Configuration of Bayesian Model Averaging Training Window to Improve Seasonal Rainfall Ensemble Prediction

R Muharsyah\textsuperscript{1,2}, T W Hadi\textsuperscript{1}, S W Indratno\textsuperscript{3}

\textsuperscript{1}Atmospheric Sciences Research Group, Faculty of Earth Sciences and Technology, Institute Technology Bandung, Bandung, 40132, Indonesia
\textsuperscript{2}Agency for Meteorology, Climatology and Geophysics of Republic of Indonesia (BMKG), Jakarta, 107720, Indonesia
\textsuperscript{3}Mathematics, Faculty of Mathematics and Natural Sciences, Institute Technology Bandung, Bandung, 40132, Indonesia

*Email: robi.muharsyah@gmail.com

Abstract. Bayesian Model Averaging (BMA) is a statistical post-processing method for producing probabilistic forecasts from ensemble prediction in the form of predictive Probability Density Function (PDF). It is known that BMA is able to improve the reliability of the probabilistic forecast of short- and medium-range rainfall forecasts. This study aims to develop the application of BMA to calibrate long-range forecast in order to improve the quality of the seasonal forecast in Indonesia. The seasonal forecast that has been used is monthly rainfall from the output of the ensemble prediction European Center for Medium-Range Weather Forecasts (ECMWF) system 4 model (ECS4). This model was calibrated against observational data at 26 stations of Agency for Meteorology, Climatology and Geophysics of Republic of Indonesia (BMKG) over Java Island in 1981 – 2018. BMA predictive PDF was generated with a Gamma distribution which was obtained based on Training Window Sequential (TW\textsubscript{S}) and Conditional (TW\textsubscript{C}) training windows. Output of BMA-TWC was slightly better than BMA-TWS. Nevertheless, both of them were superior to raw model ECS4. BMA-TWC or BMA-TWS output was varying depend on spatial and temporal, but in general, the best result was in the dry season and during the El Nino phase. BMA was able to improve the distribution characteristics of ensemble prediction. BMA also increased the skill, resolution, and reliability of probability forecast of Below Normal (BN) and Above Normal (AN). Furthermore, the reliability of BN and AN of BMA output were also have the categories of “still very useful” and “perfect” compared to raw model ECS4 that were in the “dangerous” and “not useful” categories. The reliability “still very useful” and “perfect” show that the probabilistic forecast of BN and AN event can be used for making decisions related to seasonal forecast especially over Java island.

1. Introduction

In modern climate modeling, the state of the art of seasonal forecast is probabilistic forecast (e.g. rainfall probabilistic forecast). The probabilistic forecast is useful for users if they can understand the uncertainty in that forecast. [1]. There are two sources of uncertainty; (i) the limitation of information related to the initial condition of the climate system, and (ii) the limitation of mathematical formulation in the model to represent all the processes in the atmosphere. The first sources of uncertainty can be...
overcome by making many individual predictions with a different initial condition or called as ensemble prediction [1,2].

In a poor ensemble prediction, there are bias and dispersion errors because the prediction has not been corrected or calibrated [3]. Bias is a common problem of direct output (raw) dynamical weather and climate model prediction and the dispersion errors are caused by overconfidence of the spread of ensemble member. Both of them lead to unreliable and unskillful probabilistic forecasts [4]. The seasonal forecast of the European Center for Medium-Range Weather Forecast (ECMWF) System 4 (ECS4 model) and the latest System 5 are the examples of ensemble prediction with large bias inside their raw model especially in the tropical region [5–7]. Globally, the reliability of raw model cannot be categorized as a “perfect reliability” [8]. On the other hand, the raw model has large bias locally, such as in Indonesia region (Java island), and its reliability line lay on no-skill and no-resolution [7].

The problem of bias and dispersion errors related to the reliability of probabilistic forecasts can be overcome by statistical post-processing method through bias correction or ensemble calibration [4]. The method aims to improve the quality of the forecast. Bias correction can be done using quantile mapping and will be effective if there is a significant correlation between the raw model and observation [6,7] whereas the problem of dispersion errors and reliability can be solved by the Bayesian Model Averaging (BMA) method [3]. BMA was initially used for calibration of short-range predictions from multi model outputs. The results show that BMA successfully resolved dispersion errors and improved reliability in real and simulations cases [3]. BMA is also used to calibrate predictions from the short and medium-range Ensemble Prediction Systems (EPS), for example: (i) to calibrate the prediction of 1-10 days 2 metre of Temperature (T2m) from 18 members of the EPS Meteorology Service of Canada [9]; (ii) to calibrate the prediction of 3-10 days wind speed from 51 members of EPS ECMWF [10]; and (iii) to calibrate monthly rainfall predictions from 11 members EPS Seasonal to Sub seasonal ECMWF [11]. From those examples, it can be concluded that the ability of BMA is affected by the type of predictive Probability Density Function (PDF) distribution. In addition to that, the use of the type of training window also affects the ability of the BMA [12]. Other studies reveal that the use of past predictions (re-forecast) shows better BMA results [4,9].

The utilization of BMA has been widely used to calibrate ensemble prediction in ‘short to medium range’ but few documents discuss the utilization of BMA for calibrating ensemble prediction in the ‘long-range’ or seasonal) forecast. Therefore, this study focused on BMA for calibrating seasonal forecast using monthly rainfall from the output of ECS4 model and would examine the availability of re-forecast that can be used to enhance the ability of BMA by configuring the type of training window on the BMA. This study aimed to apply BMA to calibrate seasonal forecast and to find out that BMA can be used to improve the quality of seasonal forecast for the Indonesia region, especially over Java Island.

2. Data and Methods

BMA was applied to one- and three-monthly rainfall data. Observation data were obtained from 26 stations of BMKG on the Java Island from 1981 to 2018 (456 months). Furthermore, for the calibration process, the calibrated model data was monthly rainfall from the output of the seasonal forecast of ECS4 model within1981-2018. This model had spatial resolution ~ 0.7°x0.7°, 15 ensemble members, 34 years (1981-2014) re-forecast products and four years (2015-2018) forecast. Although, there were seven (four) Lead Times (LT) for predictions of rainfall one (three) months, this study only exercised zero lead times (LT0) to be calibrated [5].

Furthermore, the methodology was using statistical approaches which were bias correction and ensemble calibration. First, as the ensemble member of raw model ECS4 (hereafter: raw model) has bias and dispersion errors, the authors applied a Quantile Mapping (QM) and then analyzed the dispersion error in the ensemble member of the bias correction output, called as BCQM. After that, we applied BMA on the raw model with two types training windows i.e. Sequential (BMA-TWS) and Conditional (BMA-TWC). Finally, we evaluated the distribution characteristics of the ensemble member raw model versus BMA output and evaluated the probabilistic forecast extreme event
condition in seasonal term i.e. Above Normal (AN) and Below Normal (BN), which were generated from ensemble member of raw model and BMA output.

2.1 Bias correction gQM and BMA calibration
There were two post processing methods used: firstly, bias correction using parametric quantile mapping with Gamma distribution (gQM) [13]. Each ensemble member of raw model was corrected based on the correction factor which obtained during the training period. Bias correction had been done with the Leave One Out Cross Validation (LOOCV) per year. Secondly, the BMA is selected as the calibration ensemble technique [3,14]. BMA is a statistical post-processing method that produces probabilistic forecasts from ensemble predictions in the form of predictive probability density functions (PDF) [3,14].

For BMA, each ensemble member \( f_k, k = 1,2, ..., K \) has posterior PDF \( g_k(y|f_k) \) which means that \( f_k \) is a condition for the occurrence of \( y \), where \( y \) is an monthly rainfall observation. BMA predictive model for ensemble predictions was expressed in equation (1),

\[
p(y|f_1, f_2, ..., f_k) = \sum_{k=1}^{K} w_k g_k(y|f_k)
\]

\( w_k \) is the weight of BMA which was obtained from the relative capabilities of the predictions of ensemble members in the training period. The Gamma distribution approach was used in the posterior PDF \( g_k(y|f_k) \) [4].

There were two ways to configure training period. BMA-TWS updated the training window with models and observations in month \( t-1 \) to calibrate predictions in the month \( t \) [3,14] whereas BMA-TWC used the same month from the previous year to update the training window [12]. Since the BMA output is PDF curves, the set of ensemble predictions for BMA was obtained from resampling the PDF curves to 15 members.

2.2 Evaluation of raw model versus BMA output
The evaluation of the distribution characteristics was done by assess the attributes of uncertainty, bias and accuracy [15]. These attributes can be measured by analyzing the shape of the Verification Rank Histogram (VRH) [16] and by calculating the Continuous Rank Probability Score (CRPS) [3,16].

Furthermore, the evaluation of probabilistic forecast of BN and AN was obtained from the assessment of the attributes of reliability, resolution, discrimination, and skill showing from Reliability Diagram (RD), the Area Under Curve of the Relative Operating Characteristic (AUC ROC) and the Brier Skill Score (BSS) [16]. Related to reliability, there were five categories of reliability (category-1: dangerous, -2: not useful, -3: marginally useful, -4: still very useful and -5 perfect) which is determined by the slope of the RD line [8]. The slope shows how useful probabilistic predictions can be used in making decisions related to seasonal forecast.

3. Result
3.1 The effect of the bias correction on the raw model ECS4
In spatial and temporal, the raw model has varied bias i.e. under- and over-estimate compare to the observed value over Java Island. Figure 1a shows value of the raw model is less than 400 mm while the observation reaches 700 mm/month. Moreover, the raw model also has a dispersion error (red bar Figure 1b). Shape of VRH of raw model is spikes on the left and right (U-shaped) meaning that many observed values are located outside of all ensemble member. In other words, uncertainty raw model to represent observation is large. U-shape of VRH also has meaning the spread of ensemble members is too narrow so that it tends to overconfidence and unreliable ensemble prediction [17].

Furthermore, we found that gQM can successfully reduce bias in the raw model which is showed by increasing R-square from 0.70 to 0.89 (Figure 1a). Then, the result of VRH of BCQM is slightly better than the VRH of raw model (Figure 1b) whereby an increasing bar in the rank-2 to rank-15. This means that the number of observations in the ensemble member of BCQM is more than the raw model. In other words, BCQM is better than raw model for presenting uncertainty of observation.
However, gQM cannot improve the reliability of ensemble members’ raw model because shape of VRH of BCQM is still U-shaped (blue bar Figure 1b). To sum up, bias correction is only correcting errors in the central tendencies for each ensemble member and unable to calibrate all ensemble members. Therefore, dispersion errors still occur within BCQM ensemble members [4]. As a result, the BMA post-processing method is needed to overcome this problem.

3.2 The capability of BMA to improve quality of raw model ECS4

The ensemble members of ECS4 raw model were generated from one single model which was treated with random perturbation by using stochastic physics [5]. Statistically, each member was indistinguishable. Therefore, BMA was carried out with the concept of exchangeable [18]. Based on this concept, $w_k$ in equation (1) is not estimated because the value of $w_k$ ($k=1,2, ..., 15$) is considered to be equal 1/15 for each member [18].

3.2.1 BMA with Sequential training window. BMA-TWS was done by taking several months as a training period before testing month. So, there were many choices for the length of the training window (TW). This study examined 12 TW lengths ranging from TW10 to TW120 months. However, there was only one TW used in the BMA-TWS calibration process. The determination of the length of optimal TW was chosen based on the average CRPS values of all stations. The smaller CRPS indicates that the BMA-TWS predictive PDF has a narrowed interval than the raw model predictive PDF. This means that the observed value is closer to BMA-TWS predictive PDF interval than the raw model.

Figure 2 shows average CRPS value for all stations. It starts from ~67 mm for TW10 months and continues to decrease to ~58 mm for TW30. Afterwards, CRPS is relatively stable. As a result, TW30 month is chosen as the optimal TW. Furthermore, by using TW30 months, we obtain a total of 426 months for the testing period (July 1983 to December 2018). Each BMA-TWS predictive PDF in the testing period was obtained by estimating the “shape” and “scale” Gamma parameters of posterior PDF. An example of calibration results of BMA-TWS predictive PDF in Darmaga station in February 2017 is showed by Figure 3a. In Figure 3a, BMA-TWS PDF predictive was plotted together with PDF of climatology rainfall for February (1981-2010) and PDF of raw model. PDF of climatology and raw model were fitted to the Gamma distribution using the Maximum Likelihood Estimator method [4].

![Figure 1](image_url)

**Figure 1.** Bias and dispersion errors in the raw model (red), BCQM (blue), BMA (black) against observed data in (a) scatter plot and (b) VRH. Both are generated from the average monthly rainfall (1981-2018) of all stations on the Java Island. Solid red line in (a) shows 400 mm/month as the maximum value of raw model.
Figure 2. Composite CRPS for all stations for each length of TW. TW30 months (yellow vertical line) is selected as the optimum length.

3.2.2 BMA with Conditional training window. BMA-TWC was carried out with LOOCV per month per year. For each calibrated testing month, a different BMA-TWC PDF was obtained (black curve in Figure 3a). This curve appears coincided with the PDF of climatology because the training month was the same as the climatology month. Overall, the BMA-TWC approach is better than the BMA-TWS. This is because (i) the annual cycle of rainfall pattern is similar with observation pattern (Figure 3b); (ii) the spread of 50% BMA-TWC PDF predictive intervals is sharper; and (iii) the accuracy (Taylor Diagram: correlation, standard deviation and RMSE, figure not shown) of the median PDF BMA-TWC versus observation is higher than BMA-TWS. Although BMA-TWC approach is better than the BMA-TWS, they are superior to the raw model.

In Figure 4, BMA-TWC is able to increase BSS of AN and BN positively ranging from 0.1 to 0.2. This is occurred almost in 25 (26) stations for BN (AN). On the other hand, BSS positive in BMA-TWS occurs in 15 (16) for BN (AN) and the rests are negative BSS. This BSS value shows that BMA-
TW C is better in reducing the error rate of probabilistic forecasts of BN and AN compared to BMA-TWS.

In Figure 5, reliability raw model in all stations shows category-1 (not useful) and category-2 (dangerous). Dangerous reliability means that probabilistic forecast of occurrence BN and AN do not reflect the actual events. In the long run, this can lead to misinformation for making a decision of the seasonal forecast. BMA-TWS can increase the reliability of the raw model, but the increasing generated by BMA-TWC is better than BMA-TWS. This is because the RD category is dominated by category-4 (still very useful) and 5 (perfect) while RD BMA-TWS is generally located in category-3 (marginally useful) and 4 (still very useful). However, both BMA-TWC and BMA-TWS can improve the reliability of the raw model. As a result, they are able to make AN and BN prediction useful in making any decisions.

Figure 4. Comparison of BSS of raw model versus BMA-TWS (upper row) and raw model versus BMA-TWC (lower row), for BN (a, b) and AN (c, d). BSS of raw model is indicated with + (positive) and • (negative) and the color on the station symbol is the BSS value for BMA-TWS and BMA-TWC

Figure 5. Comparison of Reliability Diagram category for: raw model (a, b), BMA-TWS (c, d) and BMA-TWC (e, f), probability forecast of Below Normal (upper row) and Above Normal (lower row)
Finally, BMA-TWC is proven as a statistical post-processing method to overcome the problem of bias as well as dispersion errors, compared to the gQM method. BMA-TWC reduces the bias so that the mean of BMA-TWC PDF is very close to observation with the R-square of 0.99 (black scatter plot in Figure 1a). BMA-TWC also improves the reliability of the spread of ensemble members of the raw model reflected by the flat shape of VRH (black bar Figure 1b).

### 3.3 Skill BMA in dry, wet season and ENSO phase

#### 3.3.1 Skill BMA in dry and wet season

Temporally, the ability of BMA varies every month. Based on the value of AUC ROC, BMA shows the same skill as raw model because it has value > 0.5 (Figure 4). The raw model actually has quite good resolution, discrimination and skill in the dry season compared to the rainy season in predicting AN and BN events. However, BMA is able to enhance the skill compared to the raw model especially in the dry season. Conversely, in the rainy season, the BMA is unable to improve the skill of the raw model in predicting AN and BN events.

Although the raw model has quite good resolution, discrimination and skill to predict BN and AN event, the error rate of probabilistic forecast is large. Figure 6 (red bar) shows that the BSS raw model value is always negative every month except for October. But the BMA is able to fix it, especially in the dry season, the BMA error rate was very small in predicting AN and BN events because the BSS BMA value is always positive in the JJASON period. On the other hand, in the wet season, BSS BMA is almost close to zero. It means that BMA is able to reduce the error rate of the raw model which is quite large in the rainy season even though the BMA skill is not better than the reference forecast to predict AN and BN events in the DJF-MAM period.

#### 3.3.2 Skill BMA in ENSO (El Nino) phase

One of the factors influencing rainfall in Indonesia is ENSO (El Nino Southern Oscillation). The ENSO phase, particularly El Nino, is correlated with reducing of rainfall in some Indonesia regions, especially Java Island [19]. During the El Nino phase there is a decreasing of rainfall to below normal conditions. For example, in strong El Nino 2015, almost all regions of Indonesia experienced very small rainfalls lower than the 33rd percentile (Below
Normal condition) [20].

The ability of BMA to forecast BN condition in the El Nino phase can be determined by applying BMA calibration to three-month rainfall such as JFM, FMA, etc. Therefore, it met the ENSO index category based on the ONI (Ocean Nino Index, https://origin.cpc.ncep.noaa.gov/products/analysis_monitoring/ensohtml/ONI_v5.php). The ability of BMA can be shown by a larger percentage of BN probability generating from the BMA PDF during El Nino compared to the raw model PDF. Even though the raw model PDF is able to provide predictions of BN events during El Nino Strong, the intensity of its probabilities is smaller than the BMA PDF probabilities. Generally, the predictive BMA PDF (raw model PDF) presents a probability of BN around 50% - 60% (33% - 40%) for each station. This leads to a different reliability between the raw model and BMA in predicting BN events during El Nino phase.

The reliability line of BN in raw model is flat categorized as not useful category (figure not shown). This indicates that the forecast of BN is inconsistent with the real events. However, after being calibrated with BMA, the RD category increase to category 4 (still very useful) and 5 (perfect) meaning that BMA is able to produce probability forecast of BN which is similar with the real events. Thus, this is useful for the users in making decisions of seasonal forecast.

4. Conclusion
In general, BMA is able to improve the quality of ensemble predictions of raw model ECS4 for all BMKG stations over Java Island and the results of BMA-TWC was better than the BMA-TWS. The ability of BMA has been evaluated from two steps. That shows BMA is able to improve the distribution characteristics of raw model (reduced bias, uncertainty and increased accuracy based on flat VRH and smaller CRPS) and also able to improve the quality of the raw model predictability for probabilistic forecasts of BN and AN (increased resolution and skill based on AUC ROC values > 0.5, positive BSS, especially in the dry season). The BMA also improved category of the reliability of BN and AN becoming the “still very useful” and “perfect”. In the El Nino period, the BMA also improved the reliability of probabilistic forecast of BN for all categories of El Nino which are “Strong”, “Moderate” and “Weak”. The improvement of the reliability category shows BMA could improve quality of raw model so that probability forecast of BN and AN can be used in making decisions of seasonal forecast. However, BMA calibration results were varying in temporal. BMA has not succeeded in improving the quality of raw model during the wet season. This might be caused by too large variability of rainfall in wet season. As a result, there is a need to further study discussing BMA particularly in wet season.

Acknowledgments
This study is a part of the master thesis of the first author in Atmospheric Sciences Research Group, Faculty of Earth Sciences and Technology, Institute Technology Bandung and we would like thanks to The Center of Training and Education of BMKG for supporting the scholarship.

References
[1] Slingo J and Palmer T 2011 Uncertainty in weather and climate prediction Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 369 4751–67
[2] Palmer 2000 Predicting uncertainty in forecasts of weather and climate prediction Reports on Progress in Physics 63 71–116
[3] Raftery A E, Gneiting T, Balabdaoui F and Polakowski M 2005 Using Bayesian Model Averaging to Calibrate Forecast Ensembles Monthly Weather Review 133 20
[4] Gneiting T 2014 Calibration of Medium-Range Weather Forecasts ECMWF Technical Memorandum 719 30
[5] Molteni F, Stockdale T, Balmaseda M, Balsamo G, Buizza R, Ferranti L, Magnusson L, Mogensen K, Palmer T and Vitart F 2011 The new ECMWF seasonal forecast system (System 4) ECMWF Technical Memorandum 656 51
[6] Manzanas R, Lucero A, Weisheimer A and Gutiérrez J M 2018 Can bias correction and statistical
downscaling methods improve the skill of seasonal precipitation forecasts? Climate Dynamics 50 1161–76
[7] Ratri D N, Whan K and Schmeits M 2019 A Comparative Verification of Raw and Bias-Corrected ECMWF Seasonal Ensemble Precipitation Reforecasts in Java (Indonesia) J. Appl. Meteor. Climatol. 58 1709–23
[8] Weisheimer A and Palmer T N 2014 On the reliability of seasonal climate forecasts Journal of The Royal Society Interface 11 20131162–20131162
[9] Wilson L J, Beauregard S, Raftery A E and Verret R 2007 Calibrated Surface Temperature Forecasts from the Canadian Ensemble Prediction System Using Bayesian Model Averaging Monthly Weather Review 135 1364–85
[10] Han K, Choi J and Kim C 2018 Comparison of Statistical Post-Processing Methods for Probabilistic Wind Speed Forecasting Asia-Pacific J Atmos Sci 54 91–101
[11] Kolachian R and Saghafian B 2019 Deterministic and probabilistic evaluation of raw and post processed sub-seasonal to seasonal precipitation forecasts in different precipitation regimes Theor Appl Climatol 137 1479–93
[12] Erickson M J, Colle B A and Charney J J 2012 Impact of Bias-Correction Type and Conditional Training on Bayesian Model Averaging over the Northeast United States Wea. Forecasting 27 1449–69
[13] Piani C, Haerter J O and Coppola E 2010 Statistical bias correction for daily precipitation in regional climate models over Europe Theoretical and Applied Climatology 99 187–92
[14] Sloughter J M L, Raftery A E, Gneiting T and Fraley C 2007 Probabilistic Quantitative Precipitation Forecasting Using Bayesian Model Averaging Monthly Weather Review 135 3209–20
[15] Murphy A H 1993 What Is A Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting Weather and Forecasting 8 281–93
[16] Jolliffe I T and Stephenson D B 2003 Forecast verification: a practitioner’s guide in atmospheric science (Chichester, West Sussex, England; Hoboken, NJ: J. Wiley)
[17] Hamill T M 2001 Interpretation of Rank Histograms for Verifying Ensemble Forecasts Monthly Weather Review 129 11
[18] Schmeits M J and Kok K J 2010 A Comparison between Raw Ensemble Output, (Modified) Bayesian Model Averaging, and Extended Logistic Regression Using ECMWF Ensemble Precipitation Reforecasts Monthly Weather Review 138 4199–211
[19] Supari, Tangang F, Salimun E, Aldrian E, Sopaheluwakan A and Juneng L 2018 ENSO modulation of seasonal rainfall and extremes in Indonesia Clim Dyn 51 2559–80
[20] Supari, Robi Muharsyah and Niken Wahyuni 2016 Impact of the 2015 Godzilla El Niño event on the Indonesian rainfall Scientific Journal of PPI-UKM 26–31