Performance evaluation of sub-daily ensemble precipitation forecasts

Adeleh Saedi1 | Bahram Saghafian1 | Saber Moazami2 | Saleh Aminyavari3

1Department of Civil Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran
2Department of Civil Engineering, Islamshahr Branch, Islamic Azad University, Islamshahr, Iran
3Department of Civil Engineering, Chalous Branch, Islamic Azad University, Chalous, Iran

Correspondence
Bahram Saghafian, Department of Civil Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran.
Email: b.saghafian@gmail.com

Abstract
Nowadays, major advances have been made in meteorological forecasts. For instance, ensemble forecast systems have been developed to quantify prediction uncertainty. In this research, sub-daily ensemble precipitation forecasts of five THORPEX Interactive Grand Global Ensemble (TIGGE) models from 2014 to 2018 were evaluated in 10 major basins located in north and west Iran. Furthermore, Bayesian model averaging (BMA) was used to combine five prediction models and construct a grand ensemble. The results indicate that the models had the best performance in the Karun and Western Border basins to the southwest and west, with average performance in the Sefidrood basin in the north. In terms of the prediction of precipitation depth, the European Centre for Medium-Range Weather Forecasts (ECMWF) and the UK Met Office (UKMO) models, and in terms of prediction of precipitation occurrence and non-occurrence, the National Centers for Environmental Prediction (NCEP) model, performed best. Overall, the Japan Meteorological Agency (JMA) and the China Meteorological Administration (CMA) models acquired medium scores. The BMA technique greatly improved the probability forecasts, reducing uncertainties in the numerical models. Moreover, the models' forecasts were weaker with a 6 hr lead time compared with those with 24 hr, which may be attributable to the inaccurate detection of the initiation time of precipitation by the models. In addition, the performance of the UKMO (ECMWF) models with increasing basin elevation increased (decreased), while all models better forecasted precipitation in wet years/seasons than they did in dry years/seasons. Overall, the evaluations showed that the ECMWF, UKMO and NCEP models performed well in the majority of the northern and western basins of Iran.

KEYWORDS
BMA, ensemble forecast, Iran, sub-daily, TIGGE

1 | INTRODUCTION

Meteorological forecasts offer new opportunities in hydrological applications, flood forecasting in particular. Major advances have been made at improving meteorological forecasts, including the ensemble forecasts referring to several individual forecasts developed by different physical parameterization or different initial conditions (Swinbank et al., 2016). The THORPEX Interactive Grand Global Ensemble (TIGGE) database,
which includes the ensemble forecast of 11 important meteorological centres, has provided a promising opportunity for users around the world to predict precipitation in their region of interest. One practical application of the TIGGE is to develop an early-warning product for severe weather events based on operational medium-range ensemble forecasts from four of the numerical weather centres: the European Centre for Medium-Range Weather Forecasts (ECMWF), UK Met Office (UKMO), National Centers for Environmental Prediction (NCEP), and Japan Meteorological Agency (JMA). The forecast probability of occurrence of severe weather events is defined based on each model’s climatological probabilistic density function. Several case studies using this product have demonstrated its ability to predict successfully severe events such as the heatwave in Russia in 2010, the Pakistan flood in 2010 and Hurricane Sandy in 2012 (Matsueda and Nakazawa, 2015).

Extensive research on the evaluation of ensemble precipitation forecasts has been carried out worldwide. For example, the results of ensemble forecast evaluation of five models including the ECMWF, UKMO, NCEP, JMA and China Meteorological Administration (CMA) within the TIGGE database over the Huai River basin in China indicated that the ECMWF and JMA had the best performance, while the CMA presented the poorest performance (Tao et al., 2014). In another research, the results of evaluating an ensemble precipitation forecast of five models including the ECMWF, Korea Meteorological Administration (KMA), NCEP, CMA and JMA within the TIGGE database during the rainy season before the summer in south China between 2013 and 2015 indicated that the ensemble forecast overestimated 12 hr precipitation events by 0.5–30 mm, and under-estimated precipitation events by > 30 mm. Furthermore, the JMA forecasts were found to be the poorest of the five tested models (Huang and Luo, 2017). In India, the precipitation forecasts of seven numerical models within the TIGGE database for tropical cyclones were evaluated along the coasts of the northern Indian Ocean. The results indicated that the ECMWF, UKMO and NCEP model estimations were closer to observations (Bhomia et al., 2017). In Western Africa, 2008–2012, the ECMWF and UKMO forecasts were the best among the seven TIGGE meteorological centres (Louvet et al., 2016).

For the Northern Hemisphere, among the six TIGGE meteorological centres over the summers of the period 2008–2012, the ECMWF had the best performance, while overall forecast scores were better in the middle of the Northern Hemisphere compared with the tropical regions of the Northern Hemisphere (Su et al., 2014). In Iran, assessments of 1–3 day lead times of the ECMWF, UKMO and NCEP forecasts against observed precipitations in 13 synoptic stations across eight different precipitation regimes over the period 2008–2016 revealed that the ECMWF had the best performance, while the UKMO and NCEP performed better in mountainous regions and along the coasts of Persian Gulf, respectively, whereas the model skill decreased with an increase in forecast lead time (Aminyavari et al., 2018).

The construction of a grand ensemble by combining several ensemble forecasting models is expected to improve the probabilistic skills of forecasts of severe or extreme precipitation and surface temperature events. Indeed, the grand ensemble provides more reliable forecasts compared with single-centre ensembles (Matsueda and Nakazawa, 2015). Various approaches have been devised for combining the ensemble forecasts of several models, the BMA being one of the most widely used. For example, the construction of a grand ensemble using the BMA via combining the UKMO, NCEP, ECMWF and CMA models showed that the UKMO and ECMWF had greater skills, while the grand ensemble of the four centres was recommended for heavy precipitation forecast in the study region (Liu and Xie, 2014). Qu et al. (2017) constructed a BMA grand ensemble, and through a distributed hydrologic model, they forecasted river discharge associated with 24–120 hr lead times and concluded that the BMA post-processing resulted in more reliable probabilistic discharge forecast compared with relying on raw ensemble numerical precipitation forecasts. Overall, the results of most studies suggest that the BMA can enhance the probabilistic forecast skills (Vogel et al., 2018; Zhong et al., 2018).

As vast areas of Iran are subject to flooding, a flood warning is a necessary tool for the prevention of death and sometimes damage to properties. With comprehensive performance assessment of the most widely available global numerical weather prediction (NWP) models on sub-daily scale in Iranian settings, the development of flood early-warning systems becomes feasible. So far, no study has been carried out to evaluate the NWP models on sub-daily scale in Iran, and this is therefore the main innovation of the present study.

The aim of the research is to evaluate the forecasts of the ECMWF, UKMO, NCEP, CMA and JMA (ensemble) numerical models extracted from the TIGGE database across 10 major basins in Iran. The assessment period covers 2014–2018, while the models’ skills were examined based on deterministic, dichotomous and probabilistic assessment techniques. Furthermore, a grand ensemble forecast is constructed using the BMA technique and then compared with those of five individual models.
2 | METHODOLOGY

2.1 | Data

First, observed precipitation data of 153 synoptic stations were collected. These stations are located within 10 Iranian major basins, whose general characteristics are presented in Table 1. Figure 1 shows the location of the study basins, overlaid by the selected stations. The selected basins generally receive higher precipitation among all basins in Iran and provide most of the surface water of the country. Furthermore, major floods in Iran have been reported in the selected basins. Furthermore, forecast data of the ECMWF, UKMO, NCEP, CMA and JMA models were extracted from the TIGGE database on a 6, 12, 18 and 24 hr time scale over the period 2014–2018. The characteristics of the five studied numerical models are provided in Table 2. The start time of forecasts was chosen as 0000 UTC.

The mean observed annual precipitation of the selected synoptic stations over the study period was interpolated using ordinary kriging with an exponential theoretical semi-variogram. As shown in Figure 2, maximum precipitation occurred along the Caspian Sea in the north, that is, in basins 2, 3 and 9, while the minimum precipitation was observed in basins 1, 7 and 8.

Furthermore, to study the climatic conditions of the study area, the standard precipitation index (SPI) time series were determined for all basins. First, station time series of monthly observed precipitation were averaged over each basin using Thiessen polygons. The mean monthly precipitation in each basin was then inputted into the SPEI package (Beguería and Vicente-Serrano, 2013) to determine the basin average SPI time series. The package assumes a gamma statistical distribution to fit the monthly precipitation time series.

2.2 | Evaluation techniques

The precipitation forecasts were evaluated using deterministic, dichotomous and probabilistic techniques. In the deterministic category, the Pearson correlation coefficient was used to measure the correspondence of forecasts with the observations, whereas relative root mean square error (RRMSE) was adopted to measure the error magnitude of each model. Note that in deterministic assessment, only the days with observed precipitations > 1 mm were taken into account, while the days with precipitation < 1 mm were ignored.

A dichotomous yes/no was used to measure the competence of numerical models in predicting the occurrence or non-occurrence of precipitation. A “yes” stands for the case that precipitation occurred in reality; a “no” refers to a case in which the forecasted precipitation event did not take place in reality (Wilks, 2011). In this technique, first, the contingency table (Table 3) was constructed based on the binary values of observations and forecasts. Using the bias score and equal threat score (ETS), the performance of numerical forecast models was then assessed. The bias score criterion refers to the ratio of the number of predicted events to the number of observed events. If bias score > 1, it suggests that the model overestimates the number of precipitation events, while a value < 1 represents underestimation. The ETS criterion is useful for evaluating the quality of forecasted occurrence or non-occurrence of precipitation by numerical models; it states to what extent the model has been successful in predicting the occurrence or non-occurrence of precipitation. Note that in deterministic and dichotomous assessment, the mean ensemble forecast members are used.

| ID | Basin | Area (km²) | Centre (longitude, latitude) | Mean annual precipitation in the past 50 years (mm) | Mean elevation (masl) |
|----|-------|------------|-----------------------------|---------------------------------|----------------------|
| 1  | Aras  | 39,778     | 46° 21’ E, 38° 67’ N        | 271.7                           | 1,198                |
| 2  | Haraz | 18,774     | 53° 62’ E, 36° 48’ N        | 550.9                           | 472                  |
| 3  | Between Haraz and Sefidrood | 10,893 | 50° 76’ E, 36° 52’ N        | 711.3                           | 689                  |
| 4  | Karkheh | 51,912     | 47° 69’ E, 33° 84’ N        | 419                             | 1,152                |
| 5  | Karun | 66,675     | 49° 29’ E, 32° 59’ N        | 632.2                           | 1,210                |
| 6  | Western border | 39,297 | 45° 90’ E, 34° 23’ N        | 457                             | 1,019                |
| 7  | Urmia | 51,761     | 45° 71’ E, 37° 57’ N        | 334.2                           | 1,385                |
| 8  | Sefidrood | 59,194   | 47° 31’ E, 36° 63’ N        | 350.4                           | 1,273                |
| 9  | Talesh | 7,039      | 49° 01’ E, 37° 48’ N        | 859.6                           | 217                  |
| 10 | Zohreh and Garahi | 40,820 | 49° 95’ E, 30° 01’ N        | 417.4                           | 398                  |
In the probabilistic assessment, first, the probability forecasts of each model were constructed by using a gamma distribution. The reason for choosing a gamma probability density function (PDF) is that precipitation distribution is highly skewed and the gamma fits best (Liu and Xie, 2014). For probabilistic assessment, Brier score (BS), which is a measure of the mean squared error of the probability forecasts, was adopted. The BS is defined as the sum of the three terms including reliability, resolution and uncertainty (Sene, 2010). Reliability refers to how close the forecast probabilities are to the true probabilities, given that forecast. Resolution is related to the ability to forecast to discern the occurrence or non-occurrence of the event (Sene, 2010; Wilks, 2011). Typically, the Brier skills score (BSS), which measures the forecast BS in relation to the reference BS, is used. In this research, the reference BS was generated from climatological forecasts. Further, relative operating characteristic (ROC) diagrams were used for a visual representation of the ability of the forecast models in the resolution of occurrence or non-occurrence of the event. In this diagram, the probability of detection (POD or hit rate) and false alarm ratio (FAR) are plotted against each other, and the yes/no decisions are developed with the aid of a set of ascending thresholds (in this research, 0.1–0.9 with 0.1 intervals). A perfect forecast has only 0 and 1 probabilities; in this state, only one probability threshold is calculated in the binary table. Based on this table, FAR = 0 and hit rate = 1. Furthermore, the ROC diagram

| Center                                      | Base time (UTC) | Ensemble members | Horizontal resolution archived | Forecast length (days) |
|---------------------------------------------|-----------------|------------------|-------------------------------|------------------------|
| European Centre for Medium-Range Weather Forecasts (ECMWF) | 0000/1200       | 50 + 1           | N320 (~0.28°), N160 (~0.56°) | 0–10, 10–15            |
| National Centers for Environmental Prediction (NCEP) | 0000/0600/1200/1800 | 20 + 1          | 1.0° × 1.0°                  | 0–16                   |
| UK Met Office (UKMO)                       | 0000/1200       | 23 + 1, 11 + 1, 17 + 1 | 0.83° × 0.56°                | 0–15                   |
| Japan Meteorological Agency (JMA)          | 0000/1200       | 26 + 1           | 1.25° × 1.25°                | 0–9                    |
| China Meteorological Administration (CMA)  | 0000/1200       | 14 + 1           | 0.56° × 0.56°                | 0–10                   |

Note: The UKMO changed its number of ensemble members several times: by July 2014, there were 23 members; between July 2014 and July 2017, there were 11 members; and from July 2017 to the present, they are 17 members.
includes two linear parts that intersect each other in the top left corner of the diagram. The area under the ROC curve is also considered as an assessment score, in which higher values represent more accurate forecast (Wilks, 2011). Table 4 provides a summary of the assessment criteria used in this research.

Note that the assessments were performed in two separate modes (owing to page limitations, only the results of the second state are reported). In the first mode, cell-based forecasted precipitations were downscaled into the location of each synoptic station via inverse distance weighting (IDW) interpolation method and then evaluated against observed precipitation. The mean of the assessment criteria over each basin was then examined. In the second state, following Tao et al. (2014) and Ran et al. (2018), the spatial average of the observed and forecasted precipitations was determined over each basin via Thiessen polygons, and model accuracy was then examined.

2.3 | Grand ensemble

A grand ensemble was constructed by combining the five numerical models studied here via the BMA and then
The ensemble BMA package in R software was used to generate the grand ensemble (Fraley et al., 2007). In this package, the precipitation forecasts and their corresponding observations were introduced, and based on the PDF skill of each model, a weighted co-efficient is assigned to that model in each precipitation event. By calculating the mean weighted co-efficients throughout the entire study period, the weight of each model was obtained. Eventually, the prediction values of each model were multiplied by their corresponding weight and then summed to form the grand ensemble.

3 | RESULTS AND DISCUSSION

3.1 | Annual and seasonal precipitation

At first, the performance of the NWP models was analysed based on basin elevation and flood seasonality. For this purpose, total annual precipitation was determined for all basins over 2014–2017. The basins’ average historic monthly precipitations were transformed into SPI time series and the drought/wet status of the studied four year period was determined (Figure 3).

### TABLE 4 Summary of the assessment criteria used in the research

| Verification measure                     | Formula                                                                 | Description                                                              | Perfect/no skill |
|------------------------------------------|------------------------------------------------------------------------|--------------------------------------------------------------------------|------------------|
| Pearson’s correlation co-efficient       | \( r = \frac{\sum (F-O)(O-O) \cdot \sum (F-O)}{\sqrt{\sum (F-O)^2} \cdot \sqrt{\sum (O-O)^2}} \) | Linear dependency between forecast and observation                       | 1/0              |
| Root mean square error                   | RMSE = \( \sqrt{\frac{1}{N} \sum (F-O)^2} \)                          | Closeness between forecast and observation                               | 0/1              |
| Relative root mean square error          | RRMS = \( \frac{RMSE}{O} \)                                           | To understand values of RMSE                                             | 0/1              |
| Probability of detection (hit rate)     | POD = \( A/(A+C) \)                                                    | What fraction of the observed “yes” events were correctly forecasted?    | 1/0              |
| False alarm ratio                        | FAR = \( B/(B+C) \)                                                    | What fraction of the predicted “yes” events actually did not occur       | 0/1              |
| Frequency bias                           | Bias score = \( (A+B)/(A+C) \)                                        | How did the forecast frequency of “yes” events compare with the observed frequency of “yes” events? | 1/0              |
| Equitable threat score                   | ETS = \( (A-A_{\text{random}})/(A+B+C-A_{\text{random}}) \) \( A_{\text{random}} = (A+C)(A+B)/N \) | How well did the forecasted “yes” events correspond to the observed “yes” events? | 0/1              |
| Brier score                              | BS = \( \frac{1}{N} \sum (P_F-P_O)^2 \)                              | Measure of the magnitude of the probability forecast errors             | 1/0              |
| Brier skill score                        | BSS = 1 – \( \frac{BS}{BS_{\text{ref}}} \)                           | Accuracy of the probabilistic quantitative precipitation forecast (PQPFs) compared with the climatology | 1/0              |

Note: \( F, O, P_F \) and \( P_O \) denote the forecast, corresponding observation, probability of precipitation and observed frequency, respectively; \( N \) is the amount of forecast and observation pairs; similarly, \( F \) and \( O \) denote the forecast average and observation average; \( A, B, C \) and \( D \) are obtained from the contingency table (Table 4); and \( BS_{\text{ref}} \) is the Brier score of the reference probability forecast, typically the probability of event occurrence from the climatology.

Sources: Tao et al. (2014); Aminyavari et al. (2018).
In high elevation basin 1 located in northwestern Iran, 2015 and 2016 were wet years, while in the other two more years drought dominated. Almost all NWP models overestimated precipitation in this basin. The CMA model forecasts in basin 1 were closer to observations, whereas the UKMO forecasts’ quality was next. Basin 2 in northern Iran bordering the southern Caspian Sea coastline enjoys a low altitude and moderate precipitation. According to the SPI time series, 2016 was wet, while other three study years were dry. In this basin, all models, the ECMWF in particular, overestimated the precipitation. Among the models, the JMA model was closer to the observations.

Basin 3 along the southwestern margin of the Caspian Sea is subject to relatively low altitude. Among all study basins, it receives highest precipitation after basin 9. The years 2015 and 2016 were wet and the other two years experienced drought in basin 3. The majority of models,

![Matrix plot of annual precipitation and bar plot of the standard precipitation index (SPI) in 10 basins](image-url)

**FIGURE 3** Matrix plot of annual precipitation and bar plot of the standard precipitation index (SPI) in 10 basins.
except the ECMWF, underestimated the precipitation. In this basin, the ECMWF model was relatively close to the observations.

In southwestern basin 4 with a relatively high altitude, the UKMO model was marginally better than the other models, though most of the models yielded good
FIGURE 6  Pearson correlation co-efficient within the basin lead-time matrix

FIGURE 7  Relative root mean square error (RRMSE) within the basin lead-time matrix
forecasts. In this basin, there was only one drought year in 2015, and the rest, especially 2016, were wet years.

In high altitude basin 5, average precipitation is moderate compared with those of other basins over the studied years. The ECMWF model in this basin performed better than the other models.

Basin 6, located in western Iran with a moderate elevation, was wet over the four years on the basis of the SPI, and the ECMWF model was closer to the observations. However, all models forecasted a reduction in precipitation in the basin in 2017.

In basin 7, whose average elevation is highest and its average precipitation lowest among the 10 basins, based on SPI time series, these four years, especially in 2017, were in drought. In this basin, the UKMO and ECMWF were better than the other models.

In high altitude basin 8 in central Iran whose average precipitation is similar to that of basin 7 over the study years, but the SPI index of basin 8, contrary to basin 7, shows that in some of the years studied were wet years. It has also been closer to observation among the models forecasted by the UKMO model. In the low altitude basin 9 with high precipitation, the highest precipitation occurred in this basin compared with other basins. Of course, based on the SPI, 2015 and 2016 were wet years and the other two years experienced drought in basin 9. The ECMWF model was forecasted better than the other models.

The southern basin 10 bordering the Persian Gulf has received little precipitation during the study period, while the ECMWF and NCEP models forecasted precipitation with better accuracy than the other models. Note that in all basins, 2017 was a dry year whose precipitation was forecasted poorly by all models other than the CMA.

Overall, based on the SPI index, 2014 and 2017 were drought years and 2016 wet years in most basins; and the numerical models, especially the JMA, in the wet years were closer to observation.

Also, to investigate the effect of hydrogeographical attributes such as basin elevation on the performance of the models, the basins were classified into four categories according to their elevation. Figure 4 shows the box plots of these four categories. Based on it, precipitation decreased with increasing basin elevation, and the NWP models overestimated precipitation in high-altitude basins. Among the models, the performance of the ECMWF model decreased with increasing basin elevation, and the UKMO model increased.

Figure 5 also shows the mean seasonal precipitation over the study period. The majority of precipitation occurred in winter compared with those of other seasons, although the highest precipitation in basins 3 and 9 was observed in autumn. In basins 4–6, precipitation is quite low in summer; nevertheless, this low precipitation was forecasted correctly by all models. The ECMWF model
overestimated the precipitation depth over all seasons. In general, the ECMWF model performed better in seasonal forecasts, while the UKMO model was next. Also, all NWP models have forecasted precipitation in autumn and winter better than spring and summer, especially for the JMA model.

Certainly, more analysis can be performed on the impact of climatology and hydrogeographical attributes of the study basins on the performance of the numerical models. However, owing to page limitations as well as the scope of the current research that focuses on sub-daily precipitation forecasts, further analysis is pending on future research.

3.2 | Evaluation result

The results of the deterministic evaluation by the Pearson correlation co-efficient are presented in Figure 6. It indicates that the correlation of the ECMWF and UKMO forecasts with the observations is stronger than that of other models and grand ensemble in most basins. According to this criterion, all models had the best performance in basins 5 and 6, and the poorest performance was detected in basins 1 and 10. The JMA model was poor for a 6 hr lead time and, in most basins, its correlation co-efficient was < 50%. Another point is that the forecast skill improves with an increase in lead time. Nevertheless, in some basins, the forecast skill increased up to a 12 hr lead time, while descending from a 18 to a 24 hr lead time mostly due to the absence of observed precipitation in 1800 and 2400 time slots in most stations. The reason for the weakness of numerical models in the 6 hr lead time, as compared with longer lead times, is the model’s weakness in identifying the precise time of precipitation in the diurnal cycles. Furthermore, since the forecasted precipitations are cumulative, this weakness is compensated at longer lead times and a higher correlation co-efficient is obtained. The correlation co-efficient of the grand ensemble with observed data, in most stations, was higher than those of the JMA, NCEP and CMA, and lower than those of the UKMO and ECMWF. Such an outcome was predictable based on previous research since the BMA does not greatly influence deterministic assessment; rather, it improves probability forecasts.

The error magnitude in the precipitation forecast was measured via the RRMSE (Figure 7). Owing to variation in the precipitation magnitude among basins, the RRMSE was used instead of the RMSE to simplify the...
comparison across different basins and lead times. According to Figure 7, the models had almost similar performance. The CMA model had the largest forecasting error, while the ECMWF model yielded the minimum error. Note that the CMA and UKMO models in basin 10, located in the south of Iran, did not perform well. The results also suggested a reduction in forecast error with an increase in lead time, in spite of increased precipitation cumulative value. Overall, according to deterministic assessment, the ECMWF and CMA had the best and worst performances, respectively, and the models’ skill improved by increasing the lead time. The models had better performance in high-precipitation basins compared with the other basins.
According to Figure 8, in a dichotomous assessment based on the bias score criterion, the NCEP model had better performance in most basins, apart from basins 3 and 8, and predicted the number of precipitation events correctly. Models with a 6 hr lead time underestimated the number of events, more so in the grand ensemble. The UKMO and ECMWF models overestimated the number of precipitation events when compared with observations in most basins. The models did not perform well in forecasting precipitation occurrence across eight basins.

The ETS criterion was used to assess the quality of dichotomous assessment. As shown in Figure 9, the

![Figure 11](image-url)  
**FIGURE 11** ROC diagrams for basins 6–10 associated with 6, 12, 18 and 24 hr lead times. ROC, relative operating characteristic.
NCEP model performed well in forecasting the precipitation occurrence or non-occurrence, though with average performance in basin 8. The ECMWF model acquired poor scores in forecasting the occurrence or non-occurrence of precipitation in basins 3 and 8. Although the JMA and CMA models did not have good performance in forecasting precipitation, they achieved acceptable scores in forecasting the precipitation occurrence. The grand ensemble also acquired good ETS scores in most basins and had an acceptable performance in forecasting the precipitation occurrence or non-occurrence.

Furthermore, within the dichotomous assessment framework, precipitation thresholds were used to study the model performance in heavy precipitation. With the increase in the threshold value, a bias score of < 1 was obtained for all models, which indicates that heavy precipitation events were underestimated by all models across the study basins.

The ROC curves were used to analyse the performance of models in the proper detection of the occurrence of precipitation or false warnings (Figures 10 and 11). By increasing the lead time, the diagrams increasingly shift towards the top left corner, representing more correct detection and fewer false warnings. The best model performance was observed in basin 4, where a larger area under the ROC curve was developed as compared with the other basins. On the other hand, the poorest model performance was obtained in low-precipitation basin 10. Since the results are similar in this section, identification of the best model is not straightforward. Nonetheless, the poorest model is the JMA with less correct detection (a lower POD) compared with other models in most basins. In high-precipitation basins such as in basins 2, 3 and 9, the performance of the NCEP model was poorer than that of the other basins, while the UKMO model had better results in these basins. Overall, the models in high-precipitation basins showed a variety of performance in detecting the occurrence or non-occurrence of precipitation.

The results of the probability assessment based on the BSS are shown in Figure 12, where the grand ensemble acquired the best skill scores compared with the individual models. This suggests that the BMA is extremely effective in improving probabilistic forecasts obtained by the ensemble forecasts. According to the results obtained in the probabilistic assessment, among all numerical models, the NCEP acquired the best skill score in forecasting the precipitation incidence probability. Again, the worst model performance was observed in basin 8, while the best performance was detected in basin 5. Overall, the models achieved good skill scores in most basins and predicted the precipitation probability of occurrence better than the reference forecasts (climatological forecasts).
4 | CONCLUSIONS

In this research, sub-daily ensemble precipitation forecasts of five numerical models within the THORPEX Interactive Grand Global Ensemble (TIGGE) database were evaluated across 10 major basins in Iran. Furthermore, using the Bayesian model averaging (BMA) approach, a combined forecast was constructed and compared with those of individual models. The following conclusions are drawn based on the results of the present study:

- All NWP models had better forecasts in wet years/seasons than in dry years/seasons.
- With an increase in basin elevation, precipitations generally decreased, while the performance of the European Centre for Medium-Range Weather Forecasts (ECMWF) and UK Met Office (UKMO) model(s) decreased and increased, respectively.
- The maximum correlation co-efficient was obtained by the ECMWF and UKMO models, while the Japan Meteorological Agency (JMA) model had average performance with a 6 hr lead time. The best model performance was observed in basins 5 and 6, while the worst performance occurred in basin 1, which was subject to low precipitation. Moreover, the error magnitude in forecasting precipitation was the greatest in basin 10.
- In predicting the occurrence or non-occurrence of precipitation, the National Centers for Environmental Prediction (NCEP) model was superior to the others. Nevertheless, most models underestimated the number of precipitation events with a 6 hr lead time, and somehow overestimated in other lead times. In terms of quality of forecasting the precipitation occurrence or non-occurrence, the grand ensemble performed well, while most models did not achieve good scores (in particular the ECMWF in basin 8).
- According to the ROC diagrams, the JMA model had the minimum area under the curve in most basins, representing more incorrect warnings and less correct detection in precipitation occurrence.
- In predicting the probability of precipitation occurrence, the grand ensemble yielded higher skill scores across all basins compared with the individual models. This indicates that the BMA has a great influence on improving the probability forecasts, though it was not very effective in deterministic assessments.

Overall, the present study revealed that the ECMWF, UKMO and NCEP models had good performance in forecasting sub-daily precipitation in most basins of northern and southwestern Iran, and may be further examined in the context of operational flood warning. Furthermore, to quantify uncertainty in ensemble numerical forecasting, the BMA technique is recommended in order to improve the probabilistic forecasts.

ORCID
Bahram Saghafian https://orcid.org/0000-0003-2846-2840
Saleh Aminyavari https://orcid.org/0000-0002-5319-2427

REFERENCES
Aminyavari, S., Saghafian, B. and Delavar, M. (2018) Evaluation of TIGGE ensemble forecasts of precipitation in distinct climate regions in Iran. Advances in Atmospheric Sciences, 35(4), 457–468. https://doi.org/10.1007/s00376-017-7082-6.
Beguería, S. and Vicente-Serrano, S.M., (2013) SPEI: calculation of the standardized precipitation-evapotranspiration index. R package version, 1, p.6.
Bhomia, S., Jaiswal, N. and Kishtawal, C.M. (2017) Accuracy assessment of rainfall prediction by global models during the landfall of tropical cyclones in the North Indian Ocean. Meteorological Applications, 24(3), 503–511. https://doi.org/10.1002/met.1649.
Fraley, C., Raftery, A.E., Gneiting, T. and Sloughter, J.M., 2007. EnsembleBMA: An R package for probabilistic forecasting using ensembles and Bayesian model averaging (No. TR-516). WASHING-TON UNIV SEATTLE DEPT OF STATISTICS.
Huang, L. and Luo, Y. (2017) Evaluation of quantitative precipitation forecasts by TIGGE ensembles for South China during the presummer rainy season. Journal of Geophysical Research: Atmospheres, 122(16), 8494–8516. https://doi.org/10.1002/2017JD026512.
Liu, J. and Xie, Z. (2014) BMA probabilistic quantitative precipitation forecasting over the Huaihe basin using TIGGE multimodel ensemble forecasts. Monthly Weather Review, 142(4), 1542–1555. https://doi.org/10.1175/MWR-D-13-00031.1.
Louvet, S., Sultan, B., Janicot, S., Kamsu-Tamo, P.H. and Ndiaye, O. (2016) Evaluation of TIGGE precipitation forecasts over West Africa at intraseasonal timescale. Climate Dynamics, 47(1–2), 31–47. https://doi.org/10.1007/s00382-015-2820-x.
Matsueda, M. and Nakazawa, T. (2015) Early warning products for severe weather events derived from operational medium-range ensemble forecasts. Meteorological Applications, 22(2), 213–222. https://doi.org/10.1002/met.1444.
Qu, B., Zhang, X., Pappenberger, F., Zhang, T. and Fang, Y. (2017) Multi-model grand ensemble hydrologic forecasting in the Fu river basin using Bayesian model averaging. Water, 9(2), 74. https://doi.org/10.3390/w9020074.
Raftery, A.E., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005) Using Bayesian model averaging to calibrate forecast ensembles. Monthly Weather Review, 133(5), 1155–1174. https://doi.org/10.1175/MWR2906.1.
Ran, Q., Fu, W., Liu, Y., Li, T., Shi, K. and Sivakumar, B. (2018) Evaluation of quantitative precipitation predictions by ECMWF, CMA, and UKMO for flood forecasting: application to two basins in
China. *Natural Hazards Review*, 19(2), 05018003. https://doi.org/10.1061/(ASCE)NH.1527-6996.0000282.

Sene, K. (2010) *Hydrometeorology: Forecasting and Applications*. New York, NY: Springer, p. 355. https://doi.org/10.1007/978-3-319-23546-2.

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(9), 3209–3220. https://doi.org/10.1175/MWR3441.1.

Su, X., Yuan, H., Zhu, Y., Luo, Y. and Wang, Y. (2014) Evaluation of TIGGE ensemble predictions of Northern Hemisphere summer precipitation during 2008–2012. *Journal of Geophysical Research: Atmospheres*, 119(12), 7292–7310. https://doi.org/10.1002/2014JD021733.

Swinbank, R., Kyouda, M., Buchanan, P., Froude, L., Hamill, T.M., Hewson, T.D., Keller, J.H., Matsueda, M., Methven, J., Pappenberger, F. and Scheuerer, M. (2016) The TIGGE project and its achievements. *Bulletin of the American Meteorological Society*, 97(1), 49–67. https://doi.org/10.1175/BAMS-D-13-00191.1.

Tao, Y., Duan, Q., Ye, A., Gong, W., Di, Z., Xiao, M. and Hsu, K. (2014) An evaluation of post-processed TIGGE multimodel ensemble precipitation forecast in the Huai river basin. *Journal of Hydrology*, 519, 2890–2905. https://doi.org/10.1016/j.jhydrol.2014.04.040.

Vogel, P., Knippertz, P., Fink, A.H., Schlueter, A. and Gneiting, T. (2018) Skill of global raw and postprocessed ensemble predictions of rainfall over northern tropical Africa. *Weather and Forecasting*, 33(2), 369–388. https://doi.org/10.1175/WAF-D-17-0127.1.

Wilks, D.S. (2011) Statistical Methods in the Atmospheric Sciences. 3rd ed. In: *International Geophysics*, vol. 100. San Diego, CA: Academic Press, pp.301–394. https://doi.org/10.1016/B978-0-12-385022-5.00008-7

Zhong, Y., Guo, S., Ba, H., Xiong, F., Chang, F-J. and Lin, K. (2018) Evaluation of the BMA probabilistic inflow forecasts using TIGGE numeric precipitation predictions based on artificial neural network. *Hydrology Research*, 49(5), 1417–1433. https://doi.org/10.2166/nh.2018.177.

---

**How to cite this article:** Saedi A, Saghaian B, Moazami S, Aminyavari S. Performance evaluation of sub-daily ensemble precipitation forecasts. *Meteorol Appl*. 2020;27:e1872. https://doi.org/10.1002/met.1872