ADVANCES IN ATMOSPHERIC SCIENCES, VOL. 35, APRIL 2018, 457–468

• **Original Paper** •

Evaluation of TIGGE Ensemble Forecasts of Precipitation in Distinct Climate Regions in Iran

Saleh AMINYAVARI[1], Bahram SAGHAFIAN*[1], and Majid DELAVAR[2]

[1]*Department of Technical and Engineering, Science and Research Branch, Islamic Azad University, Tehran* 1477893855*, Iran*
[2]*Department of Water Resources Engineering, Tarbiat Modares University, Tehran* 14115-336*, Iran*

(Received 14 April 2017; revised 13 July 2017; accepted 7 August 2017)

ABSTRACT

The application of numerical weather prediction (NWP) products is increasing dramatically. Existing reports indicate that ensemble predictions have better skill than deterministic forecasts. In this study, numerical ensemble precipitation forecasts in the TIGGE database were evaluated using deterministic, dichotomous (yes/no), and probabilistic techniques over Iran for the period 2008–16. Thirteen rain gauges spread over eight homogeneous precipitation regimes were selected for evaluation. The Inverse Distance Weighting and Kriging methods were adopted for interpolation of the prediction values, downscaled to the stations at lead times of one to three days. To enhance the forecast quality, NWP values were post-processed via Bayesian Model Averaging. The results showed that ECMWF had better scores than other products. However, products of all centers underestimated precipitation in high precipitation regions while overestimating precipitation in other regions. This points to a systematic bias in forecasts and demands application of bias correction techniques. Based on dichotomous evaluation, NCEP did better at most stations, although all centers overpredicted the number of precipitation events. Compared to those of ECMWF and NCEP, UKMO yielded higher scores in mountainous regions, but performed poorly at other selected stations. Furthermore, the evaluations showed that all centers had better skill in wet than in dry seasons. The quality of post-processed predictions was better than those of the raw predictions. In conclusion, the accuracy of the NWP predictions made by the selected centers could be classified as medium over Iran, while post-processing of predictions is recommended to improve the quality.

Key words: ensemble forecast, NWP, TIGGE, evaluation, post-processing

Citation: Aminyavari, S., B. Saghafian, and M. Delavar, 2018: Evaluation of TIGGE ensemble forecasts of precipitation in distinct climate regions in Iran. *Adv. Atmos. Sci.*, **35**(4), 457–468, https://doi.org/10.1007/s00376-017-7082-6.

1. Introduction

Nowadays, meteorological forecasts are produced using numerical models. Precipitation is one of the widely demanded meteorological factors. Improvement of Quantitative Precipitation Forecasts (QPFs) is the main objective of forecast centers and a major challenge for the meteorological research communities. Deterministic predictions have limitations in atmospheric conditions and change in initial conditions; thus, Ensemble Prediction Systems (EPSs) have been produced to enhance numerical and probabilistic prediction skill (Sene, 2010). EPSs involve different individual predictions produced by different physical parameterizations or different initial conditions. In the 1990s, EPSs were used practically in calculating the chaotic nature of climate processes, which could significantly reduce the uncertainties that had existed previously (Buizza et al., 2005). The first EPSs started in 1992 using data from ECMWF and NCEP (Zapata, 2010). The WMO organized THORPEX to further improve ensemble forecasts of severe meteorological events with one-day to two-week lead times (Shapiro and Thorpe, 2004). The THORPEX executive phase lasted from 2005 to 2014 (Swinbank et al., 2016) but later extended until 2019. TIGGE encompasses EPSs of 10 numerical weather prediction (NWP) centers whose data are made available by the China Meteorological Administration (CMA) and ECMWF centers. When various models that produce EPSs from different weather centers are aggregated, the probabilistic nature of the ensemble precipitation forecasts is better retained and accounted for (Bao et al., 2011).

A number of researchers have evaluated the TIGGE data in different regions. For instance, Zhao et al. (2011) showed that ECMWF was slightly better compared to NCEP and CMA in China region, whereas for lead times of over five days, none of the centers presented reliable predictions.

* Corresponding author: Bahram SAGHAFIAN
Email: b.saghafian@srbaiu.ac.ir

Based on ensemble forecasting data of the CMA, UKMO, ECMWF, NCEP and JMA in the TIGGE datasets in the Northern Hemisphere, Zhi et al. (2011) investigated the multi-model ensemble (MME) precipitation forecasting techniques and concluded that the bias-removed ensemble mean forecast was more skillful and more stable than each individual model. Liu and Fan (2014) post-processed TIGGE precipitation predictions using Bayesian Model Averaging (BMA) and showed that the post-processed prediction skill was better compared to that of the raw predictions. Moreover, UKMO and ECMWF yielded better predictions compared to those of NCEP and CMA. For the Northern Hemisphere, Su et al. (2014) showed that the ECMWF product was better compared to those of other centers, while in central parts of the Northern Hemisphere better prediction skill was achieved compared to those of the equatorial regions.

Louvet et al. (2016) reported that ECMWF and UKMO provided better results in West Africa compared to those of other centers. Luitel et al. (2016) evaluated the precipitation products, driven by North Atlantic tropical cyclone activities, of five prediction centers in the TIGGE database and concluded that the predictions were more suitable for lead times up to 48 h. In evaluating the prediction accuracies of TIGGE data over South Korea at six operational forecast centers, Lee et al. (2016) showed that ECMWF and KMA (Korea Meteorological Administration) performed well, while CMC and CMA did poorly, in forecasts.

Some researchers have used databases other than TIGGE in applications of MME forecasts. For instance, Fan et al. (2012) evaluated the prediction ability of the three DEMETER models (CNRM, UKMO and ECMWF) as well as the MME in seasonal predictions of the East Asian summer monsoon. The interannual increment prediction approach was applied to improve the prediction ability of the models and it was concluded that the direct outputs of the models were better able to predict than its original form. Liu and Fan (2014) applied two statistical downscaling schemes based on three different DEMETER GCMs to predict station rainfall. The downscaling model based on any single predictor demonstrated lower prediction skill than the multi-predictor downscaling models.

In the context of regional studies conducted in or around Iran, Sodoudi et al. (2010) showed that ECMWF, to some extent, could better predict the location of precipitation bands in mountainous and high-elevation regions compared to those in desert plain. In addition, ECMWF provided better results in the Zagros Mountains, to the west of Iran, compared to the Alborz Mountains to the north of Iran. Gevorgyan (2013) concluded that changes in precipitation amounts throughout Armenia were not modeled properly by ECMWF precipitation data. Mohammad and Suma (2016) evaluated the 3-h precipitation product of the ECMWF's ERA-Interim over Iran and concluded that this product had adequate performance in precipitation prediction in the Zagros Mountains, southern shores of the Caspian Sea, and Northeast Iran. Raziei and Sotoudeh (2017) evaluated ERA-Interim data over Iran and concluded that, at most stations, sufficient accuracy was achieved. However, ECMWF underpredicted the precipitation at Caspian littoral stations, due to the inability of ERA-Interim to accurately predict heavy precipitation in the region. Javanmard et al. (2016) evaluated TIGGE database predictions in the Karoon river basin, located in the southwest of Iran, over the period 2008–09. The results showed that ECMWF performed better compared to products of other centers. Moreover, after post-processing by the Bagging, Adaboost, and BMA methods, they concluded that post-processed predictions performed better compared to raw predictions.

The accuracy of numerical ensemble precipitation predictions within the TIGGE database has not been evaluated over the whole country so far. This study aims to assess the TIGGE ensemble predictions of three meteorological centers—ECMWF, NCEP and UKMO—over the period 2008–16, with lead times of one to three days, covering 13 rain gauges from eight homogenous precipitation regimes as classified by Modarres (2006). The reason for selecting these three particular products out of ten available centers within the TIGGE database was due to their better abilities reported in previous studies. Evaluations were performed using (i) deterministic, (ii) dichotomous (yes/no) and (iii) probabilistic techniques. Finally, to assess the possible improvement in predictions, the ensemble predictions were post-processed using the BMA method, which constituted a grand ensemble prediction.

2. Data and methods

2.1. *Data*

The 2008–16 50-km prediction products were extracted from the TIGGE database at the ECMWF with lead times of one, two and three days over Iran. Among the centers in the TIGGE database, three (ECMWF, NCEP and UKMO) were selected. The characteristics of the aforementioned centers are provided in Table 1. Observed data were extracted for 13 synoptic stations in Iran, spread over eight different regions as classified by Modarres (2006). Table 2 presents the characteristics of the stations. Modarres (2006) classified eight homogenous precipitation regimes over Iran, based on the application of Ward's technique to the annual and monthly precipitation of the selected rainfall stations. These eight regimes/clusters cover 90% of the precipitation variance within Iran. The first cluster (G1) is the largest and includes stations in arid and semi-arid regions in central Iran. The second cluster (G2) involves highland margins of G1, while G3 represents the northwestern cold region. The fourth cluster (G4) includes areas along the Persian Gulf coast south of Iran, while the sixth and the eighth clusters (G6, G8) involve areas located along the coast of the Caspian Sea. The major difference between the G6 and G8 regions in the north is the amount of precipitation decreasing from west to east. The fifth and seventh clusters (G5, G7) encompass regions in the Zagros Mountains, where precipitation in G5 is higher than

Table 1. Characteristics of selected prediction centers within the TIGGE database (Su et al., 2014).

Center	Base time (UTC)	No. of ensemble members	Horizontal resolution archived	Forecast length (days)	Initial perturbation method	Model uncertainty
ECMWF[a]	00/12	50+1	N320 (~ 0.28°) N160 (~ 0.56°)	0–10, 10–15	EDA-SVINI	SPPT+SPBS
NCEP[b]	00/06/12/18	20+1	1.0° × 1.0°	0–16	BV-ETR	SPPT
UKMO[c]	00/12	23+1 (11+1)	0.83° × 0.56°	0–15	ETKF	RP+SKEB

[a]The ECMWF EPS used a horizontal resolution of N200 (~ 0.45°) for 0–10 day forecasts and N128 (~ 0.7°) for 10–15 day forecasts before 26 January 2010. The ensemble of data assimilation and the initial-time singular vectors (EDA-SVINI) used as the initial perturbation method. The stochastic perturbation of physics tendency (SPPT) has been applied to account for model uncertainties. The spectral stochastic backscatter scheme (SPBS) was also introduced into the ECMWF EPS on 9 November 2010.
[b]The NCEP EPS uses the bred vector-ensemble transform with rescaling (BV-ETR) to generate initial perturbations.
[c]The UKMO EPS uses the ensemble transform Kalman filter (ETKF) as the initial perturbation strategy. Random parameters (RPs) and stochastic kinetic energy backscatter (SKEB) schemes are used to represent model uncertainties. Number of ensemble members of UKMO was 23+1 before 17 August 2014.

Table 2. Characteristics of the selected stations for evaluation.

Group	Stations	Elevation (m)	Longitude (N)	Latitude (E)
G1	Esfahan	1550.4	51°40′	32°37′
	Semnan	1127	53°25′	35°35′
	Zahedan	1370	60°53′	29°28′
G2	Mashhad	999.2	59°38′	36°16′
	Shahrekord	2048.2	50°51′	32°17′
	Tehran	1190.2	51°19′	35°41′
G3	Tabriz	1361	46°17′	38°05′
G4	Ahvaz	22.5	48°40′	31°20′
	BandarAbbas	9.8	56°22′	27°13′
G5	Sanandaj	1373	47°00′	35°20′
G6	Babolsar	−21	52°39′	36°43′
G7	Ilam	1337	46°26′	33°38′
G8	Rasht	−8.6	49°37′	37°19′

in G7. The geographic distribution of the cluster regions is shown in Fig. 1. To make a direct comparison with precipitation spatial variation, Fig. 2 displays the average annual precipitation from 1984 to 2014. Interpolation of NWP predicted values at stations was implemented using the Inverse Distance Weighting (IDW) and Kriging methods.

2.2. *Evaluation techniques*

The evaluations were performed using deterministic, dichotomous (yes/no), and probabilistic approaches. For deterministic evaluation, four common criteria were adopted, including the Pearson correlation coefficient (Pearson's r), root-mean-square error (RMSE), mean absolute error (MAE), and the relative root-mean-square error (RRMSE). Furthermore, yes/no binary assessment criteria, including the probability of

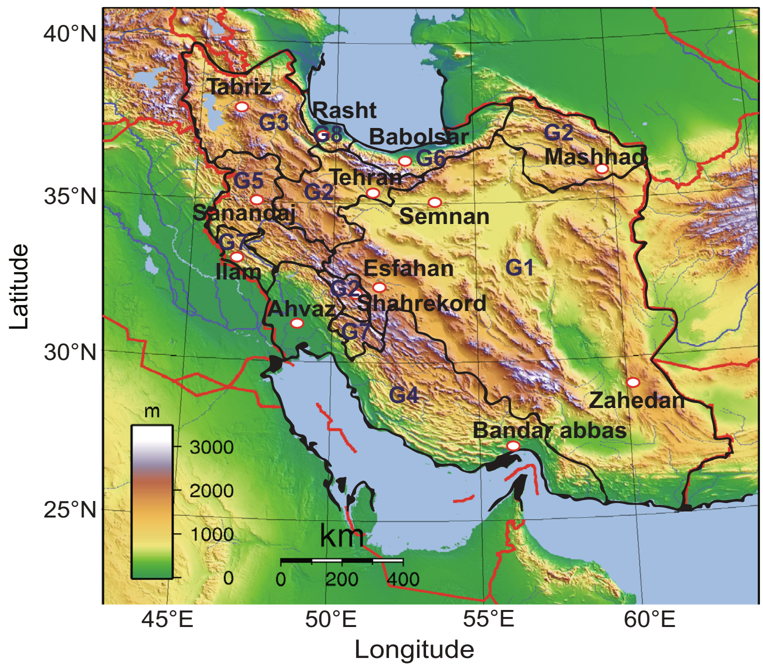

Fig. 1. Classification map of Iran's precipitation regimes according to Modarres (2006) overlaid on the topography (red circles are the selected stations in each region).

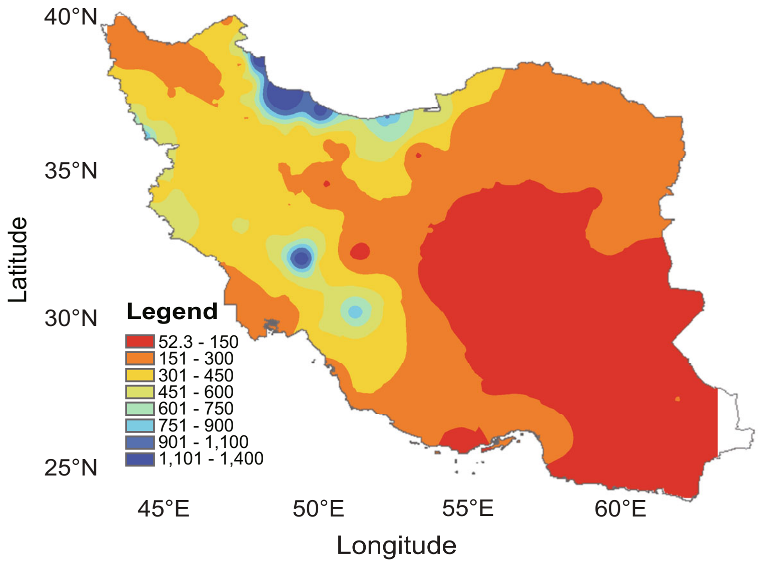

Fig. 2. Mean annual precipitation of Iran.

Table 3. Formulation of the evaluation criteria used in this study (Fan et al., 2008; Tao et al., 2014).

Verification measure	Formula	Description	Perfect/no skill
Pearson's correlation coefficient	$r=\frac{\sum(F-\overline{F})(O-\overline{O})}{\sqrt{\sum(F-\overline{F})^2}\sqrt{\sum(O-\overline{O})^2}}$	Linear dependency between forecast and observation	1/0
Mean absolute error	$\mathrm{MAE}=\frac{1}{N}\sum\lvert F-O\rvert$	Closeness between forecast and observation	0/
Root-mean-square Error	$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum(F-O)^2}$	Closeness between forecast and observation	0/
Relative root-mean-square error	$\mathrm{RRMSE}=\frac{\mathrm{RMSE}}{\overline{O}}$	To understand values of RMSE	0/
Probability of detection	$\mathrm{POD}=A/(A+C)$	What fraction of the observed "yes" events were correctly forecasted?	1/0
False alarm ratio	$\mathrm{FAR}=B/(B+C)$	What fraction of the predicted "yes" events actually did not occur	0/1
Frequency bias	$\mathrm{BIAS}=(A+B)/(A+C)$	How did the forecast frequency of "yes" events compare to the observed frequency of "yes" events?	1/
Equitable threat score	$\mathrm{ETS}=(A-A_{\mathrm{random}})/(A+B+C-A_{\mathrm{random}})$ $A_{\mathrm{random}}=(A+C)(A+B)/N$	How well did the forecasted "yes" events correspond to the observed "yes" events?	1/0
Brier score	$\mathrm{BS}=\frac{1}{N}\sum(P_F-P_O)^2$	Measure of the magnitude of the probability forecast errors	0/1
Brierskill score	$\mathrm{BSS}=1-\frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}}$	Accuracy of the PQPFs compared to the climatology	$1/\leqslant 0$
Continuous ranked probability score	$\mathrm{CRPS}=\int(P_F(x)-P_O(x))^2dx$	How well did the probability forecast predict the category that the observation fell into?	0/1

Notes: 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, $\overline{F}$ and $\overline{O}$ denote the forecast average and observation average. A, B, C and D are obtained from the contingency table, table 4. BS_{ref} is the Brier score of the reference probability forecast, typically the probability of event occurrence from the climatology.

detection (POD), false alarm rate (FAR), bias score (BIAS), and equitable threat score (ETS), were used for the dichotomous evaluations. Finally, the Brier score (BS), Brier skill score (BSS), continuous ranked probability score (CRPS), and the area under the relative operating characteristic (ROC) (ROC.Area) were adopted for the probabilistic evaluation. All criteria formulations are given in Table 3 and contingency table are shown in Table 4.

Table 4. 2×2 contingency table.

		Event observed	
		Yes	No
Event forecast	Yes	A	B
	No	C	D

2.3. *BMA*

BMA combines predictions from several statistical models with variable weighting coefficients. This method was used for ensemble predictions by Raftery et al. (2005) to predict air temperature, surface and sea level pressure (Liu and Fan, 2014). The probability distribution function (PDF) of BMA is as follows (Raftery et al., 2005):

$$P(y|f_1,\cdots,f_K)=\sum_{k=1}^{K}w_k g_k(y|f_k)\ ,$$

where y is the prediction coefficient; $g_k(y|f_k)$ is the conditional PDF of y based on f_k, which is the best member of the ensemble prediction; w_k is the posterior probability of forecast k which is non-negative with a summation equal to one and K is the number of models being combined. Since there were large numbers of zero precipitation events, the computational PDF in this paper was set to a Gamma distribution function, which was selected due to its high skewness. Detailed information regarding the calculation of $g_k(y|f_k)$ and w_k may be found in the literature (e.g. Raftery et al., 2005; Liu and Xie, 2014). This study took advantage of the ensemble BMA package in the R software.

3. Results and discussion

In what follows, the results of all evaluations associated with each of the eight regions are described. Due to a large number of results (prediction evaluation at 13 stations from 2008 to 2016), only the evaluation of 24-h precipitation at all stations is provided, and then the forecasts are evaluated for different lead times at the end of the section. Since in most parts of Iran precipitation is low in the dry seasons, the evaluations were carried out and reported for the wet seasons only. The wet seasons in Iran generally take place from November to April.

As previously noted, the IDW and Kriging methods were used for spatial interpolation of precipitation forecasts. Nevertheless, the results of these two methods showed no significant difference. Hence, in what follows, only the IDW results are presented.

3.1. *Total annual QPF evaluation*

According to Modarres (2006), the G1 region is the dominant precipitation regime in Iran and has a high coefficient of variation with low precipitation in a predominantly arid and semi-arid climate condition. Due to the extent of this region, three stations (Esfahan, Semnan, and Zahedan) were selected. Figure 3 presents the total annual precipitation associated with this region. In most years, all centers overestimated the annual precipitation, while ECMWF offered better precipitation predictions at Semnan and Esfahan compared to that at Zahedan. On the contrary, NCEP performed better in predicting the annual precipitation at Zahedan but comparatively poorly at Esfahan and Semnan.

In the G2 region, which essentially constitutes mountainous areas upstream of the G1 region, three stations were selected: Mashhad, Shahrekord, and Tehran. Similar to G1, all centers overestimated the annual precipitation for most years at Mashhad and Tehran. At Shahrekord, which receives higher precipitation than the other two stations, UKMO underestimated, whereas the other two generally overestimated, the precipitation.

Some centers showed different performance in predicting precipitation in the wet seasons compared with those of the whole year. For example, UKMO, which performed better than the other two models at Shahrekord, was the weakest for the wet season. The total NCEP predicted precipitation over the study period was significantly different from the total observed precipitation at Tehran.

In the G3 region, which encompasses cold regions in northwestern Iran, the station at Tabriz was studied. According to Fig. 3, NCEP predictions were the poorest in all years, except in 2010 and 2011, compared to those of the other centers, while better predictions were achieved by ECMWF compared to those of UKMO and NCEP.

In the G4 region, the stations at Ahvaz and Bandar Abbas were selected. Based on Fig. 3, although all three centers overestimated the annual precipitation, UKMO did quite poorly. For Sanandaj station in the G5 region, similar to other regions, all centers overestimated the annual precipitation. Predictions made by UKMO were better compared to those of ECMWF. Moreover, poorer predictions were made by ECMWF in 2008 and 2009.

In the rainy climate of the G6 region, the station at Babolsar was selected. Based on Fig. 4, the centers overestimated and underestimated precipitation in different years. At Ilam in G7, which generally receives more precipitation than G5, NCEP was the poorest of all the centers, whereas ECMWF's predictions were better than those of UKMO in most years. As shown in Fig. 4, in the G8 region, receiving higher precipitation than the G6 region, ECMWF offered better predictions compared to those of the other centers, while NCEP's was the poorest, underestimating the precipitation in all years.

Overall, the products of all the centers underestimated the precipitation in the relatively wetter climate regions but overestimated the precipitation in dryer climate areas. This implies a systematic bias in forecasts and demands application of bias correction techniques, such as quantile mapping.

3.2. *QPF deterministic evaluation*

For the deterministic evaluation, this study adopted four criteria: the correlation coefficient (r), MAE, RMSE, and RRMSE, whose formulations are presented in Table 3. The results are shown in Fig. 5. Due to limitations in displaying

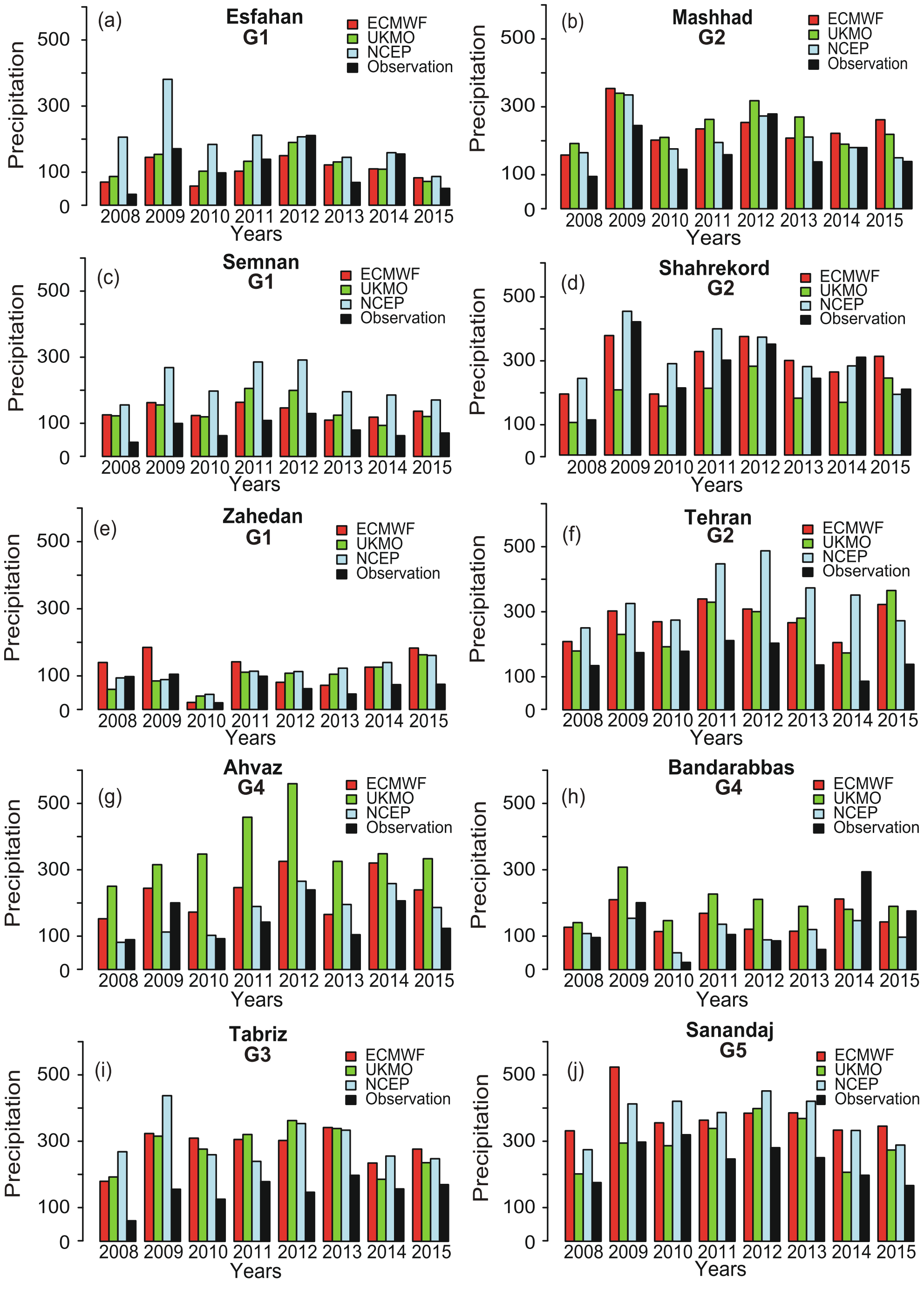

Fig. 3. Total observed and predicted annual precipitation (mm yr^{-1}) of three centers at rain gauge stations selected in precipitation regions: (a) Esfahan; (b) Mashhad; (c) Semnan; (d) Shahrekord; (e) Zahedan; (f) Tehran; (g) Ahvaz; (h) Bandarabbas; (i) Tabriz; (j) Sanandaj.

all examined cases, the average performance of the stations in each cluster is presented. Moreover, the results of each station are presented in Table 5.

At Esfahan and Semnan in the G1 region, ECMWF and NCEP yielded the best and poorest scores, respectively. In contrast, at Zahedan, ECMWF and NCEP were the poorest and the best predicting centers, respectively. All in all, in this region, ECMWF was the best and NCEP was the poorest.

In the G2 region, and based on the correlation coefficient, ECMWF at all three selected stations produced the best scores, while NCEP was the poorest. At Shahrekord, UKMO performed well, but was poorest at Mashhad.

In the cold climate of the G3 region, based on all three indicators, ECMWF was the best and NCEP was the poorest of all. In the hot and dry G4 region, NCEP yielded smaller prediction errors compared to those of the other centers, while UKMO performed comparatively poorly in terms of the deterministic evaluation scores.

Table 5. Summary of the evaluation results for stations at a lead time of one day. Bold numbers represent the best score among the three centers.

Station	Model	Correlation	MAE	RMSE	RRMSE	BIAS	ETS	FAR	BS	BSS	CRPS
ESFAHAN	ECMWF	0.55	**0.66**	**2.18**	**3.78**	**3.52**	**0.15**	**0.71**	**0.21**	**−0.69**	**0.54**
	NCEP	**0.62**	0.93	2.45	4.87	3.81	0.13	0.74	0.28	−1.3	0.76
	UKMO	0.54	0.71	2.23	4.13	4.09	0.11	0.75	0.26	−1.11	0.59
SEMNAN	ECMWF	**0.61**	**0.71**	**1.87**	**4.17**	5.52	0.06	0.81	0.4	−2.67	**0.57**
	NCEP	0.48	1.15	2.47	5.49	**5.15**	**0.07**	**0.8**	0.4	−2.62	0.95
	UKMO	0.55	0.77	1.93	4.32	5.83	0.05	0.82	**0.39**	**−2.55**	0.6
ZAHEDAN	ECMWF	**0.62**	**0.53**	**1.76**	**4.62**	5.72	0.12	0.8	0.22	−1.87	**0.39**
	NCEP	0.6	0.57	2.12	5.84	**4.57**	**0.18**	**0.74**	0.18	**−1.41**	0.42
	UKMO	0.6	0.54	1.88	4.84	6.03	0.1	0.81	0.23	−2	0.4
MASHHAD	ECMWF	**0.68**	**1.03**	**2.21**	**2.48**	3.04	0.11	0.67	0.34	−0.97	0.83
	NCEP	0.63	**1.03**	2.33	2.6	**2.67**	**0.17**	**0.62**	**0.29**	**−0.68**	**0.81**
	UKMO	0.62	1.13	2.5	2.86	3.28	0.08	0.69	0.36	−1.1	0.89
SHAHRE KORD	ECMWF	**0.74**	1.22	**3.18**	**2.25**	3.08	0.14	0.67	0.26	−0.67	1.01
	NCEP	0.7	1.4	3.64	2.56	**2.79**	**0.18**	**0.64**	0.27	−0.7	1.16
	UKMO	0.7	**1.18**	3.34	2.35	3.03	0.15	0.66	**0.25**	**−0.59**	**1**
TEHRAN	ECMWF	**0.65**	1.2	**2.4**	**2.77**	3.42	**0.1**	0.7	0.37	−1.15	0.95
	NCEP	0.55	1.62	3.14	3.7	**3.32**	**0.1**	**0.69**	0.4	−1.36	1.37
	UKMO	0.61	**1.15**	2.51	2.94	3.58	0.08	0.71	**0.36**	**−1.07**	**0.92**
TABRIZ	ECMWF	**0.63**	**1.22**	**2.46**	**3.17**	**2.82**	**0.12**	**0.63**	**0.34**	**−0.8**	**0.98**
	NCEP	0.55	1.41	2.83	3.67	2.85	0.11	**0.63**	0.37	−0.96	1.18
	UKMO	0.55	1.28	2.62	3.38	3.05	0.08	0.66	0.36	−0.87	1.04
AHVAZ	ECMWF	0.6	1.04	3.39	4.24	3.27	0.18	0.68	0.21	−0.67	0.81
	NCEP	**0.64**	**0.89**	**2.93**	**3.5**	**2.76**	**0.22**	**0.64**	**0.19**	**−0.45**	**0.73**
	UKMO	0.56	1.68	5	6.51	4.18	0.1	0.75	0.32	−1.53	1.33
BANDAR ABBAS	ECMWF	**0.58**	0.89	4.1	8.08	6.62	0.08	0.84	0.19	−1.61	0.75
	NCEP	0.56	**0.81**	**3.96**	**5.93**	**5.05**	**0.15**	**0.78**	**0.17**	**−1.3**	**0.67**
	UKMO	0.57	1.12	4.85	9.51	6.97	0.08	0.84	0.22	−2.05	0.92
SANANDAJ	ECMWF	**0.69**	1.54	3.25	2.42	2.86	0.12	0.65	0.32	−0.69	1.22
	NCEP	0.68	1.56	3.35	2.47	**2.64**	**0.16**	**0.62**	**0.28**	**−0.5**	1.27
	UKMO	0.65	**1.36**	**3.17**	**2.39**	2.73	0.14	0.63	0.29	−0.55	**1.1**
BABOLSAR	ECMWF	**0.74**	**2.2**	**5.76**	**2.06**	2.52	0.13	0.6	0.34	−0.65	**1.95**
	NCEP	0.61	2.44	6.55	2.41	**2.41**	**0.14**	**0.59**	**0.32**	**−0.55**	2.21
	UKMO	0.7	2.39	6.1	2.22	3.04	0.04	0.67	0.47	−1.31	2.11
ILAM	ECMWF	**0.72**	1.89	**4.46**	**2.17**	2.52	0.19	0.6	0.26	−0.36	1.5
	NCEP	0.68	**1.65**	4.66	2.24	**2.15**	**0.27**	**0.54**	**0.19**	**−0.02**	**1.43**
	UKMO	0.67	1.79	4.67	2.27	2.51	0.19	0.6	0.23	−0.2	1.46
RASHT	ECMWF	**0.78**	**2.88**	**6.79**	**1.72**	2.33	0.06	0.57	0.39	−0.69	**2.52**
	NCEP	0.72	2.92	7.81	1.96	**1.97**	**0.18**	**0.5**	**0.28**	**−0.2**	2.67
	UKMO	**0.78**	2.93	6.89	1.78	2.44	0.03	0.59	0.47	−1.01	2.57

In the G5 region, of all three centers, UKMO resulted in smaller prediction error, whereas NCEP performed the poorest. In the G6 rainy region, ECMWF and NCEP had the best and poorest scores, respectively. However, in this region, due to higher precipitation relative to other areas in Iran, large prediction errors were produced by all three centers.

At Ilam in the G7 region, ECMWF's predictions were slightly better than those of UKMO; NCEP was the poorest of all. In G8, based on the correlation coefficient and RMSE, ECMWF was the best and UKMO was the poorest.

In general, based on deterministic evaluation, ECMWF in most regions of Iran, UKMO in mountainous regions, and NCEP in southern Iran, provided better results compared to other centers. In addition, TIGGE numerical precipitation predictions at Ilam within the G7 region performed best among all examined stations in terms of annual precipitation.

3.3. *QPF dichotomous (yes/no) evaluation*

This study used four indicators (POD, FAR, ETS and BIAS) for dichotomous evaluation. The evaluation results are

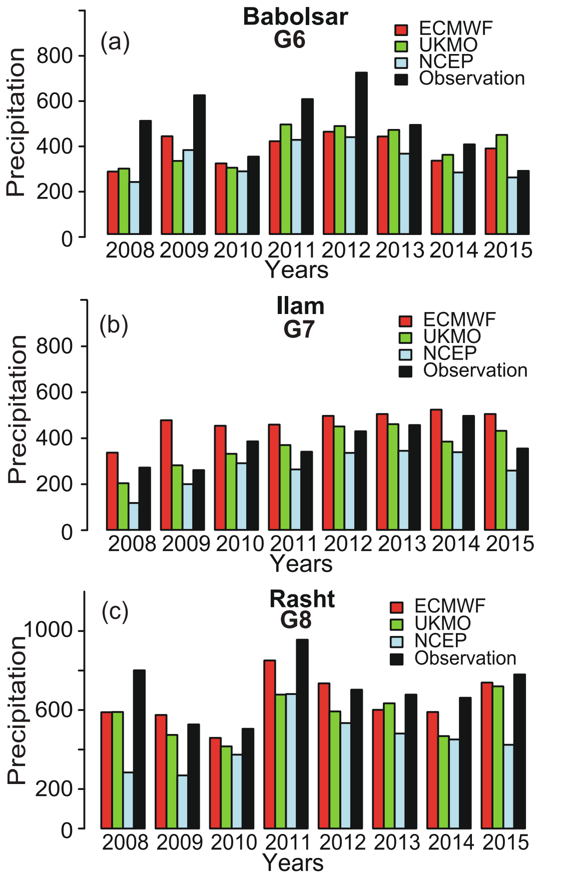

Fig. 4. Total observed and predicted annual precipitation (mm yr^{-1}) of three centers at three rain gauge stations selected in precipitation regions: (a) Babolsar in the G6 region; (b) Ilam in the G7 region; (c) Rasht in the G8 region.

shown in Fig. 5. According to the BIAS criteria, which is the ratio of the number of predicted precipitation events to observed precipitation events, NCEP and UKMO respectively offered the best and poorest predictions of the number of precipitation days. ECMWF showed smaller BIAS in the G3 region compared to that of NCEP. All centers overestimated the number of precipitation days.

Based on the ETS score, which measures the fraction of forecast events that were correctly predicted, NCEP achieved comparatively better scores at all stations, except in the G3 region. In addition, the prediction quality of UKMO was poor. However, the very low scores of ETS at most stations represents an inappropriate prediction accuracy of the number of precipitation events.

According to Fig. 5d, POD values are high, which is due to a high BIAS score at most stations. Of all centers, UKMO, due to the higher values of BIAS compared to those of other centers, yielded better POD, while NCEP had the lowest scores. Based on FAR, which represents the number of false alarms in precipitation events, UKMO was the poorest and NCEP, in most regions, was better than other centers. The number of false identifications was quite high in the G1 and G4 regions, most likely due to the rarity of precipitation events in these regions. In conclusion, the number of precipitation events predicted by all three centers was higher than observed, while NCEP had better scores in most regions.

3.4. *QPF probabilistic evaluation*

In this section, the gamma PDF was used to represent the QPF distribution. Four common methods (ROC.Area, CRPS, BS and BSS) were used for the probabilistic evaluation and the results are presented in Fig. 6. BS, which is a function

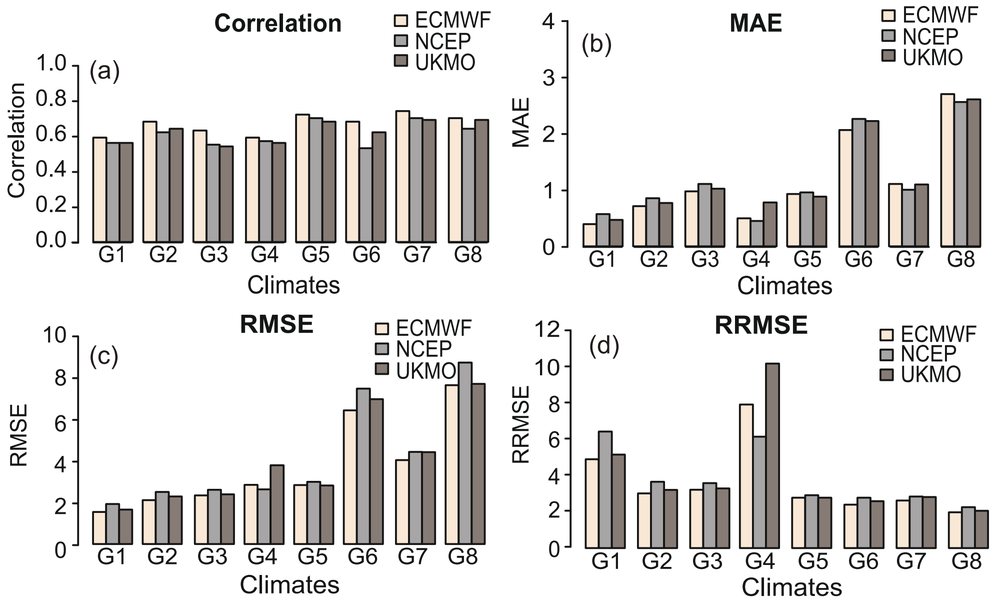

Fig. 5. Results of the deterministic evaluation of three centers for eight precipitation regions in Iran between observations and forecasts: (a) correlation coefficient; (b) mean absolute error (mm d^{-1}); (c) root-mean-square error (mm d^{-1}); (d) relative root-mean-square error.

of resolution, uncertainty and reliability, measures the mean squared probability error. BSS, which expresses the BS skill score relative to the reference BS, is usually determined by climatology predictions. CRPS evaluates the accuracy of the probabilistic forecast distribution. The ROC curve is a measure of the prediction's isolation skill in occurrence/non-occurrence of precipitation. The area under the curve is also an evaluation criterion. The values closer to 1.0 represent higher confidence in predictions.

Figure 7 shows the average probabilistic evaluations over the eight study years. Based on BS, precipitation at stations in the G4 region was better predicted than that at other selected stations. However, based on BSS, predictions were poor due to, as previously mentioned, the rarity of precipitation events. In all regions, based on BSS, NCEP showed better prediction capability compared to ECMWF, except in G1 and G3, whereas UKMO was the poorest based on both BS and BSS. Moreover, based on CRPS, UKMO and ECMWF had higher scores in some regions while NCEP did poorly compared to other models. Based on ROC.Area, ECMWF and NCEP yielded the highest and lowest scores, respectively.

As a whole, according to the probabilistic evaluations in

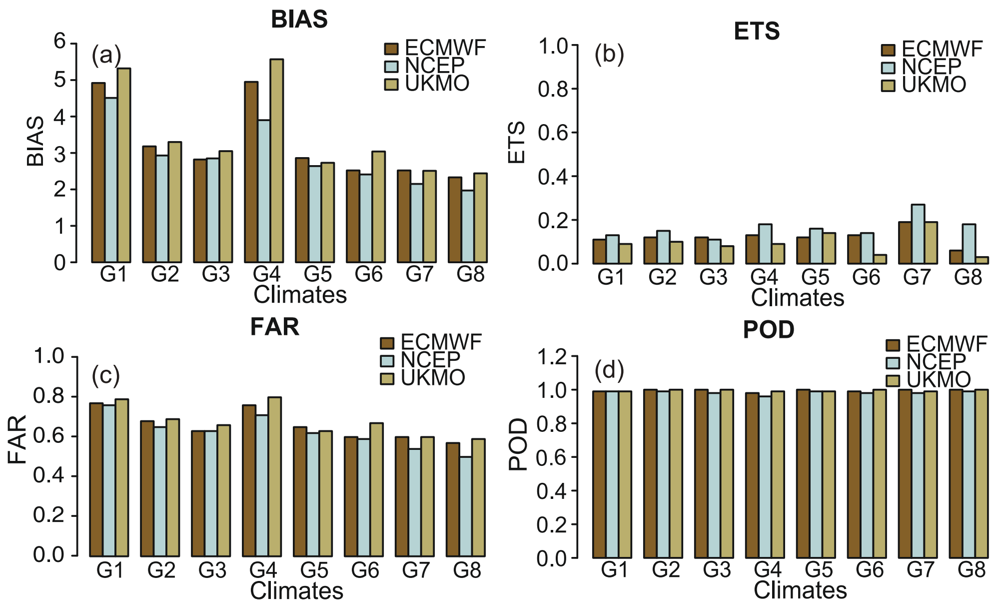

Fig. 6. Dichotomous (yes/no) evaluation of three centers for eight precipitation regions in Iran between observations and forecasts: (a) bias score (frequency bias); (b) equitable threat score (Gilbert skill score); (c) false alarm ratio; (d) probability of detection (hit rate).

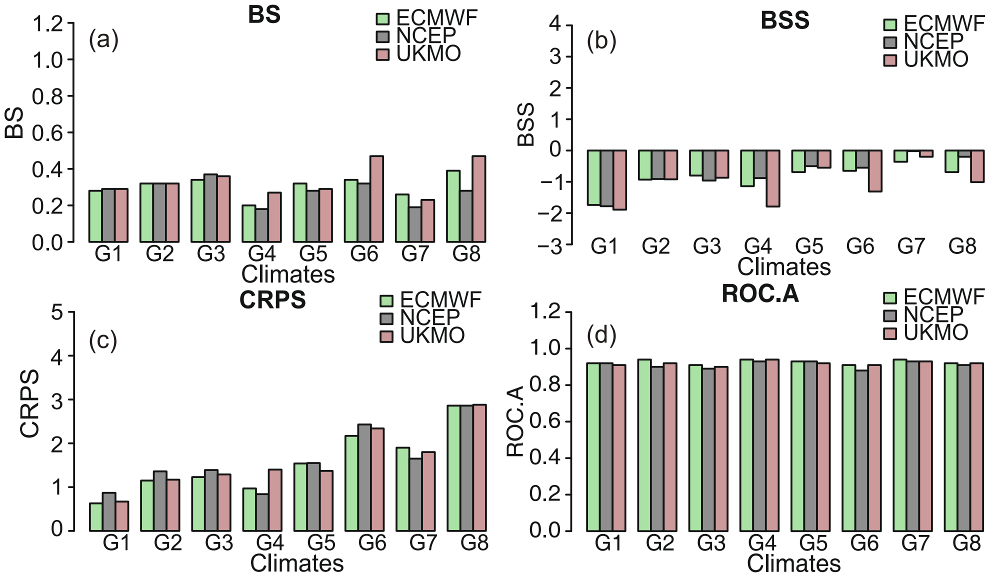

Fig. 7. Results of the probabilistic evaluation of three centers for eight precipitation regions in Iran between observation and forecasts: (a) Brier score; (b) Brier skill score; (c) continuous ranked probability score; (d) area under the relative operating characteristic (ROC) curve.

Table 5, precipitation at Semnan and Zahedan in the G1 region, as well as Bandar Abbas in G4, were poorly predicted. Mashhad, Zahedan, Ilam had better scores than those of other stations. ECMWF and NCEP performed almost the same, while UKMO performed poorer in the probability of precipitation occurrence/non-occurrence criteria.

Summary results are presented in Table 5, showing ECMWF performed better in all regions. UKMO had slightly better performance compared to NCEP in precipitation prediction. However, according to the dichotomous evaluation, NCEP performed better in almost all regions and could predict precipitation occurrence/non-occurrence better than other centers. Figure 8 presents the evaluation results for lead times of between one and three days. The results clearly illustrate that the precipitation prediction skill decreases with an increase in lead time. This reduction is quite obvious based on CRPS. According to Fig. 8, region G7 had the best scores, while the poorest performance in precipitation prediction was achieved in G1 and G4.

Also, Fig. 9 compares the performance of the models in the dry and wet seasons. Only the results of the rainy regions of G6 and G8 are presented because other regions receive very little precipitation in the dry season. Based on Fig. 9, all models performed better in the wet than in the dry season, whereas UKMO failed in the G8 region for the dry season.

Overall, the results indicate that better numerical prediction performance is expected in regions with high precipitation.

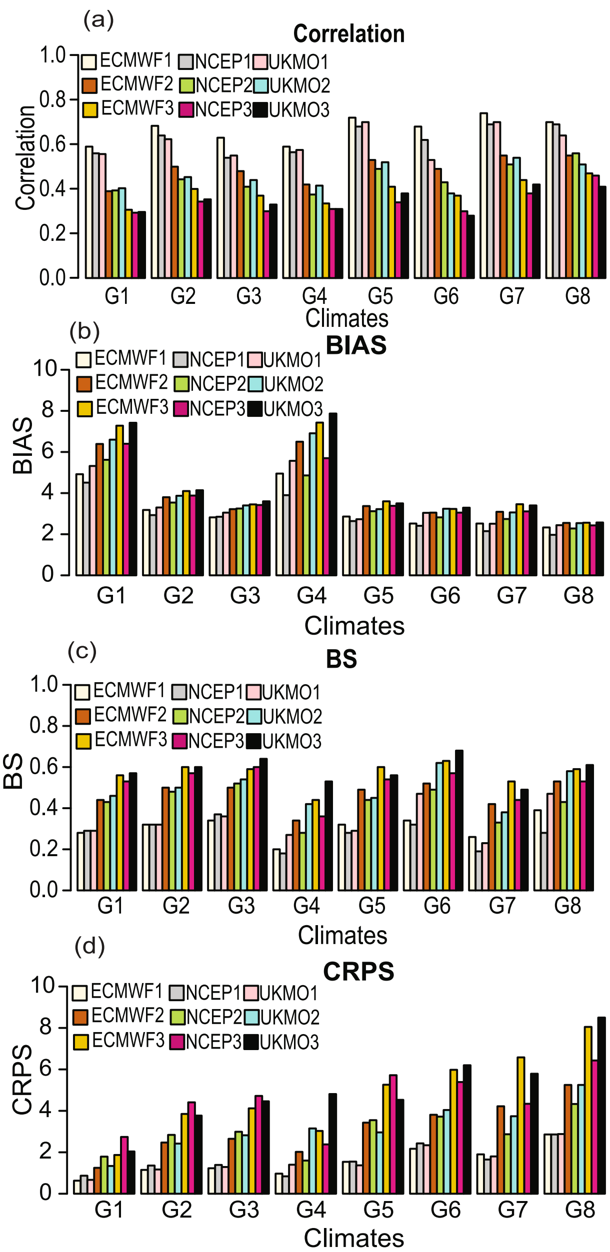

Fig. 8. Results of the three prediction centers' assessments for eight precipitation regions with different lead times between observation and forecasts: (a) correlation coefficient; (b) bias score; (c) Brier score; (d) continuous ranked probability score.

4. Grand ensemble prediction

A grand ensemble that includes EPS forecasts from several forecasting centers may improve the accuracy of numerical weather forecasts by taking uncertainties in the initial conditions, lateral boundary conditions, and model physics into account. The ensemble is potentially able to provide a better representation of the probable distribution of true predictions

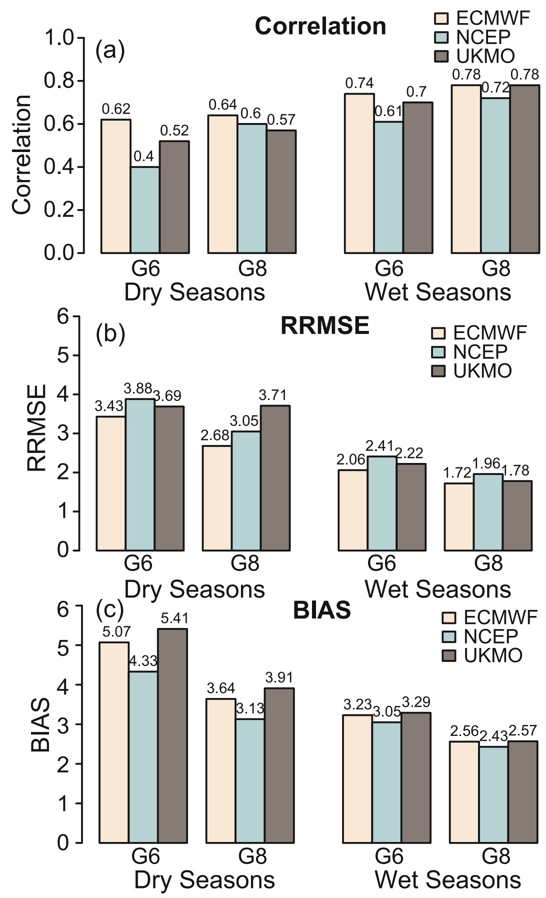

Fig. 9. Comparison of dry season and wet season ensemble predictions for rainy climates: (a) correlation coefficient; (b) relative root-mean-square error; (c) bias score.

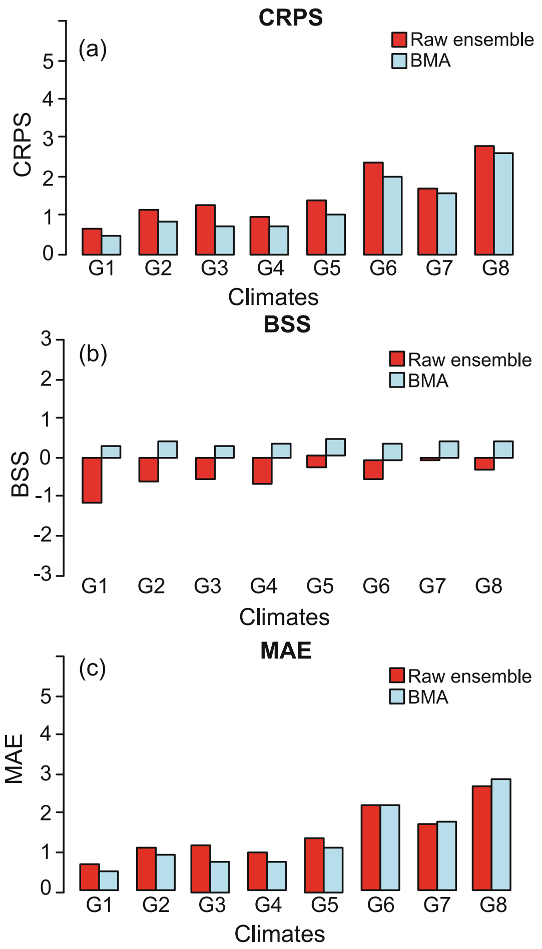

Fig. 10. Comparison of grand raw and post-processed ensemble predictions: (a) continuous ranked probability score; (b) Brier skill score; (c) mean absolute error (mm d^{-1}).

(Liu and Fan, 2014). For this purpose, ensemble predictions of the three centers were combined to constitute a grand ensemble prediction using two techniques. First, the products of the three centers were combined with the same weighting coefficients and without post-processing (raw) to constitute a grand ensemble prediction. In the second scenario, the predictions of each of the three centers were post-processed with variable weighting coefficients via BMA. Using the ensemble BMA package in the R software package, observed precipitation and ensemble predictions associated with the three centers from 2008 to 2016 were selected and, by a training period of 30 days based on the cube root of the ensemble mean, were post-processed.

To evaluate the performance of the BMA prediction model in both deterministic and probabilistic forecasts, MAE was used to measure the former skill, while CRPS and BSS were selected to measure the latter skill. Figure 10 presents the results, showing the prediction capability improved after post-processing. This implies that raw TIGGE ensemble predictions must be post-processed to be used in hydrological applications. Furthermore, the grand ensemble prediction showed better performance compared to individual model predictions.

5. Summary and conclusions

In this paper, TIGGE numerical ensemble precipitation predictions of the UKMO, NCEP and ECMWF centers for the Iran region were extracted with lead times of one, two and three days over the period 2008–16. To spatially break down the evaluation process, eight precipitation regions, as classified by Modarres (2006), were adopted. Deterministic, dichotomous (yes/no), and probabilistic evaluations were carried out for 13 selected stations in eight homogenous regions. The major findings were as follows:

(1) Comparison of the observed annual precipitation in the wet season at each station with the predicted values indicates that, in rainy regions, such as G6 and G8, the predicted precipitation overestimated the observed in most years. Conversely, annual precipitation was underestimated in other regions subject to a drier climate. ECMWF's predictions were closer to the observations in most regions, while UKMO predicted the annual precipitation quite well, mainly in mountainous regions, such as at the stations of Shahrekord and Sanandaj. Interestingly, UKMO underestimated precipitation in regions of high precipitation but overestimated observations in low-precipitation regions. NCEP predicted annual precipitation better than UKMO and ECMWF over the rim of the Persian Gulf in the G4 region.

(2) Based on deterministic evaluation, ECMWF performed best at most stations, while UKMO had better scores in mountainous regions, such as at Shahrekord and Sanandaj. Additionally, NCEP performed best in the G4 region.

(3) According to dichotomous (yes/no) evaluations and the BIAS indicator, all centers over-predicted the number of precipitation events, being much higher at some stations, such as in Bandar Abbas, with rare precipitation. In general, UKMO performed very poorly in precipitation occurrence/non-occurrence, compared to those of the other two centers. Moreover, NCEP performed better compared to UKMO and ECMWF.

(4) According to probabilistic evaluations, which represent the occurrence probability, reliability, and prediction quality, ECMWF had better scores, with NCEP coming close and UKMO rated last. Based on BSS, all centers were weaker than their climatology.

(5) Two- and three-day lead time predictions were also evaluated and, as expected, these predictions showed poorer skill compared to those of the one-day predictions.

(6) Comparing the wet and dry seasons in rainy regions, the evaluations showed that all three centers had better skill in the wet than in the dry season. This may indicate that it is easier to predict rainfall occurrence during the wet season.

(7) Ensemble predictions of the three centers were post-processed using BMA, constituting a grand ensemble prediction. The evaluation results showed that the quality of predictions improved considerably over the raw ensemble prediction.

All in all, it can be stated that, over Iran, ECMWF performs better compared to UKMO and NCEP. Overall, however, the evaluation scores returned a "medium" result, suggesting that precipitation predictions must be post-processed before application in operational forecasts.

REFERENCES

Bao, H.-J., L.-N. Zhao, Y. He, Z.-J. Li, F. Wetterhall, H. L. Cloke, F. Pappenberger, and D. Manful, 2011: Coupling ensemble weather predictions based on TIGGE database with Grid-Xinanjiang model for flood forecast. *Advances in Geosciences*, **29**, 61–67, https://doi.org/10.5194/adgeo-29-61-2011.

Buizza, R., P. L. Houtekamer, G. Pellerin, Z. Toth, Y. J. Zhu, and M. Z. Wei, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.*, **133**, 1076–1097, https://doi.org/10.1175/MWR2905.1.

Fan, K., H. J. Wang, and Y. J. Choi, 2008: A physically-based statistical forecast model for the middle-lower reaches of the Yangtze River Valley summer rainfall. *Chinese Science Bulletin*, **53**, 602–609, https://doi.org/10.1007/s11434-008-0083-1.

Fan, K., Y. Liu, and H. P. Chen, 2012: Improving the Prediction of the East Asian summer monsoon: New approaches. *Wea. Forecasting*, **27**, 1017–1030, https://doi.org/10.1175/WAF-D-11-00092.1.

Gevorgyan, A., 2013: Verification of daily precipitation amount forecasts in Armenia by ERA-Interim model. *International Journal of Climatology*, **33**, 2706–2712, https://doi.org/10.1002/joc.3621.

Javanmard, M., M. Delavar, and S. Morid, 2016: Evaluation and uncertainty analysis of the results of the global weather forecast models to apply in flood warning systems (case study: Karoon River basin, Iran). M.S. thesis, Tarbiat Modares University.

Lee, S.-M., J.-E. Nam, H.-W. Choi, J.-C. Ha, Y. H. Lee, Y.-H. Kim, H.-S. Kang, and C. Cho, 2016: A study on the predictability of the transition day from the dry to the rainy season over South Korea. *Theor. Appl. Climatol.*, **125**, 449–467, https://doi.org/10.1007/s00704-015-1504-0.

Liu, J. G., and Z. H. Xie, 2014: BMA probabilistic quantitative precipitation forecasting over the Huaihe Basin Using TIGGE multimodel ensemble forecasts. *Mon. Wea. Rev.*, **142**, 1542–1555, https://doi.org/10.1175/MWR-D-13-00031.1.

Liu, Y., and K. Fan, 2014: An application of hybrid downscaling model to forecast summer precipitation at stations in China. *Atmos. Res.*, **143**, 17–30, https://doi.org/10.1016/j.atmosres.2014.01.024.

Louvet, S., B. Sultan, S. Janicot, P. H. Kamsu-Tamo, and O. Ndiaye, 2016: Evaluation of TIGGE precipitation forecasts over West Africa at intraseasonal timescale. *Climate Dyn.*, **47**, 31–47, https://doi.org/10.1007/s00382-015-2820-x.

Luitel, B., G. Villarini, and G. A. Vecchi, 2016: Verification of the skill of numerical weather prediction models in forecasting rainfall from U.S. landfalling tropical cyclones. *J. Hydrol.*, https://doi.org/10.1016/j.jhydrol.2016.09.019.

Modarres, R., 2006: Regional precipitation climates of Iran. *Journal of Hydrology* (*New Zealand*), **45**, 13–27.

Mohammad, D., and Z. K. Suma, 2016: Evaluation of spatiotemporal accuracy of precipitation of European Center for medium-range weather forecasts (ECMWF) over Iran. *Physical Geography Research Quarterly*, **47**, 651–675, https://doi.org/10.22059/jphgr.2015.56054.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, https://doi.org/10.1175/MWR2906.1.

Raziei, T., and F. Sotoudeh, 2017: Investigation of the accuracy of the European Center for Medium Range Weather Forecast (ECMWF) in forecasting observed precipitation in different climates of Iran. *Journal of the Earth and Space Physics*., **43**, 133–147, https://doiorg/10.22059/jesphys.2017.57958.

Sene, K., 2010: *Hydrometeorology*. Springer, 345 pp.

Shapiro, M. A., and A. J. Thorpe, 2004: THORPEX international science plan. WMO/TD No. 1246, WMO.

Sodoudi, S., A. Noorian, M. Geb, and E. Reimer, 2010: Daily precipitation forecast of ECMWF verified over Iran. *Theor. Appl. Climatol.*, **99**, 39–51, https://doi.org/10.1007/s00704-009-0118-9.

Su, X., H. L. Yuan, Y. J. Zhu, Y. Luo, and Y. Wang, 2014: Evaluation of TIGGE ensemble predictions of Northern Hemisphere summer precipitation during 2008–2012. *J. Geophys. Res.*, **119**, 7292–7310, https://doi.org/10.1002/2014JD021733.

Swinbank, R., and Coauthors, 2016: The TIGGE project and its achievements. *Bull. Amer. Meteor. Soc.*, **97**, 49–67, https://doi.org/10.1175/BAMS-D-13-00191.1.

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

Zapata, J. A. V., 2010: Evaluation of hydrological ensemble prediction systems for operational forecasting. PhD dissertation, Université Laval.

Zhao, L.-N., F.-Y. Tian, H. Wu, D. Qi, J.-Y. Di, and Z. Wang, 2011: Verification and comparison of probabilistic precipitation forecasts using the TIGGE data in the upriver of Huaihe Basin. *Advances in Geosciences*, **29**, 95–102, https://doi.org/10.5194/adgeo-29-95-2011.

Zhi, X. F., L. Zhang, and Y. Q. Bai, 2011: Application of the multi-model ensemble forecast in the QPF. 2011 *International Conference on Information Science and Technology* (*ICIST*), Nanjing, IEEE, 657–660.