Improving prediction quality of sea surface temperature (SST) in Niño3.4 region using Bayesian Model Averaging

Robi Muharsyah¹, Dian Nur Ratri¹, Damiana Fitria Kussatiti¹

¹Agency for Meteorology, Climatology and Geophysics of Republic of Indonesia (BMKG), Jakarta, 107720, Indonesia

Email: robi.muharsyah@gmail.com

Abstract. Prediction of Sea Surface Temperature (SST) in Niño3.4 region (170 W - 120 W; 5S - 5N) is important as a valuable indicator to identify El Niño Southern Oscillation (ENSO), i.e., El Niño, La Niña, and Neutral condition for coming months. More accurate prediction Niño3.4 SST can be used to determine the response of ENSO phenomenon to rainfall over Indonesia region. SST predictions are routinely released by meteorological institutions such as the European Center for Medium-Range Weather Forecasts (ECMWF). However, SST predictions from the direct output (RAW) of global models such as ECMWF seasonal forecast is suffering from bias that affects the poor quality of SST predictions. As a result, it also increases the potential errors in predicting the ENSO events. This study uses SST from the output Ensemble Prediction System (EPS) of ECMWF seasonal forecast, namely SEAS5. SEAS5 SST is downloaded from The Copernicus Climate Change Service (C3S) for period 1993-2020. One value representing SST over Niño3.4 region is calculated for each lead time (LT), LT0-LT6. Bayesian Model Averaging (BMA) is selected as one of the post-processing methods to improve the prediction quality of SEAS5-RAW. The advantage of BMA over other post-processing methods is its ability to quantify the uncertainty in EPS, which is expressed as probability density function (PDF) predictive. It was found that the BMA calibration process reaches optimal performance using 160 months training window. The result show, prediction quality of Niño3.4 SST of BMA output is superior to SEAS5-RAW, especially for LT0-LT2. In term deterministic prediction, BMA shows a lower Root Mean Square Error (RMSE), higher Proportion of Correct (PC). In term probabilistic prediction, the error rate of BMA, which is showed by the Brier Score is lower than RAW. Moreover, BMA shows a good ability to discriminating ENSO events which indicates by AUC ROC close to a perfect score.

1. Introduction

El Niño (La Niña) or known as ENSO, is a phenomenon in the equatorial Pacific Ocean characterized by sea surface temperature (SST) anomalies in the Niño3.4 region that is above (below) the threshold of +0.5°C (-0.5°C) [1]. For the Indonesian region, more accurate ENSO prediction is essential, because ENSO is one of the factors affecting rainfall in Indonesia. Many studies have discussed the relationship between ENSO phenomena to rainfall in Indonesia. They have a similar conclusion that El Niño (La Niña) leads to decrease (increase) rainfall in most of the Indonesia region [2–4]. For example, Strong El Niño in 2015 with the highest SST anomaly (SSTA) reach 2.8°C in November 2015, caused most of Indonesia to experience a rainfall less than its climatology [5].
Indonesia Agency for Meteorology, Climatology, and Geophysics (BMKG) routinely monitor the latest condition and predict SST in the Niño3.4 region for several months ahead [6]. ENSO prediction from various international institutions such as ECMWF (Europe), JMA (Japan), BoM (Australia), and CPC NOAA (USA) was collected to be compared among each other. The goal is to obtain a consensus regarding SST (warmer or cooler) in the Pacific Ocean for the coming months. One of the models that are often used in considering the ENSO prediction is the SST prediction from the ECMWF seasonal forecast system, SEAS5. Operationally, BMKG subscribed and gained access to ECMWF forecast products under a specific license since 2014. Many studies revealed that skill of SST released by ECMWF both SEAS4 (previous version of SEAS5) and SEAS5 is better, such as when it was compared to CFSv2 [7]. Kim et al. 2012 shown that the SEAS4 SST is more accurate than the CFSv2 prediction model [7]. Recently, Johnson et al. 2019, also proved that the prediction SST in Niño3.4 using SEAS5-corrected has better skills than SEAS4, which indicated reduced bias, higher correlation, lower amplitude, and lower RMSE (Figure 1) [8].

The results presented by Johnson et al. 2019, have applied Bias Correction (BC) method to remove bias in the SEAS5. BC is the most popular statistical post-processing method used to improve the seasonal forecast model [9–11]. However, several studies were reported that BC could not be solving the overconfident problem (under-dispersive or over-dispersive) indirect output (raw) of ensemble prediction, such as SEAS5. The overconfident problem makes a probabilistic prediction which generated from ensemble prediction became unreliable [10,11]. Unfortunately, the study reported by Johnson et al. 2019, we did not find any evaluation that showed the ability of SEAS5 for probabilistic forecast, particularly to predict the probability of ENSO events per category separately: El Niño, La Niña, and Neutral.

Instead of BC, this study aims to apply a statistical post-processing method as an alternative to correct the model’s bias. The method, called as “the calibration technique”, can to quantify the uncertainties inside ensemble predictions [10,12]. These uncertainties are measured by using probabilistic predictions that can be obtained from ensemble predictions of SEAS5 SST. Therefore, a calibration method is needed in order to determine the “best” predictions of the probability of ENSO events. The popularity of the calibration method has grown in recent years [13–21]. The methods include Bayesian Join Probability [10,20], Forecast Assimilation [14,16], and Bayesian Model Averaging [13,18,19,21]. So, it
is expected that the implementation of the calibration method will improve the prediction quality of SEAS5 SST in Niño3.4.

2. Data and Methods

2.1 Model data
We use the Ensemble Prediction System (EPS) of ECMWF seasonal forecast SEAS5 [8]. The SEAS5 real-time forecast has been regularly utilized as the operational forecast since released in November 2017, replacing SEAS4 [22]. SEAS5 possesses improved related to the ocean-atmosphere model used in the Integrated Forecasting System ECMWF. The SEAS5 and SEAS4 atmosphere models have the same initialization. However, the initialization of the sea-ice and ocean models of SEAS5 uses an upgraded operational model, OCEAN5. This ocean system is composed of ORAS5 for historical reanalysis and OCEAN5-RT for daily real-time analysis. Unperturbed and perturbed atmospheric initial conditions, as well as stochastic model perturbations, are applied to generate ensemble members. The spatial resolution of SEAS5 is about 36 km makes it the only high-resolution global seasonal forecast model. However, this resolution is available for the restricted data category, which can only be accessed using a specific license. Details on the configuration of SEAS5 and its comparison to other global seasonal forecast models are available at https://wmolc.org/contents2/index/ECMWF.

![Niño3.4 region](https://www.ncdc.noaa.gov/teleconnections/enso/indicators/sst/).

Figure 2. Niño3.4 region is located in the central Pacific Ocean with domain 170 W – 120 W; 5S – 5N. In this study, Seas Surface Temperature (SST) from the output of SEAS5, in particular over the Niño3.4 region (170 W – 120 W; 5S – 5N), is selected to be analyzed (see Figure 2). The monthly average of the SEAS5 SST is downloaded from The Copernicus Climate Change Service (C3S), for the reforecast period of January 1993 - October 2020 (334 months) (https://cds.climate.copernicus.eu/). For each month, the SEAS5 SST has 25 ensemble members with a forecast of up to seven months ahead, known as lead-time (LT0-LT6). C3S provides the SEAS5 reforecast for free, but it has lower resolution around 1°x1° so for the Niño3.4 domain box, there are 561 (51 x 11) grid cells. The fldmean function in Climate Data Operator is used to get a single value of SST in the Niño3.4 region. This uncalibrated SEAS5 SST is hereinafter referred to as RAW.
2.2 Observed data
The Observed value of the SST Niño3.4 comes from ERSST version 5. This data can be downloaded at https://www.cpc.ncep.noaa.gov/data/indices/ersst5.Niño.mth.91-20.ascii. The Extended Reconstructed Sea Surface Temperature (ERSST) dataset is a global monthly SST dataset derived from the International Comprehensive Ocean-Atmosphere Dataset (ICOADS). ERSSTv5 is used as the input for calculating the Ocean Niño Index (ONI) used for ENSO monitoring by CPC NOAA (USA).

2.3. Bayesian Model Averaging
BMA is one of the affine kernel dressing (AKD) methods that can be used to calibrate ensemble predictions [23]. The main objective of the BMA method is to generate probabilistic predictions are expressed in predictive PDF, namely BMA PDF [23–26]. BMA PDF is obtained based on information in the past (training window) between each member of the ensemble predictions against observed value.

Each ensemble member has a posterior PDF which contributing to the BMA PDF. Let is y, observed value and \( f = f_1, f_2, ..., f_k \) is ensemble prediction of SEAS5 SST, where \( k = 25 \). Each ensemble member, \( f_k \) has a posterior PDF, \( g_k(y|f_k) \), so BMA PDF predictive model can be expressed in equation (1). In addition, the mean and median attributes of BMA PDF are used as deterministic predictions of BMA [24].

\[
p(y|f_1, f_2, ..., f_k) = \sum_{k=1}^{25} w_k g_k(y|f_k)
\]  

(1)

The main source of the ensemble member of SEAS5 is derived from a single model with random perturbation [8,22]. This condition leads to consequence on the weight of \( w_k \) in equation (1). In other words, statistically, each member of the predictions indistinguishable. Therefore, in this study, BMA is carried out with the concept of exchangeable [26]. As a result, the estimation of weight \( w_k \) is not carried out because \( w_k \) is assumed to be the same for \( k = 1, 2, ..., 25 \), which is 1/25 for each member.

Next, a normal distribution is selected to get the posterior BMA PDF, \( g_k(y|f_k) \) [24]. Determination of the type distribution was adjusted to the various variables to be calibrated, for instance Precipitation and Wind Speed (Gamma) and Temperature (Normal) [27]. Theoretically, the BMA calibration process is starting by estimate the components that build up the posterior BMA PDF which derived from each member, \( f_k \). For Normal distribution, there are two components i.e., mean \( \alpha_k + b_k f_k \) and standard deviation, \( \sigma \). The coefficients \( \alpha_k, b_k \) and \( \sigma \) will be estimated using the Maximum Likelihood (ML) approach, the Expectation-Maximization (EM) algorithm, and the minimum Continuous Rank Probability Score (CRPS) estimation [24]. Details about BMA theory can be found in Raftery et al 2005 [24]. In this study, the BMA calibration was carried out using ensemble BMA library in the R package [28].

2.4. BMA Training Window
One of the crucial parts of the BMA calibration process is determining the training window (TW) [21]. In this study, the BMA training window is performed sequentially. The reason is that the variations of SST Niño3.4 tend to be normally distributed, and there is no significant variation between seasons. Moreover, the Sequential Training Window (STW) is the original method was developed when BMA was introduced by Raftery, et al. 2005 to calibrate T2m and MSLP. To calibrate the time \( T^n \), the STW will use the observation and model on the \( T-1, T-2, ..., T-n \) months, where \( n \) is the length of TW. For example, if TW 10 months is chosen, the BMA process starts from the first ten months of the data, i.e., period Jan 1993 - Oct 1993 with Nov 1993 as the target month to be calibrated. Then it shifted one month ahead, February 1993 - November 1993 as the training period and December 1993 as the target month, and so on for TW 20, 30, 40, etc. The consequence is that the longer the TW is selected, the
shorter the testing period (number of target month) will be obtained. The optimal length of TW is selected based on the minimum value of CRPS [24]. The smaller CRPS indicates that the BMA PDF has a narrower interval than the RAW PDF, which is means the observed value closer to the BMA PDF interval than the RAW PDF.

2.5 Forecast Evaluation
The evaluation focuses on comparing the uncalibrated SEAS5 SST, i.e., RAW versus calibrated SST, i.e., BMA output. Detailed information about all the scores which are used for the evaluation can be found in Jolliffe I T and Stephenson D B 2003 [29]. In general, there are two steps:

1. Evaluate the deterministic predictions of the ensemble mean RAW (RAW-mean) against the mean and the median of the BMA PDF (BMA-mean and BMA-median). The Evaluation includes the similarity pattern using Pearson correlation score and the accuracy using the Root Mean Square Error (RMSE). We also compute categorical accuracy using a proportion of correct (PC) [28] between the RAW-mean, the BMA-mean, and the BMA-median to predict the three categories of ENSO i.e., El Niño, La Niña, and Neutral.

2. Evaluate the probabilistic predictions for each ENSO category. For BMA, probability of El Niño, La Niña, and Neutral events are generated from BMA PDF, meanwhile for RAW, the probability are calculated based on the number of members in each ENSO categories divided by 25 members. Brier Score (BS) and Relative Operating Characteristic (ROC) scores are used to evaluating the probabilistic prediction of ENSO. BS is the mean of squares error probabilistic prediction of an event [29], BS measures the magnitude error for dichotomous events. For example, "El Niño occurs" or "El Niño not occur" the BS is getting better if it is close to zero. BS formula follows equation (2), where \( N \) is the number of samples, \( F_i \) is the probability prediction of an event, and \( O_i \) is the categorical observed value, 1 (occurs) and 0 (not occur). ROC measure the ability of the prediction system to distinguish between an event and the corresponding non-events, which can be described in terms of hit rate and false alarm rate, respectively [29]. The ROC score is measured by the Area Under Curve (AUC). If the AUC ROC > 0.5, it means that the prediction system well discriminating an event, for example, "El Niño" or "not El Niño" and AUC ROC equal to 1 means perfect score in discrimination.

\[
BS = \frac{1}{N} \sum_{i=1}^{N} (F_i - O_i)^2
\]  

Figure 3. CRPS for each experiment of TW10 months up to TW200 months.
3. Result

3.1. Optimal Training Window Selection
This study examines 20 training window lengths ranging from TW10 to TW200 months. It has an interval per 10 months. However, there is only one selected as “the best” TW to be evaluated further. Figure 3 shown that TW160 is the best TW. The CRPS value decrease from 0.11 in the first TW10 to 0.08 in TW160, then increases toward TW200. The smallest CRPS for TW160 indicates that the BMA PDF has the narrowest interval compared to other TWs. This means that the observed value is closest to the BMA PDF when using TW160. We also do the same experiment for LT1 to LT6 and found optimal TW ranging from 150-180. However, only TW160 is selected for all LTs. We assume that LT0 is the best LT compared to other LTs.

3.2. Comparison of BMA PDF and Spread RAW
Based on the optimal TW160 months, we obtain 174 BMAPDF through the testing period (the target months period is May 2006 to October 2020). For each target month, various BMA PDF is found, both in density (curve height) and variance (curve width). Figure 4 is an example of BMA PDF for the first three LT, LT0, LT1, and LT2 to predict Strong El Niño (October 2015) and Strong La Niña (October 2010). SST observed values for both extreme events were 29.08°C and 25.67°C, respectively.

For example, when Strong El Niño occurred (Figures 4a, 4b and 4c), we found that the BMA-mean and the BMA-median are relatively closer to the observed value than the RAW-mean for LT0 and LT1. In these figures, the RAW-mean is consistently far to the right side of the observation and located at the end of the BMA PDF curve. This means that the RAW-mean tends to be overestimated or positively biased. Moreover, the bias in RAW is getting wider from LT0 to LT2. On other hands, in Strong La Niña events (Figures 4d, 4e, and 4f), the ability of BMA similar to RAW. The RAW-mean, BMA-mean or BMA-median very close to the observed value. In fact, on LT2, RAW-mean was seemed closer to the observed value than either BMA-mean, or median BMA. Overall, ability of the BMA is better when the SST Niño3.4 tend to El Niño rather than La Niña Figure 4 (a – f). Besides, we found the observed value of SST is consistent within a range P25 to P75 of BMA PDF. In other words, 50% of the BMA PDF able to capture observed value, which means that BMA can improve RAW spreads, as seen in Figures 4a. and 4b.

If the results in Figure 4 are displayed as a time series throughout the testing period, we will get Figure 5. Nevertheless, the SST value in figure 5 has been converted into SST anomaly (ASST) Niño3.4. By using ASST, we can easily identify the performance of BMA and RAW for each ENSO categories. In general, spread RAW tends too narrow and far from the observed value for LT0, in particular during El Niño and Neutral phase. Meanwhile, in LT1 and LT2, the spread of RAW too wide even though it can cover the observed value for each ENSO category. Too narrow or too wide spread RAW show that SEAS5 RAW suffers from overconfident prediction. This condition can be a common problem for the ensemble prediction from the GCM output model [10,11,13,21,24].
Figure 4. BMA predictive PDF (black curve) for (a, d) LT0, (b, e) LT1, and (c, f) LT2 to predict SST in Strong El Niño in October 2015 (upper row) and Strong La Niña in October 2010 (lower row). Black dash vertical lines are percentile 25th (P25) and 75th (P75) of BMA PDF, red dash vertical lines are minimum (RAW-min) and maximum (RAW-max) of RAW ensemble member. Solid vertical lines with blue, green, red, and orange colors are the mean BMA PDF (BMA-mean), the median or the P50 BMA PDF (BMA-median), the RAW ensemble means (RAW-mean), and the observed value, respectively.

3.3. Evaluation of deterministic prediction of RAW versus BMA
Firstly, deterministic predictions between the RAW versus the BMA are measured from the accuracy of categorical predictions for guessing ENSO categories. The results show the BMA-median and the BMA-mean have higher PC than the RAW-mean for LT0 and LT1, with BMA-median being the superior among all. However, BMA accuracy is significant in the first three LTs, while for LT3 to LT6, PC BMA-mean and PC BMA-median drops below the PC RAW-mean.

Furthermore, the accuracy of the RAW and the BMA are measured using Correlation and RMSE. We compare the patterns between the RAW-mean, the BMA-mean, and the BMA-median with respect to the pattern of the observed value. They are very similar to the observed pattern, as shown in figure 5. Similarity pattern has consequences correlation score for the RAW-mean, the BMA-mean, and the BMA-median, they are the same (Figure 7a). For example, in LT0, they correlate around 0.99, then gradually drop to 0.78 in LT6, which is consistent with previous studies [8,31].

Although at LT0, the correlation shows a perfect score (0.99), the RAW-mean is not close to the observed value (see Figure 5a). In LT0, the RAW-mean tends to overestimate, especially when the ASST in the Neutral and El Niño category (SSTA>0), while in the La Niña, the RAW-mean is slightly closer to the observed value. Overall, in Figures 5 (a-c), both the BMA-mean and the BMA median are better than the RAW-mean due to eliminating bias; however, it is only valid for the first three LTs. Figure 7b summarizes the comparison of the RMSE for the entire testing period. We can see that the RMSE BMA-median for LT1 and LT2 is around 0.1-0.3; this is lower than the RAW-mean around 0.3-0.4. Besides, for all LTs, RMSE increases gradually, which is also found in other studies [8,31].
Figure 5. SSTA in Niño3.4 region for each target month through testing period for (a) LT0, (b) LT1, (c) LT2. The red shading area shows the range minimum (RAW-min) and maximum (RAW-max) of the RAW ensemble members. For solid lines with blue, green, red, and orange colors are mean BMA PDF (BMA-mean), median or P50 BMA PDF (BMA-median), RAW ensemble means (RAW-mean) and observed value, respectively. The horizontal dash line is threshold ±0.5 to separate El Niño, Neutral, and La Niña categories.
3.4. Evaluation of probabilistic prediction of RAW versus BMA

The evaluation of probabilistic predictions for each ENSO category separately; El Niño, La Niña, and Neutral is important to be investigated. In general, ENSO prediction is released in terms of their “chance” or “probability” (https://iri.columbia.edu/our-expertise/climate/forecasts/enso/current/). The results show that the BMA is superior to the RAW in predicting the probability of El Niño and Neutral events for all LTs. The BS value of the BMA < 0.1 for the probability of El Niño events in particular for LT0 and LT. This means that the error rate of BMA in predicting El Niño events is less than 10%. For LT2 to LT6, BMA has a maximum of 20% error rate; meanwhile, the error rate of the RAW is up to 30%. Similar to El Niño, for Neutral events, the BMA shows better performances than the RAW. Nevertheless, for La Niña events, there is a slightly difference between the BMA and the RAW; the BMA is only superior to the RAW model for LT0, while for LT1 to LT6, both BMA and RAW have a similar average error rate in predicting La Niña events which are around 10% -20%.

Furthermore, this study also shows how well the BMA or the RAW in discriminating ENSO per events occurred by measure the AUC ROC. In general, both of AUC ROC scores of the BMA and the RAW is more than 0.5. In fact, for predictions of La Niña events, both of them have AUC ROC >0.8. The AUC ROC > 0.5 indicates that the BMA and the RAW have the same abilities in discriminating "events" or "not-events" for each ENSO category for all LTs. The BMA is relatively able to improve the RAW, especially for LT0 (El Niño and La Niña) and for LT0 - LT2 in Neutral events. Interestingly, we found that the probability of ENSO occurrences generated from the BMA PDF looks very close to being perfect (AUC ROC close to 1). This means that the BMA probabilistic prediction has a higher confidence level than the RAW.
4. Conclusion

Besides BC methods, BMA can be used as alternative post-processing to improve the quality of SST as one of the outputs of SEAS5. For the deterministic prediction of SST in the Niño3.4 region, BMA can provide accurate predictions from LT0 to LT2. Meanwhile, for probabilistic predictions, SEAS5-RAW relatively good to distinguish which ENSO categories will occur. However, if we compared to BMA, BMA is more superior to RAW for all LTs.

The output of BMA deterministic prediction is highly recommended to be used as the ASST Niño3.4 index. It is found that BMA is a promising output for two or three months ahead. This index is one of the routine climate services information products of BMKG that commonly been released to users and stakeholders regarding the ENSO prediction for coming several months.

In addition, another advantage of BMA is BMA does not need a real-time forecast for the target month. In its implementation, BMA only requires updating the observed values. For example, to produce SST prediction for June, so ideally around the 1st or the 2nd of June, the SST observation for May is already available to be downloaded from various sources such as ERSSTv5. Based on the SST observation in May coupled with the SST SEAS5 real-time forecast, which is already available for May, the BMA calibration process can be carried out to obtain prediction SST Niño3.4 for June to December. So, in its routine operation, by using the BMA method, BMKG can release prediction ASST Niño3.4 in the early month, does not have to wait for a real-time forecast of SEAS5, which usually available at 8th every month [8,22].

Figure 8. Brier Score (BS) of probabilistic prediction of ENSO categories (a) El Niño, (b) La Niña dan (c) Neutral for LT0 to LT6.

Figure 9. AUC ROC of probabilistic prediction of ENSO categories (a) El Niño, (b) La Niña dan (c) Neutral for LT0 to LT6.
5. Reference

[1] Wang C, Deser C, Yu J-Y, DiNezio P and Clement A 2017 *El Niño and Southern Oscillation (ENSO): A Review Coral Reefs of the Eastern Tropical Pacific Coral Reefs of the World* vol 8, ed P W Glynn, D P Manzello and I C Enockhs (Dordrecht: Springer Netherlands) pp 85–106

[2] Lestari D O, Sutriyono E and Iskandar I 2018 *ITB J. Publ.* 50 257–72

[3] Supari, Tangang F, Salimun E, Aldrian E, Sopaheluwakan A and Juneng L 2018 *Clim. Dyn.* 51 2559–80

[4] Alsepan G and Minobe S 2020 *J. Clim.* 33 5271–91

[5] Supari, Robi Muharsyah and Niken Wahyuni 2016 *Sci. J. PPI-UKM* 3 26–31

[6] Anon 2020 *Buku Prakiraan Musim Hujan 2020/2021 Di Indonesia* (Jakarta: Bidang Analisis Variabilitas Iklim BMKG)

[7] Kim H-M, Webster P J and Curry J A 2012 *Clim. Dyn.* 39 2957–73

[8] Johnson S J, et al 2019 *Geosci. Model Dev.* 12 1087–117

[9] Maraun and Widmann 2018 *Statistical Downscaling and Bias Correction for Climate Research* (United Kingdom: Cambridge University Press)

[10] Zhao T, Bennett J C, Wang Q J, Schepen A, Wood A W, Robertson D E and Ramos M-H 2017 *J. Clim.* 30 3185–96

[11] Manzanas R, Lucero A, Weisheimer A and Gutiérrez J M 2018 *Clim. Dyn.* 50 1161–1176

[12] Slingo J and Palmer T 2011 *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* 369 4751–67

[13] Wang Q J, Schepen A and Robertson D E 2012 *J. Clim.* 25 5524–37

[14] Rodrigues L R L, Doblas-Reyes F J and Coelho C A dos S 2014 *Clim. Dyn.* 42 597–616

[15] Wilks D S 2018 *Q. J. R. Meteorol. Soc.* 144 76–84

[16] Rodrigues L R L, Doblas-Reyes F J and Coelho C A S 2018 *Clim. Dyn.* 53, 7305–7320

[17] Manzanas R, Gutiérrez J M, Bhend J, Hemri S, Doblas-Reyes F J, Penabad E and Brookshaw A 2020 *Clim. Dyn.* 54 2869–82

[18] N W 2018 *J. Phys. Math.* 09

[19] Han K, Choi J and Kim C 2018 *Asia-Pac. J. Atmospheric Sci.* 54 91–101

[20] Wang Q J, Shao Y, Song Y, Schepen A, Robertson D E, Ryu D and Pappenberger F 2019 *Environ. Model. Softw.* 122 104550

[21] Muharsyah R, Hadi T W and Indratno S W 2020 *IOP Conf. Ser. Earth Environ. Sci.* 572 012034

[22] Molteni F, Stockdale T, Balmaseda M, Balsamo G, Buizza R, Ferranti L, Magnusson L, Mogensen K, Palmer T and Vitart F 2011 *ECMWF Tech. Memo.* 656 51

[23] Brcker J and Smith L A 2008 *Tellus A* 60 663–78

[24] Raftery A E, Gneiting T, Balabdaoui F and Polakowski M 2005 *Mon. Weather Rev.* 133 20

[25] Vrugt J A, Diks C G H and Clark M P 2008 *Environ. Fluid Mech.* 8 579–95

[26] Schmeits M J and Kok K J 2010 *Mon. Weather Rev.* 138 4199–211

[27] Gneiting T 2014 *ECMWF Tech. Memo.* 719 30

[28] Fraley C, Raftery A E, Gneiting T and Slaughter J M 2018 *ensembleBMA: An R package for Probabilistic Forecasting using Ensembles and Bayesian Model Averaging R Package Version 515 19

[29] 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)

[30] Wang J, Chen J, Du J, Zhang Y, Xia Y and Deng G 2018 *Mon. Weather Rev.* 146 781–96

[31] Barnston A G, Tippett M K, Ranganathan M and L’Heureux M L 2019 *Clim. Dyn.* 53 7215–34