Comparison of the BMA and EMOS statistical methods for probabilistic quantitative precipitation forecasting

Zohreh Javanshiri | Maede Fathi | Seyyedeh Atefeh Mohammadi

Abstract
The main approach to probabilistic weather forecasting has been the use of ensemble forecasting. In ensemble forecasting, the probability information is generally derived by using several numerical model runs, with perturbation of the initial conditions, physical schemes or dynamic core of the numerical weather prediction (NWP) models. However, ensemble forecasting usually tends to be under-dispersive. Statistical post-processing has, therefore, become an essential component of any ensemble prediction system aiming to improve the quality of numerical weather forecasts as they seek to generate calibrated and sharp predictive distributions of future weather quantities. Different versions of the ensemble model output statistics (EMOS) and the Bayesian model averaging (BMA) post-processing methods are used in the present paper to calibrate 24, 48 and 72 hr forecasts of 24 hr accumulative precipitation. The ensemble employs the weather and research forecasting (WRF) model with eight different configurations which were run over Iran for six months (September 2015–February 2016). The results reveal that the BMA and EMOS-censored, shifted gamma (CSG) techniques are substantially successful at improving the raw WRF ensemble forecasts; however, each approach improves different aspects of the forecast quality. The BMA method is more accurate, skilful and reliable than the EMOS-CSG method, but has poorer discrimination. Moreover, it has better resolution in predicting the probability of high-precipitation events than the EMOS-CSG method.

Keywords
Bayesian model averaging, ensemble forecasting, ensemble model output statistics, verification

1 INTRODUCTION
Precipitation forecasting is an essential tool for the optimal management of water resources and flood forecasting. The numerical weather prediction (NWP) models play a major role in weather forecasting. It predicts the future state of the weather by the mathematical modelling of the atmosphere’s behaviour based on its current condition. However, the accuracy of the NWP models is still a challenging issue and its improvement is the main goal of the operational prediction centres.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2021 The Authors. Meteorological Applications published by John Wiley & Sons Ltd on behalf of the Royal Meteorological Society.
In weather forecasting by using the NWP models, there are several sources of uncertainty, such as the intrinsic chaotic behaviour of the atmosphere dynamic system, errors in the observational data and initial conditions that are almost impossible to remove (Yang et al., 2012). To quantify these uncertainties, ensemble systems are used. Instead of only one deterministic forecast, an ensemble system is created by several forecasts obtained from perturbation of the initial conditions, physical schemes or dynamic core of the NWP models. Ensemble systems are widely used by the meteorological communities, especially for medium-range weather forecasts as well as for short- and even ultra-short-range weather prediction (Toth and Kalnay, 1993; Houtekamer and Derome, 1995; Molteni et al., 1996; Toth and Kalnay, 1997; Buizza et al., 2005; Charron et al., 2010; Iversen et al., 2011; Bouallégue et al., 2013).

A probabilistic forecasting of flood and extreme precipitation can be created by an ensemble system. However, ensemble forecasting is under-dispersive and not calibrated, especially for meteorological parameters near ground level. To overcome this problem, several statistical methods (Wilks, 2006; Wilks and Hamill, 2007; Williams et al., 2014; Vannitsem et al., 2019; Vannitsem et al., 2020) can be used to post-process the ensemble forecasting. By using the ensemble post-processing methods, the biases in both location and dispersion are eliminated by a historical database of ensemble forecast errors, and then a predictive probability density function (PDF) can be estimated. The most popular ensemble post-processing methods are Bayesian model averaging (BMA) (Raftery et al., 2005; Sloughter et al., 2007, 2010; Soltanzadeh et al., 2011; Baran, 2014) and ensemble model output statistics (EMOS) (Gneiting et al., 2005; Wilks and Hamill, 2007; Thorarinsdottir and Gneiting, 2010; Scheuerer, 2014; Baran and Lerch, 2015; Scheuerer and Hamill, 2015).

In the BMA method (Raftery et al., 2005), a PDF is first fitted to every ensemble member according to the behaviour of the considered weather parameter, and then the predictive PDF is estimated by weighted averaging of the PDFs fitted to the ensemble members. Weights are estimated according to the performance of each ensemble member. In the EMOS method (Gneiting et al., 2005), the distribution parameters are a linear combination of the ensemble members. In both methods, the coefficients are estimated by forecasts and their corresponding observational data in a training period.

Baran et al. (2014) compared the BMA and EMOS post-processing methods for probabilistic forecasting of 2 min temperature and 10 min wind speed of the Hungarian Meteorological Service. They concluded that both post-processing methods improve the calibration of probabilistic forecasts, and the BMA has a slightly better performance than the EMOS method. Schmeits and Kok (2010) compared the raw ensemble, BMA and extended logistic regression (LR) for precipitation probabilistic forecasts in the Netherlands. The results revealed that the difference in the skill of the BMA and LR is not statistically significant. Diaz et al. (2019) used the BMA and EMOS post-processing methods for 2 min temperature probabilistic forecasts. They showed that these post-processing methods result in a significant decrease in the continuous ranked probability scores (CRPS) of probabilistic forecasts.

In the present paper, the 24, 48 and 72 hr ahead probabilistic forecasting of 24 hr accumulated precipitation is evaluated over Iran. The BMA and EMOS methods are used for the post-processing of the ensemble forecasting and estimating the predictive PDF. In the BMA, a gamma-distribution is fitted as the predictive PDF of the individual ensemble members. In the EMOS, the generalized extreme-value (GEV) distribution and the censored, shifted gamma (CSG) distribution are fitted. To implement the BMA and EMOS models, the ensemble-BMA (Fraley et al., 2018) and EMOS (Yuen et al., 2018) R packages are used. The ensemble system has eight members created by running the weather and research forecasting (WRF) model over Iran by different physical configurations during the autumn and winter of 2015–16. The results are verified by rank and the probability integral transform (PIT) histograms, reliability diagram, relative operating characteristic (ROC) diagram, ROC skill score (RSS), Brier score (BS), Brier skill score (BSS) and the CRPS (Wilks, 2011).

The paper is organized as follows. Section 2 introduces the configurations and domains of the WRF model. Section 3 describes the BMA and EMOS methods. The results of the different verification tools are discussed in Section 4. Finally, the summary and conclusions are provided in Section 5.

## 2 | THE ENSEMBLE SYSTEM AND DATA

Forecasts of the WRF model with eight different configurations of 24 hr accumulated precipitation for the 24, 48 and 72 hr lead times are used. The different physical configurations of the WRF model are summarized in Table 1.

The initial conditions of the WRF run are the global forecast system (GFS) forecasts with a 0.5° horizontal resolution. The WRF model is run with two nested domains. The larger domain has a 15 km horizontal resolution and covers an area in the southwest of the Middle East from
The smaller domain has a horizontal resolution of 5 km and covers an area between 23°–41° N and 42°–65° E (the whole of Iran). Initialization time is 1800 UTC between September 1, 2015 and February 28, 2016. The WRF forecasts are bilinearly interpolated to synoptic meteorological stations spread all over Iran. The observational data used in the present study consist of the 24 hr cumulative precipitation

| Table 1 | Different physics combinations of the ensemble members |
|---------|-------------------------------------------------------|
| Ensemble member | Boundary layer | Surface layer | Land surface | Cumulus physics | Microphysics | Short/long wave radiations |
| 1 | Yonsei University scheme (YSU) | Monin–Obukhov–Janjic scheme | Rapid update cycle (RUC) land-surface model | Grell–Freitas ensemble scheme | Ferrier (new Eta) microphysics | Ferrier (new Eta) microphysics |
| 2 | Mellor–Yamada–Janjic turbulent kinetic energy (TKE) scheme | Monin–Obukhov–Janjic scheme | Unified Noah land-surface model | Betts–Miller–Janjic scheme | Ferrier (new Eta) microphysics | Dudhia/Rapid radiative transfer model (RRTM) scheme |
| 3 | Mellor–Yamada–Janjic TKE scheme | Monin–Obukhov–Janjic scheme | Unified Noah land-surface model | Grell–Freitas ensemble scheme | Kessler scheme | Dudhia/RRTM scheme |
| 4 | YSU scheme | MM5 Monin–Obukhov scheme | Unified Noah land-surface model | Kain–Fritsch (new Eta) scheme | WSM 5-class scheme | Dudhia/RRTM scheme |
| 5 | Mellor–Yamada–Janjic TKE scheme | Monin–Obukhov–Janjic scheme | Unified Noah land-surface model | Betts–Miller–Janjic scheme | WSM 5-class scheme | Dudhia/RRTM scheme |
| 6 | Mellor–Yamada–Janjic TKE scheme | Monin–Obukhov–Janjic scheme | Unified Noah land-surface model | Kain–Fritsch (new Eta) scheme | Lin et al. scheme | Goddard/RRTM scheme |
| 7 | MRF | MM5 Monin–Obukhov scheme | Unified Noah land-surface model | Modified Kain–Fritsch scheme | Kessler scheme | Dudhia/RRTM scheme |
| 8 | Mellor–Yamada–Janjic TKE scheme | Monin–Obukhov–Janjic scheme | Unified Noah land-surface model | Kain–Fritsch (new Eta) scheme | WSM 5-class scheme | Dudhia/RRTM scheme |

**Figure 1** Two nested domains of the weather and research forecasting (WRF) model and the locations of all the synoptic meteorological stations used in the study.
measured at 0600 UTC in 349 synoptic meteorological stations spread all over Iran. Figure 1 shows the two nested domains of the WRF model and the locations of all the synoptic meteorological stations used in the study.

In our implementation, as in Gneiting et al. (2005), the training period is a sliding training window of the most recent $N$ days before the forecast for which the ensemble output and the verifying observation were available.

To determine the suitable length of the sliding training period, the CRPS of the probabilistic forecasts is computed for different sliding training periods from 50 to 110 days for the BMA method. The same test data set is used to evaluate all the training periods; that is, from all the available 180 days between September 1, 2015 and February 29, 2016, the first 110 days are considered as the training period and the remaining 70 days as the test period consisting of all the days between December 21, 2015 and February 29, 2016. As shown in Figure 2, the CRPS decreases for the training periods from 50 to 87 days. Since an 87 day window is selected as the training period, the period from November 28, 2015 to February 28, 2016 is considered as the test period, and a sliding window of the 87 previous days is used as the training period in order to fit the PDFs of the post-processing models.

The observational data used in the present study consist of the 24 hr cumulative precipitation measured at 0600 UTC in 349 irregularly spaced synoptic meteorological stations spread across Iran.

3 | METHODS AND VERIFICATION SCORES

3.1 | BMA

The BMA method provides probabilistic forecasts as predictive PDFs of the weather quantity based on an ensemble prediction system (Raftery et al., 2005; Sloughter et al., 2007). The ensemble forecast, $f_1 \cdots f_M$, for precipitation $x$, at a given time and location is considered. Supposing that each ensemble member $f_k$ corresponds to a component PDF $h_k(x|f_k; \theta_k)$, where $\theta_k$ are the parameters to be estimated, the BMA predictive PDF of $x$ is:

$$p(x|f_1 \cdots f_M; \theta_1 \cdots \theta_M) = \sum_{k=1}^{M} w_k h_k(x|f_k; \theta_k),$$

(1)

FIGURE 2  Comparison of the lengths of the training periods for accumulated precipitation forecast with continuous ranked probability score (CRPS)

FIGURE 3  Verification rank histogram of the eight member weather and research forecasting (WRF) ensemble forecasts of 24 hr accumulated precipitation for the 24, 48 and 72 hr lead times, September 1, 2015–February 28, 2016
where \( w_k \) is the weights based on the predictive performance of forecast \( f_k \) during the training period; and \( \sum_{k=1}^{M} w_k = 1 \).

To model precipitation given its properties (positive probability at 0 and skewness), Sloughter et al. (2007) proposed a mixture of a discrete component at 0 and a gamma-distribution as a component PDF; therefore, the BMA model is:

\[
p(x|f_1 \cdots f_M) = \sum_{k=1}^{M} w_k \left\{ p(X = 0|f_k) I[x = 0] + p(X > 0|f_k) g_k(x|f_k) I[x > 0] \right\}, \tag{2}
\]

for the forecast PDF of the cube root of precipitation accumulation, in which \( I \) is the indicator function and \( p(x = 0|f_k) \) is specified by:

---

**FIGURE 4** Probability integral transform (PIT) histograms for the Bayesian model averaging (BMA), ensemble model output statistics-censored, shifted gamma (EMOS-CSG) and ensemble model output statistics-generalized extreme-value (GEV) post-processed forecasts of 24 hr accumulated precipitation for the 24, 48 and 72 hr lead times
logit \( P(x = 0 | f_k) = \log \frac{P(x = 0 | f_k)}{P(x > 0 | f_k)} = a_{0k} + a_{1k} \frac{1}{f_k} + a_{2k} \delta_k, \)

\( \delta_k = 1 \) if \( f_k = 0 \), and 0 otherwise; and \( g_k(x | f_k) \) is a gamma PDF with mean \( \mu_k = b_{0k} + b_{1k} f_k^{1/3} \) and variance \( \sigma_k^2 = c_0 + c_1 f_k \). The parameters \( a_{0k}, a_{1k} \) and \( a_{2k} \) are
estimated by LR, with the yes–no precipitation as the dependent variable and $f_k^{1/3}$ and $\delta_k$ as the two predictor variables.

The parameters $b_{0k}$ and $b_{1k}$ are determined by linear regression of the cube root of the non-zero observation on the corresponding cube root of ensemble members; while $w_k$, $c_0$ and $c_1$ are assessed by the maximum likelihood (Fisher, 1922) technique from the training data.

The cdf function of “ensembleBMA” package in R is used to compute the validation criteria. As mentioned in the package guideline, if the model has been applied to a power transformation of the data, the output is transformed appropriately (Fraley et al., 2018).

3.2 | EMOS

The EMOS predictive PDF is a single parametric distribution in which parameters are functions of the ensemble members. Scheuerer (2014) modelled precipitation by a GEV distribution, which is left-censored at 0, as follows:

$$G(y) = \begin{cases} G(y) & y \geq 0 \\ 0 & y < 0 \end{cases},$$

where $G(y)$ is the cumulative distribution function (CDF) of the GEV distribution with parameters $\mu$, $\sigma$ and $\xi$, which characterize the location, scale and shape of the
GEV. The ensemble members are linked to the mean $m$ and variance $\sigma^2$ of the underlying GEV distribution via the following equations:

$$m = \alpha_0 + \alpha_1 f_1 + \cdots + \alpha_M f_M + \alpha_{M+1} I_{\{f=0\}}, \quad (5)$$

$$\sigma = \beta_0 + \beta_1 MD(f), \quad (6)$$

where $I_{\{f=0\}}$ and $MD(f)$ are the ratio of zero precipitation members and Gini’s mean difference, respectively. The parameters $\alpha_0, \alpha_1, \cdots, \alpha_M+1, \beta_0$ and $\beta_1$ are estimated by minimizing the mean CRPS over the specified training period.

As an alternative to the above-mentioned approach, Scheuerer and Hamill (2015) suggested model precipitation by using the CSG distribution with shape $k$, scale $\theta$ and shift $\delta$ parameters as follows:

$$\tilde{F}_{k\theta\delta}(y) = \begin{cases} F_k \left( \frac{y-\delta}{\theta} \right), & y \geq 0 \\ 0, & 0 < y \end{cases}, \quad (7)$$

where $F_k$ is the CDF of a gamma distribution with the unit scale and shape parameter $k$. Baran and Nemoda (2016) proposed that the ensemble members are linked to the mean $\mu$ and variance $\sigma^2$ of the underlying gamma distribution via the following equations:
FIGURE 8  Relative operating characteristic (ROC) diagrams for the probabilistic forecasts of 24 hr cumulative precipitation using the raw ensemble forecast, and the Bayesian model averaging (BMA), ensemble model output statistics-censored, shifted gamma (EMOS-CSG) and ensemble model output statistics-generalized extreme-value (GEV) methods for the 24, 48 and 72 hr lead times.
\[ \mu = a_0 + a_1 f_1 + \cdots + a_M f_M, \quad (8) \]

\[ \sigma^2 = b_0 + b_1 \tilde{f}. \quad (9) \]

The parameters \( a_0, a_1, \ldots, a_M, b_0 \) and \( b_1 \) are estimated by minimizing the mean CRPS over the specified training period.

### 4 | VERIFICATION RESULTS OF ENSEMBLE CALIBRATION USING THE BMA AND EMOS

In the present research, several verification methods such as the PIT and rank histograms, reliability and ROC diagrams, RSS, BS, BSS and CRPS are used to estimate the performance of the ensemble post-processing methods described in Section 3.2. The test period is from November 28, 2015 to February 28, 2016. In this period, the predictive PDFs of 24 hr accumulated precipitation are estimated by the BMA and EMOS methods, and the observational data are available at 349 irregularly spaced synoptic meteorological stations all over Iran (Figure 2).

Figure 3 shows the verification rank histograms of the raw ensemble forecasts of precipitation for the 24, 48 and 72 hr lead times. They are not uniformly distributed and the ensemble systems are under-dispersive. In other words, the ensemble members are not spread out enough and resemble each other too much (Wilks, 2011).

The PIT diagram (Wilks, 2011) is a common tool to assess the calibration of PDF forecasts. This diagram is the histogram of the predictive CDF evaluated at verifying observations. The PIT histograms of the BMA and two versions of the EMOS predictive PDFs are shown in Figure 4. In comparison with the rank histogram of the raw ensemble forecast (Figure 3), it is obvious that all the implemented post-processing methods can significantly eliminate the under-dispersion exhibited by the raw ensemble forecast. However, the BMA and EMOS-CSG models are more uniformed than the EMOS-GEV model. The results of Kolmogorov–Smirnov tests for uniformity of the PIT values are presented in Table 2. The BMA models produce the best PIT histograms for all the lead times. However, the under-representation of the last bin in the PIT histograms shows that the BMA model has a positive bias in predicting 0 values. This is because most of the observations are 0, and in order to plot PIT, the

| Threshold | 24 hr | 48 hr | 72 hr |
|-----------|-------|-------|-------|
|           | AUC   | 95% CI| AUC   | 95% CI| AUC   | 95% CI|
| (a) BMA   |       |       |       |       |       |       |
| > 0.1 mm  | 0.939 | (0.935, 0.943) | 0.927 | (0.923, 0.932) | 0.901 | (0.895, 0.906) |
| > 2.5 mm  | 0.953 | (0.950, 0.957) | 0.944 | (0.939, 0.948) | 0.921 | (0.916, 0.927) |
| > 5 mm    | 0.956 | (0.952, 0.960) | 0.947 | (0.942, 0.951) | 0.928 | (0.922, 0.934) |
| > 10 mm   | 0.962 | (0.957, 0.966) | 0.958 | (0.953, 0.963) | 0.943 | (0.936, 0.949) |
| > 15 mm   | 0.966 | (0.959, 0.972) | 0.967 | (0.963, 0.972) | 0.955 | (0.948, 0.963) |
| > 25 mm   | 0.977 | (0.969, 0.984) | 0.975 | (0.968, 0.982) | 0.966 | (0.958, 0.975) |
| (b) EMOS-CSG |       |       |       |       |       |       |
| > 0.1 mm  | 0.938 | (0.934, 0.942) | 0.933 | (0.929, 0.937) | 0.908 | (0.902, 0.913) |
| > 2.5 mm  | 0.952 | (0.948, 0.955) | 0.946 | (0.942, 0.950) | 0.931 | (0.926, 0.936) |
| > 5 mm    | 0.954 | (0.950, 0.958) | 0.953 | (0.949, 0.956) | 0.941 | (0.936, 0.946) |
| > 10 mm   | 0.957 | (0.951, 0.962) | 0.960 | (0.955, 0.965) | 0.951 | (0.945, 0.957) |
| > 15 mm   | 0.960 | (0.953, 0.967) | 0.967 | (0.962, 0.972) | 0.958 | (0.951, 0.965) |
| > 25 mm   | 0.971 | (0.961, 0.980) | 0.977 | (0.971, 0.984) | 0.972 | (0.965, 0.980) |
| (c) EMOS-GEV |       |       |       |       |       |       |
| > 0.1 mm  | 0.909 | (0.903, 0.914) | 0.890 | (0.884, 0.895) | 0.844 | (0.837, 0.851) |
| > 2.5 mm  | 0.936 | (0.931, 0.941) | 0.933 | (0.928, 0.938) | 0.912 | (0.906, 0.918) |
| > 5 mm    | 0.941 | (0.935, 0.946) | 0.942 | (0.937, 0.947) | 0.919 | (0.912, 0.926) |
| > 10 mm   | 0.948 | (0.942, 0.955) | 0.953 | (0.947, 0.959) | 0.936 | (0.928, 0.944) |
| > 15 mm   | 0.948 | (0.939, 0.957) | 0.962 | (0.955, 0.968) | 0.947 | (0.938, 0.957) |
| > 25 mm   | 0.967 | (0.957, 0.977) | 0.971 | (0.963, 0.980) | 0.952 | (0.937, 0.967) |
value of the CDF at 0 should be chosen randomly from the interval between 0 and the value of the CDF at 0. For the ideal model, the CDF at 0 (when observation is 0) is expected to be near 1 or the highest frequency of CDF values at 0 be in the last bin near 1. In the BMA model, the highest frequency of CDF values at 0 is not for the last bin (between 0.9 and 1). The EMOS-CSG models are also relatively good for the 48 and 72 hr forecasts, while the fit of the EMOS-GEV model is rather poor for the 48 and 72 hr forecasts.

The reliability diagram (Wilks, 2011) is a common diagnostic graph to assess the reliability and resolution of a probabilistic forecast. In the reliability diagram, the frequency of an observed event is plotted against its forecast probability. In this diagram, the curve of a well-calibrated probabilistic forecast is close to the bisector and far from the “no resolution” line. The democratic voting method
(Wilks, 2011) is used to determine the probabilistic forecast obtained from the raw ensemble.

Figures 5–7 show the reliability diagram of the probabilistic forecasts of 24 hr accumulated precipitation estimated by the democratic voting (raw ensemble forecast), BMA, EMOS-CSG and EMOS-GEV methods for the 24, 48 and 72 hr lead times in different thresholds of 0.1, 2.5, 5, 10, 15 and 25 mm. As shown, the predictive PDFs estimated by the raw ensemble and the EMOS-GEV method are not well calibrated and are close to the “no resolution” line, especially in high thresholds. The reliability curve gets close to the bisector by using the BMA and EMOS-CSG methods on the thresholds of 0.1, 2.5, 5 and 10 mm. In other words, the reliability of light and medium rainfalls for the 24 hr accumulative rainfall is considerable. In the thresholds of 15 and 25 mm, the reliability curve of the BMA method is better calibrated than that of the EMOS-CSG method. The BMA and EMOS-CSG methods have higher resolution than the raw output of the ensemble system and the EMOS-GEV method in all the lead times. Generally, the BMA method used for the 24, 48 and 72 hr forecasts provides more reliable probabilistic forecasts than the EMOS-CSG and EMOS-GEV methods and the raw ensemble forecast over studied domain.

The ROC diagram (Wilks, 2011) measures the ability of the forecast to discriminate between events and non-events and demonstrates the hit rate (H) versus the false-alarm rate (F). Corresponding pairs (F, H) are depicted on the graph and connected to each other by several lines. The pairs (0, 0) and (1, 1) mean “never forecasting” and “always forecasting”, respectively. In the good discrimination forecasts, the ROC curves approach the top-left corner; however, in poor discrimination ones, these curves get close to the F = H diameter (Wilks, 2006).

The ROC diagrams of the calibrated forecasts using the BMA, EMOS-CSG, EMOS-GEV methods, and the raw ensemble forecast for the 24, 48 and 72 hr lead times in different thresholds are shown in Figure 8. Table 3 shows the 95% confidence intervals for the area under the relative operating characteristic curve (AUC) for BMA, EMOS-CSG and EMOS-GEV methods. The confidence intervals are determined by bootstrapping methods. The ROC diagrams show that the BMA and EMOS methods, for all the lead times in different thresholds, have better discrimination than the raw ensemble forecasts. Hence, the order of the methods from good to poor discrimination is EMOS-CSG, BMA and EMOS-GEV, respectively. Table 3 shows that discrimination of EMOS-CSG is significantly poorer than the BMA and EMOS-CSG.

The RSS compares the area under the ROC curve of a probabilistic forecast with the area under the ROC curve of a reference forecast. RSS is defined as:

$$RSS = \frac{A - A(\text{reference})}{A(\text{perfect}) - A(\text{reference})},$$  \hspace{1cm} (10)
where A is the area under the ROC curve of the probabilistic forecast; and A (reference) is the area under the reference forecast (the raw ensemble forecast the present study). A (perfect) is the total area of the square and its value is 1.

The RSS values computed for the BMA and EMOS methods in different thresholds for the 24, 48 and 72 hr lead times are shown in Figure 9. The higher RSS shows better discrimination in comparison with the raw ensemble forecasts. The RSS values in all thresholds for the 24, 48 and 72 hr forecasts in the BMA and EMOS-CSG methods are higher than the RSS values of the EMOS-GEV method. The RSS values are not high in the low thresholds because the raw ensemble forecasts have good discrimination in these thresholds, as shown in Figure 8.

It should be noted that the ROC summarized over all stations and the verification period. Therefore, the ROC and RSS results may suffer from the Hamill and Jarus (2007) effect.

The BS (Brier, 1950) is a common verification method for the probabilistic forecasts of binary events (Wilks, 2011). This score is the mean square of the differences between forecast probabilities and their corresponding binary events, and is calculated as:

$$BS = \frac{1}{N} \sum_{n=1}^{N} (y_n - o_n)^2,$$

where N is the number of events; y_n is the forecast probability; and o_n is the binary view of the nth event. If the event is observed, o_n = 1, otherwise o_n = 0. A lower BS is desirable, so BS = 0 a perfect forecast.

The BSS assesses the performance of a probabilistic forecast compared with a reference forecast:

$$BSS = \frac{BS - BS(\text{reference})}{BS(\text{perfect}) - BS(\text{reference})} = 1 - \frac{BS}{BS(\text{reference})},$$

where BS(perfect) = 0 and the raw ensemble forecast is considered as the reference forecast in the present study. The BSS of a perfect probabilistic forecast is 1.

The results of BS and BSS values for the probabilistic forecasts of 24 hr accumulative precipitation in different thresholds for the 24, 48 and 72 hr lead times are presented in Figures 10 and 11. Table 4 also shows the BS values and the 95% confidence interval for BS, which are

| Threshold   | 24 hr BS | 24 hr 95% CI | 48 hr BS | 48 hr 95% CI | 72 hr BS | 72 hr 95% CI |
|-------------|----------|--------------|----------|--------------|----------|--------------|
|             |          |              |          |              |          |              |
| (a) BMA     |          |              |          |              |          |              |
| > 0.1 mm    | 0.077    | (0.076, 0.080) | 0.084    | (0.083, 0.087) | 0.096    | (0.095, 0.098) |
| > 2.5 mm    | 0.055    | (0.054, 0.056) | 0.059    | (0.058, 0.061) | 0.066    | (0.065, 0.067) |
| > 5 mm      | 0.043    | (0.042, 0.045) | 0.047    | (0.046, 0.048) | 0.051    | (0.050, 0.052) |
| > 10 mm     | 0.028    | (0.027, 0.029) | 0.029    | (0.028, 0.030) | 0.032    | (0.030, 0.034) |
| > 15 mm     | 0.018    | (0.018, 0.019) | 0.019    | (0.017, 0.019) | 0.020    | (0.019, 0.021) |
| > 25 mm     | 0.009    | (0.008, 0.009) | 0.009    | (0.009, 0.010) | 0.010    | (0.009, 0.011) |
| (b) EMOS-CSG|          |              |          |              |          |              |
| > 0.1 mm    | 0.079    | (0.078, 0.081) | 0.085    | (0.083, 0.087) | 0.095    | (0.092, 0.097) |
| > 2.5 mm    | 0.054    | (0.052, 0.056) | 0.058    | (0.057, 0.059) | 0.062    | (0.061, 0.063) |
| > 5 mm      | 0.043    | (0.042, 0.045) | 0.046    | (0.044, 0.046) | 0.048    | (0.046, 0.050) |
| > 10 mm     | 0.028    | (0.027, 0.030) | 0.029    | (0.027, 0.030) | 0.030    | (0.029, 0.031) |
| > 15 mm     | 0.018    | (0.017, 0.019) | 0.019    | (0.018, 0.020) | 0.019    | (0.019, 0.021) |
| > 25 mm     | 0.009    | (0.008, 0.010) | 0.009    | (0.009, 0.010) | 0.009    | (0.009, 0.010) |
| (c) EMOS-GEV|          |              |          |              |          |              |
| > 0.1 mm    | 0.095    | (0.094, 0.097) | 0.102    | (0.101, 0.105) | 0.122    | (0.121, 0.123) |
| > 2.5 mm    | 0.058    | (0.056, 0.059) | 0.060    | (0.058, 0.062) | 0.069    | (0.068, 0.070) |
| > 5 mm      | 0.046    | (0.044, 0.047) | 0.047    | (0.045, 0.048) | 0.052    | (0.050, 0.054) |
| > 10 mm     | 0.030    | (0.029, 0.031) | 0.029    | (0.028, 0.030) | 0.032    | (0.031, 0.034) |
| > 15 mm     | 0.019    | (0.019, 0.021) | 0.019    | (0.017, 0.020) | 0.021    | (0.020, 0.022) |
| > 25 mm     | 0.009    | (0.008, 0.010) | 0.009    | (0.008, 0.010) | 0.010    | (0.010, 0.011) |
obtained by a bootstrapping procedure. The BS values of the BMA and EMOS-CSG are significantly less than the BS values of the EMOS-GEV and raw forecast for all the lead times, but the BS values of the BMA and EMOS-CSG methods are not significantly different. In addition, the BS values in all the forecast lead times are almost the same. As shown in Figure 11, the BSS values of the BMA, EMOS-CSG and EMOS-GEV methods for the 24, 48 and 72 hr forecasts are positive, and thus more calibrated than the raw ensemble forecasts. As indicated, the BMA and EMOS-CSG method are more skilful than the EMOS-GEV methods, especially in the 24 hr lead time forecasts.

The CRPS is calculated to assess the overall performance of the probabilistic forecasts. The CRPS formula is as follows:

$$CRPS(F(y), x) = \int_{-\infty}^{+\infty} \{F(y) - I_{y \geq x}\}^2 dy,$$ (14)

where $F(y)$ is the CDF of the forecast parameter $y$; and $I_{y \geq x}$ is a heaviside function that equals 1 when $y \geq x$, and 0 otherwise.

Table 5 gives the CRPS values of the probabilistic forecasts estimated by the BMA and EMOS models and the raw ensemble and 95% bootstrapping confidence intervals for the CRPS values. It shows that all three methods can improve the probabilistic forecasts in comparison with the raw ensemble forecasts. The CRPS values of the BMA and EMOS-CSG methods are significantly less than of the EMOS-GEV in all the lead times.

|          | BMA 95% CI | EMOS-CSG 95% CI | EMOS-GEV 95% CI | Raw ensemble |
|----------|------------|-----------------|-----------------|--------------|
| 24 hr    | 0.836 (0.785, 0.009) | 0.827934 (0.797, 0.861) | 0.895739 (0.862, 0.934) | 0.994 |
| 48 hr    | 0.868 (0.831, 0.913) | 0.863548 (0.830, 0.899) | 0.897305 (0.863, 0.931) | 1.020 |
| 72 hr    | 0.923 (0.892, 0.963) | 0.905009 (0.870, 0.940) | 1.013482 (0.973, 1.054) | 1.076 |

5 | CONCLUSIONS

In the present paper different versions of the ensemble model output statistics (EMOS) and Bayesian model averaging (BMA) methods are compared in order to improve the calibration of precipitation forecasts of the weather and research forecasting (WRF) system in Iran. First, the weaknesses of the WRF raw ensemble system were demonstrated as being under-dispersive and thus uncalibrated. Second, the BMA and two versions of the EMOS were applied and some standard measures for comparing them were computed.

Overall, the results showed that the BMA and EMOS-censored, shifted gamma (CSG) techniques are significantly successful at improving the raw WRF ensemble forecasts. However, each approach improves different aspects of the forecast quality. The BMA method is more accurate, skilful and reliable than the EMOS-CSG method, but has poorer discrimination. Moreover, the BMA method gave better resolution in predicting the probability of high-precipitation events than the EMOS-CSG method.

Future studies could investigate the manner in which modern post-processing techniques such as machine learning or blending approaches are performed. As discussed by Vannitsem et al. (2020), post-processing methods such as the BMA and EMOS can be easily implemented, but neural networks or analogues are much more flexible. Therefore, another possibility is to assess the trade-off between complexity and flexibility.

ORCID

Zohreh Javanshiri  https://orcid.org/0000-0002-3542-1310
Maede Fathi  https://orcid.org/0000-0002-5923-5782

REFERENCES

Baran, S. (2014) Probabilistic wind speed forecasting using Bayesian model averaging with truncated normal components. *Computational Statistics and Data Analysis*, 75, 227–238. https://doi.org/10.1016/j.csda.2014.02.013.

Baran, S., Horanyi, A. and Nemoda, D. (2014) Comparison of the BMA and EMOS statistical methods in calibrating temperature and wind speed forecast ensembles. *Idojárás*, 118(3), 217–241.

Baran, S. and Lerch, S. (2015) Log-normal distribution based EMOS models for probabilistic wind speed forecasting. *Quarterly Journal of the Royal Meteorological Society*, 141, 2289–2299.

Baran, S. and Nemoda, D. (2016) Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting. *Environmetrics*, 27(5), 280–292.

Bouallégue, B.Z., Theis, S. and Gebhardt, C. (2013) Enhancing COSMO-DE ensemble forecasts by inexpensive techniques. *Meteorologische Zeitschrift*, 22, 49–59.
Brier, G.W. (1950) Verification of forecasts expressed in terms of probabilities. *Monthly Weather Review*, 78, 1–3.

Buizza, R., Houtekamer, P.L., Toth, Z., Pellerin, G., Wei, M. and Zhu, Y. (2005) A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Monthly Weather Review*, 133, 1076–1097.

Charlon, M., Pellerin, G., Spacek, L., Houtekamer, P.L., Gagnon, N., Mitchell, H.L. and Michelin, L. (2010) Toward random sampling of model error in the Canadian ensemble prediction system. *Monthly Weather Review*, 138, 1877–1901.

Díaz, M., Nicolis, O., Marin, J.C. and Baran, S. (2019) Statistical post-processing of ensemble forecasts of temperature in Santiago de Chile. *Meteorological Application*, 27(1), e1818. https://doi.org/10.1002/met.1818.

Fisher, R.A. (1922) On the mathematical foundations of theoretical statistics. *Philosophical Transactions. Royal Society of London*, 222A, 309–368.

Fraley, C., Raftery, A.E., Sloughter, J.M. and Gneiting, T. (2018) ensembleBMA: Probabilistic Forecasting using Ensembles and Bayesian Model Averaging. University of Washington. R package version 5.1.5. Available at: https://CRAN.R-project.org/package=ensembleBMA [Accessed 18th January 2018].

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

Hamill, T.M. and Jarus, J. (2007) Measuring forecast skill: is it real skill or is it the varying climatology? *Quarterly Journal of the Royal Meteorological Society*, 132, 2905–2923.

Houtekamer, P.L. and Derome, J. (1995) Methods for ensemble prediction. *Monthly Weather Review*, 123, 2181–2196.

Iversen, T., Deckmyn, A., Santos, C., Sattler, K., Bremnes, J.B., Feddersen, H. and Frogner, I.-L. (2011) Evaluation of “GLAMEPS”—A proposed multimodel EPS for short range forecasting. *Tellus A: Dynamic Meteorology and Oceanography*, 63, 513–530.

Molteni, F., Buizza, R., Palmer, T. and Petroliaqis, T. (1996) The ECMWF ensemble prediction system: methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, 125, 73–119.

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

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(11), 4199–4211.

Sloughter, J.M., Gneiting, T. and Raftery, A.E. (2010) Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *Journal of the American Statistical Association*, 105, 25–35.

Sloughter, J.M., Raftery, A.E., Gneiting, T. and FRALEY, C. (2007) Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review*, 135, 3209–3220.

Soltanzadeh, I., Azadi, M. and Vakili, G.A. (2011) Using Bayesian model averaging (BMA) to calibrate probabilistic surface temperature forecasts over Iran. *Annales de Geophysique*, 29, 1295–1303.

Thorarinsdottir, T.L. and Gneiting, T. (2010) Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173, 371–388.

Toth, Z. and Kalnay, E. (1993) Ensemble forecasting at NMC: the generation of perturbations. *Bulletin of the American Meteorological Society*, 74, 2317–2330.

Toth, Z. and Kalnay, E. (1997) Ensemble forecasting at NCEP and the breeding method. *Monthly Weather Review*, 125, 3297–3319.

Vannitsen, S., Bremnes, J.B., Demaeyer, J., Evans, G.R., Flowerdew, J., Henri, S., Lerch, S., Roberts, N., Theis, S., Atencia, A., Zied Ben Bouallégue, Bhend, J., Dabernig, M., De Cruz, L., Hieto, L., Mestre, O., Moret, L., Iris Odak Plenkovici, Schmeits, M., Taillardat, M., Van den Bergh, J., Van Schaeybroeck, B., Whan, K. and Ylhaiasi, J. (2020) Statistical post-processing for weather forecasts—review, challenges and avenues in a big data world, 2020. [Preprint] Available at: http://arxiv.org/abs/2004.06582 [Accessed 19th November 2020].

Vannitsen, S., Wilks, D.S. and Messner, J. (2019) *Statistical Post-processing of Ensemble Forecasts*. Amsterdam, Netherlands: Elsevier.

Wilks, D.S. (2006) Comparison of ensemble MOS methods in the Lorenz 96 setting. *Meteorological Applications*, 13, 243–256.

Wilks, D.S. (2011) *Statistical Methods in the Atmospheric Sciences*, 3rd edition. New York: Academic Press.

Wilks, D.S. and Hamill, T.M. (2007) Comparison of ensemble-MOS methods using GFS reforecasts. *Monthly Weather Review*, 135, 2379–2390.

Williams, R.M., Ferro, C.A.T. and Kwasniok, F. (2014) A comparison of ensemble post-processing methods for extreme events. *Quarterly Journal of the Royal Meteorological Society*, 140, 1112–1120. https://doi.org/10.1002/qj.2198.

Yang, C., Yan, Z. and Shao, Y. (2012) Probabilistic precipitation forecasting based on ensemble output using generalized additive models and Bayesian model averaging. *Acta Meteorologica Sinica*, 26, 1–12. https://doi.org/10.1007/s13351-012-0101-8.

Yuen, R.A., Baran, S., Fraley, C., Gneiting, T., Lerch, S., Scheuerer, M. and Thorarinsdottir, T.L. (2018) ensembleBMA: Ensemble Model Output Statistics. R package version 0.8.2. Available at: https://cran.r-project.org/web/packages/ensembleBMA/ensembleBMA.pdf [Accessed 22nd March 2018].