*, **137**, 2365–2379.

Received October 10, 2016; Revised January 9, 2017; Accepted January 9, 2017