1 Comparison of probabilistic post-processing approaches for 2 improving NWP-based daily and weekly reference evapotranspiration 3 forecasts

4 Hanoi Medina, Di Tian

5 Department of Crop, Soil, and Environmental Sciences, Auburn University, Auburn, AL 36849

6 *Correspondence to*: Di Tian (tiandi@auburn.edu)

7 **Abstract:** Reference evapotranspiration (ETo) forecasts play an important role in agricultural, environmental, and water
8 management. This study evaluated probabilistic post-processing approaches, including the nonhomogeneous Gaussian
9 regression (NGR), affine kernel dressing (AKD), and Bayesian model averaging (BMA) techniques, for improving daily and
10 weekly ETo forecasting based on single or multiple numerical weather predictions (NWP) from The International Grand
11 Global Ensemble (TIGGE), including the European Centre for Medium-Range Weather Forecasts (ECMWF), the National
12 Centers for Environmental Prediction Global Forecast System (NCEP), and the United Kingdom Meteorological Office
13 forecasts (UKMO). We found that the NGR, the AKD and the BMA methods greatly improved the skill and reliability of the
14 ETo forecasts compared to a linear regression bias correction method, due to the considerable adjustments on the spread of
15 ensemble forecasts. The methods were especially effective when applied over the weekly NCEP forecasts, followed by UKMO
16 forecasts. The post-processed weekly forecasts had much lower rRMSE (between 8-11%) than the persistence-based weekly
17 forecasts (22%), and the post-processed daily forecasts (13-20%). Compared with the single model ETo forecasts based on
18 ECMWF, multi-model ensemble ETo forecasts showed higher skill at short lead times (1 or 2 days) and over the southern and
19 western regions of the United States. The improvement was higher at the daily timescale than at the weekly timescale. The
20 NGR and AKD methods performed the best, but the NGR method is more simple and computationally efficient than the other
21 methods. In summary, the study demonstrated that the three probabilistic approaches generally outperform conventional
22 procedures based on the simple bias correction of single model forecasts, with the NGR post-processing of the ECMWF and
23 ECMWF-UKMO forecasts providing the most efficient ETo forecasting.

24 1. Introduction

25 Reference crop evapotranspiration (ETo) represents the weather driven component of the water transfer from plants and soils
26 to the atmosphere. It plays a fundamental role in estimating mass and energy balance over land surface as well as in agronomic,
27 forestry, and water resources management. In particular, ETo forecasting is important for aiding water management decision
28 making (such as irrigation scheduling, reservoir operation, etc.) under uncertainty by identifying the range of future plausible
29 water stress and demand. However, ETo forecasting is highly uncertain due to the chaotic nature of weather systems. In

30   addition, ETo estimation requires full sets of meteorological data which is usually not easy to obtain. Due to the improvement

31   of numerical weather predictions (NWPs), studies have been recently emerged to forecast ETo using outputs of NWPs over

32   different regions of the world (Silva et al., 2010; Tian and Martinez, 2012 a, 2012b, and 2014; Perera et al., 2014; Pelosi et al.,

33   2016; Chirico et al., 2018; Medina et al., 2018). Operationally, experimental ETo forecast products are being developed, such

34   as Forecast Reference EvapoTranspiration (FRET) product (https://digital.weather.gov/), as part of the U.S. National Weather

35   Service (NWS) National Digital Forecast Database (NDFD) (Glahn and Ruth, 2003), and the Australian Bureau of

36   Meteorology's Water and Land website (http://www.bom.gov.au/watl), which provides current and forecasted ETo at the

37   continental scale.

38   The improved performance of NWPs during recent years is largely due to the improvement of physical, statistical

39   representations of the major processes in the models, and the use of ensemble forecasting (Hamill et al., 2013, Bauer et al.,

40   2015). Nevertheless, the NWP forecasts still commonly show systematic inconsistencies with measurements, which are often

41   caused by inherent errors of NWPs or local land-atmospheric variability which is not well resolved in the models. Post-

42   processing methods, defined as any form of adjustment to the model outputs in order to get better predictions (eg., Hagedorn

43   et al., 2012), is highly recommended to attenuate, or even eliminate, those inconsistencies (Gneitting et al., 2005; Raftery et

44   al., 2005). However, most post-processing procedures only considered single-model predictions (i.e., predictions generated by

45   a single NWP model), and addressed errors in the mean of the forecast distribution while ignored those in the forecast variance

46   (Gneiting, 2014). These procedures regularly adopted some form of model output statistics (MOS, Glahn and Lowry, 1972;

47   Klein and Glahn, 1974) methods, focusing on correcting current ensemble forecasts based on the bias in the historical forecasts.

48   As no forecast is complete without an accurate description of its uncertainty (National Research Council of the National

49   Academies 2006), the dispersion of the forecast ensemble often misrepresent the true density distribution of the forecast

50   uncertainty (Krzysztofowicz 2001; Smith 2001; Hansen 2002). The ensemble forecasts are, for example, commonly under-

51   dispersed (e.g. Buizza et al. 2005; Leutbecher and Palmer, 2008), which make the probabilistic predictions overconfident

52   (Wilks 2011). Therefore, a new generation of probabilistic techniques has been proposed to also address dispersion errors of

53   the ensembles (Hamill and Colucci 1997; Buizza et al. 2005, Pelosi et al., 2017), in some cases through the manipulation of

54   multi-model weather forecasts. The nonhomogeneous Gaussian regression (NGR, Gneiting et al., 2005), the Bayesian model

55   averaging, (BMA, Raftery et al., 2005; Fraley et al., 2010) and the family of kernel dressing (Roulston and Smith 2003; Wang

56   and Bishop 2005), such as the affine kernel dressing (AKD, Brocker and Smith 2008), are emerging probabilistic techniques

57   (Gneiting, 2014), with the NGR and the BMA methods being especially designed for multi-model post-processing.

58   Studies suggest that the post-processing of NWP-based ETo forecasts are crucial for informing decision making (e.g. Ishak et

59   al., 2010). Medina et al. (2018) compared single and multi-model NWP-based ETo forecasts and the results showed that the

60   performance of the multi-model ensemble ETo forecasts considerably improved through a simple bias-correction post-

61   processing, and that the bias-corrected multi-model forecasts were in general better than the single model forecasts. In reality,

62   while most applications for the ETo forecasting have involved some form of post-processing, these have been often limited to

63   simple MOS procedures of single-model ensembles (e.g., Silva et al., 2010; Perera et al., 2014). Poor treatments of uncertainty

64   and variability is considered as a main issue affecting users' perceptions and adoptions of weather forecasts (Mase and
65   Prokopy, 2014). The appropriate representation of the second and higher moments of the ETo forecast probability density is
66   especially important to predict extreme values. Therefore, the use of probabilistic post-processing techniques such as the NGR,
67   the AKD and BMA, may greatly enhance the overall performance of the ETo forecasts compared to the simple MOS
68   procedures.
69   Only a few studies have considered probabilistic methods for post-processing ETo forecasts. These include the works of Tian
70   and Martinez (2012a, 2012b, and 2014), and more recently Zhao et al (2019). The former authors showed the Analog Forecast
71   (AF) method to be useful for the post-processing ETo forecasts based on Global Forecast System (GFS, Hamill et al., 2006)
72   and Global Ensemble Forecast System (GEFS, Hamill et al., 2013) reforecasts. Tian and Martinez (2014) found that water
73   deficit forecasts produced with the post-processed ETo forecasts had higher accuracy than those produced with climatology.
74   On other hand, Zhao et al. (2019) improved the skill and the reliability of the Australian BoM model using a Bayesian joint
75   probability (BJP) post-processing approach, which is based on the parametric modelling of the joint probability distribution
76   between forecast ensemble means and observations. However, a main disadvantage of both the AF and the BJP methods
77   compared to the aforementioned emerging probabilistic approaches is that, while they transform the spread of the ensembles,
78   they rely on the mean of retrospective reforecasts, thus neglecting information about their dispersion. The AF approach also
79   requires long time series of retrospective forecasts, and may be unsuitable for extreme events forecasting (e.g. Medina et al.,
80   2019). The AKD, NGR and BMA methods produce continuous predictive density distributions, which may be useful for the
81   decision making (Gneiting, 2014), and perform commonly well with relatively short training datasets (Geiting et al., 2005;
82   Raftery et al., 2005; Wilks and Hamill, 2007). The use of novel forecasting strategies relying on the postprocessing of single
83   and multi-model forecasts with these emerging probabilistic techniques provide good opportunities for improving the ETo
84   predictions.
85   While ETo forecasts based on global medium range NWP have been mostly focused on the daily timescale (Perera et al., 2014;
86   Silva et al., 2010; Tian and Martinez, 2012a, b, 2014; Medina et al., 2018), weekly ETo forecasts are also important for users.
87   Studies show that both daily and weekly forecasts have increasing influence on the decision makers in agriculture (Prokopy et
88   al., 2013; Mase and Prokopy, 2014) and water resource management (Hobbins et al., 2017). For example, irrigation is
89   commonly scheduled considering both daily and weekly basis. The weekly evapotranspiration forecasts are useful for
90   planning water allocation from reservoirs, especially in cases of shortages. Weekly ETo anomalies can also be useful to provide
91   warnings of wild-fires (Castro et al., 2003) and evolving flash drought conditions (Hobbins et al., 2017). Therefore, accounting
92   for the post-processing of both daily and weekly ETo predictions provides a more comprehensive view of the capabilities of
93   these forecasting approaches than considering only daily predictions while better s the user's actual needs.
94   In this paper, we are addressing several scientific questions which have not been adequately studied in previous literature,
95   including, how effective are the new probabilistic post-processing methods compared with the traditional MOS bias correction
96   methods for post-processing ETo forecasts? Is it worth implementing the probabilistic post-processing for multi-model rather
97   than single-model ensemble forecasting? For the first time, this work aims to evaluate and compare multiple novel strategies

98 for post-processing both daily and weekly ETo forecasts using the emerging probabilistic approaches. The study represents a

99 major step forward with respect to Medina et al. (2018), which evaluated the performance of raw and linear regression bias

100 corrected daily ETo forecasts produced with single and multi-model forecasts. It provides a broad characterization of the

101 performance for different probabilistic post-processing strategies but also diagnoses the causes of high and low performance.

102 **2 Methods and Datasets**

103 **2.1 The probabilistic methods**

104 The NGR, AKD and BMA techniques follow a common strategy: they yield a predictive probability density function (PDF)

105 of the post-processed forecasts y given the raw forecasts x and some fitting parameters θ ($p(y|x,\theta)$). The parameters θ are

106 fitted using a training dataset of ensemble forecasts and observations, as in the MOS techniques. Below is a brief description

107 of each technique.

108 **2.1.1 Non-Homogeneous Gaussian Regression**

109 The NGR (Gneiting et al., 2005) produces a Gaussian predictive (PDF) based on the current ensemble (of typically multi-

110 model) forecasts. If x_{ij} denote the jth ($j = 1, ..., m_i$) ensemble forecast member of model i ($i = 1, ..., n$), then

111 $p(y|x,\theta) \sim \mathcal{N}(\mu, v)$, where the mean:

112 $\mu = a + \sum_{i=1}^{n} b_i \bar{x}_i$ (1)

113 is a linear combination of the mean ensemble forecasts $\bar{x}_i$ and the variance:

114 $v = c + dS^2$ (2)

115 is a linear function of the ensemble variance S^2. The fitting parameters a, b_i, c and d are determined by minimizing the

116 continuous rank probability score (CRPS) using the training set of forecasts and observations. Notice that parameters a, c and

117 d are indistinguishable among exchangeable members; therefore the b_i can be seen as a weighting parameters that reflect the

118 better or worse performance of one model compared to the others. The NGR technique is implemented in R (R Core Team)

119 using the packages ensembleMOS (Yuen et al., 2018),

120 **2.1.2. Affine Kernel Dressing**

121 The affine kernel dressing method (Bröcker and Smith, 2008) only considers single model ensemble forecasts. It

122 estimates $p(y|x,\theta)$ using a mixture of normally distributed variables:

123 $p(y|x,\theta) = \frac{1}{m\sigma} \sum_{j=1}^{m} K\left(\frac{y-z_j}{\sigma}\right)$ (3)

124 where K represents a standard normal density kernel ($K(\xi) = 1/\sqrt{2\pi} \exp(-1/2\xi^2)$), centered at z_j, such that:

125 $z_j = ax_j + r_1 + r_2\bar{x}$ (4)

126 and,

127     $\sigma^2 = h_s^2\big(s_1 + s_2 u(\mathbf{z})\big)$ (5)

128   where h_s is the Silversman's factor (Bröcker and Smith, 2008), $u(\mathbf{z})$ is the variance of $\mathbf{z}$ and a, r_1, r_2, s_1, s_2 are fitting

129   parameters obtained by minimizing the mean Ignorance score. For clarity we use the same nomenclature for the parameters as

130   in the original study. From Eqs. 4 and 5 we can obtain that the predictive variance v is a function of the ensemble variance S^2

131   (Brocker and Smith, 2008):

132     $v = h_s^2 s_1 + a^2(1 + h_s^2 s_2)S^2 = c^* + d^* S^2$ (6)

133   Here, S^2 represents the variance of the ensemble of exchangeable members.

134   The AKD technique is implemented through the SpecsVerification R package (Siegert, 2017).

2.1.3 Bayesian Model Averaging

136   The BMA method (Raftery et al. 2005, Fraley et al., 2010) also produces a mixture of normally distributed variables, as the

137   AKD method, but based on multi-model forecasts. In this case the predictive PDF is given by a weighted sum of component

138   PDFs, $g_i\big(y\,|\,x_{i,j}; \theta_i\big)$, one per each member:

139     $p(y|x,\theta) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} w_i g_i\big(y\,|\,x_{i,j}, \theta_i\big)$ (7)

140   such that weights and the parameters are invariable among members of the same model and

144     $\sum_{i=1}^{n} m_i w_i = 1$

141   In the study the component PDFs are assumed normal as for the affine kernel dressing method. Estimates of w_is and θ_is are

142   produced by maximizing the likelihood function using an Expectation Maximization algorithm (Casella and Berger, 2002).

143   The BMA technique is implemented through the ensembleBMA R package (Fraley et al., 2016).

2.2 Measurement and forecast datasets

146   ETo observations and forecasts were computed with the FAO-56 PM equation (Allen et al., 1998), from daily meteorological

147   data as inputs. They covered the same period, between May and August from 2014 to 2016. The observations used daily

148   measurements of minimum and maximum temperature, minimum and maximum relative humidity, wind speed, and surface

149   incoming solar radiation from 101 U.S. Climate Reference Network (USCRN) weather stations. The USCRN stations are

150   distributed over nine climatologically consistent regions in CONUS (Fig. 1). The ETo forecasts used daily maximum and

151   minimum temperature, solar radiation, wind speed, and dew point temperature reforecasts of European Centre for Medium-

152   Range Weather Forecasts model (ECMWF) outputs, United Kingdom Meteorological office model (UKMO) outputs, and

153   National Centers for Environmental Prediction model (NCEP) from The International Grand Global Ensemble (TIGGE;

154   Swinbank et al. 2016) database at each of these stations, considering a maximum lead time of 7 days. The weekly forecasts

155   accounted for the sum of the daily predictions generated a specific day of each week, and the weekly observations considered

156   the sum of the daily observations over the corresponding forecasting days, such that the weekly observations were independent

each other. In the study, we used the nearest neighbor approach to interpolate the forecasts to the USCRN stations, which does not account for the effects of elevation. While the use of interpolation techniques considering the effects of elevation (e.g. van Osnabrugge et al., 2019) may correct part of the forecasts errors before the post-processing, it could also affect the multivariate dependence of the weather variables. Hagedorn et al. (2012) showed that the post-processing can not only address the discrepancies related to the model's spatial resolution, but also serve as a means of downscaling the forecasts.

2.3 Post-processing schemes

2.3.1 Training and verification periods

The training data for the daily post-processing comprehended the pairs of daily forecasts and observations corresponding from 30 days prior to the forecast initial day, as in Medina et al. (2018). Instead, the training data for the weekly post-processing comprehended all the other pairs of weekly forecasts and observations available for the forecast location, similarly as in the case of a leave one out cross validation framework. In the study both the daily and weekly forecasts were verified for events over June-August, 2014-2016.

2.3.2 Baseline approaches

Linear regression bias correction (BC) of the ECMWF forecast was used as a baseline approach for measuring the effectiveness of the NGR, the AKD and the BMA methods considering both daily and weekly forecasts. Here, the current forecasts bias is estimated as a linear function of the forecasts mean, and the members of the ensemble are shifted accordingly. The function is calibrated using the forecasts mean and the actual biases from a retrospective set of forecast and observations. Persistence is also used as a baseline approach for weekly forecasts, considering its applicability in productive systems. In this case the ETo for a current week is estimated as the observed ETo during the previous week.

2.3.3 Forecasting Experiments

Table 1 summarizes the daily and weekly NWP-based ETo forecasting experiments based on different post-processing methods and model combinations. The analyses of the daily forecasts make more emphasis on the differences among post-processing methods. They include an examination of the effect of the duration of the training period on the forecasts assessments as well as the regression weights from the tested post-processing methods. Whereas, the weekly forecasts make more emphasis on the differences among the several single and multi-model ETo forecasts under baseline and probabilistic post-processing.

2.4 Forecast verification metrics

In this study we use several metrics to evaluate deterministic and probabilistic forecast performance of the post-processed ETo forecasts. For consistency purposes, the metrics of the tested methods were assessed using 50 random samples, i.e., same

186 number as members in the bias corrected ECMWF forecasts. Deterministic ETo forecast was produced by taking the average

187 of the ensemble members. The deterministic forecast performance was assessed using the bias or mean error, the root mean

188 square error (RMSE) and the correlation (ρ), which are common measures of agreement in many studies. Both the relative and

189 the absolute bias and RMSE are calculated and reported.

190 The probabilistic forecast performance was assessed using the spread-skill relationship (see Wilks, 2011) and the forecast

191 coverage as measurements of the forecast reliability, and the Brier Skill Score as a measurement of the skill. Reliability here

192 refers to the statistical consistency (as in Toth et al. 2003), which is met when the observations are statistically indistinguishable

193 from the forecast ensembles (Wilks, 2011). The spread-skill relations are represented as binned-type plots (e.g., Bosi et

194 al., 2017), accounting for the mean of the ensemble standard deviation deciles (as an indication of the ensemble spread) against

195 the mean RMSE of the forecasts in each decile over the verification period. The plots include the correlation between these

196 two quantities. Calibrated ensembles should show a 1:1 relationship between the standard deviations and the RMSE. If the

197 forecasts are unbiased and the spread is small compared to the RMSE, then the ensembles tend to be under-dispersive. The

198 inverse of the spread provides an indication of sharpness, which is the level of "compactness" of the ensemble (Wilks, 2011).

199 In addition to the spread skill relationship, we also report the ratio between the observed and nominal coverage (hereinafter

200 referred as coverage ratio). The coverage of a $(1 - \alpha)100\,\%$, $\alpha \in (0,1)$, central prediction interval is the fraction of

201 observations from the verification data set lying between $\alpha/2$ and $1 - \alpha/2$ quantiles of the predictive distribution. It is

202 empirically assessed by considering the observations lying between the extreme values of the ensembles. The nominal or

203 theoretical coverage of a calibrated predictive distribution is $(1 - \alpha)100\,\%$. A calibrated forecast of m ensemble members

204 provides a nominal coverage of about $(m - 1)/(m + 1)\,100\,\%$ central prediction interval (e.g., Beran and Hall, 1993). For

205 example, an ensemble of 50 members provides 96% central prediction interval. The ratio between the observed and nominal

206 coverages provides a quantitative indicator of the quality of the forecasts dispersion under unbiasedness: a ratio lower (larger)

207 than 1 suggest that the forecasts tend to be under (over) dispersive. Finally, the BSS represents a traditional skill-score

208 relationship that adopts the Brier score (Wilks, 2011), as the accuracy measure. This study we compute the BSS associated

209 to the tercile events of the ETo forecasts (upper or 1st, middle or 2nd, and lower or 3rd terciles), exactly as in Medina et al.

210 (2018).

3 Results

3.1 Comparing the NGR, AKD and BMA methods at daily scale

3.1.1 Deterministic forecast performance

214 Figure 2 shows the relative bias and RMSE as well as the correlation of the forecasts post-processed using different approaches

215 over the southeast (SE) and northwest (NW) region. In general, the probabilistic post-processing methods add no additional

216 skill to the deterministic forecast performance compared to the simple bias correction. While the RMSE are relatively high,

217   the bias is very low, which indicates that the errors are mostly random. The BMA and the simple linear regression methods

218   provided lower bias than the NGR and AKD methods. Instead, the BMA method provided higher RMSE and lower correlations

219   than the other three methods at long lead times.

3.1.2 Probabilistic forecast performance

221   The spread-skill relationship in Figure 3 shows that the probabilistic post-processing methods considerably improved the

222   reliability of the ETo forecasts compared with the linear regression bias correction (Figure 3). The former methods tend to

223   correct evident shortcomings of the ensemble raw forecasts which are unresolved by the simple post-processing, i.e., the

224   considerable under-dispersion at short lead times, and the poor consistency between the ensemble spread and the RMSE at

225   longer lead times. The adjustments had a low cost in terms of sharpness, judging by the range of ensemble spreads for the

226   different line plots, but seemed slightly insufficient. The correlations between the ensemble standard deviation and the RMSE

227   are fairly low, suggesting a limited predictive ability of the spread (Wilks, 2011). Nonetheless, they were consistently higher

228   for probabilistic post-processing methods, compared to the linear regression method, and at short lead times, compared to the

229   long lead times. The performance was low sensitive to the type of probabilistic post-processing, independent of the single or

230   multi-model forecasts strategy, although the BMA post-processing provided slightly lower correlations, especially for longer

231   lead times. The coverage ratios in table 2 provide quantitative insights about the forecasts under-dispersion for the different

232   strategies. The simple bias corrected ECMWF forecasts provided a mean coverage ratio of 77%, but it can be as low as the

233   50%. The other forecasts provided coverage ratios of over 91%. The ratios were slightly better (i.e., closer to one) using the

234   BMA method than with the NGR and the AKD methods, and using single ECMWF forecasts than with the ECMWF- UKMO

235   and the ECMWF-NCEP- UKMO forecasts.

236   The NGR and AFK methods provided better Brier skill score (BSS) than the BC method for the three categories of ETo values,

237   with improvements being higher for the middle tercile, than for the lower and upper terciles (Figure 4). The BMA based skill

238   scores tended to decrease with lead time. On west regions (SW, W and NW) and at short lead days the multi-model forecasts

239   post-processed with the NGR were the most skillful; in the other cases the ECMWF forecasts post-processed with the NGR

240   and the AKD methods tended to be best.

3.1.3 Summary of average performance for daily forecast

242   Table 3 shows the average performance for the lead days 1 and 7, by weighting the values of each metric according to the

243   number of stations in each region. The ECMWF- UKMO forecasts post-processed with the NGR method were best at short

244   lead times (1-2 days), while the ECMWF forecasts post-processed with the AKD and the NGR methods were the first and

245   second best at the longer lead times. The BMA method performed well at short lead times but poorly at long times, while the

246   simple bias correction method performed well for deterministic forecasts, but poorly for the probabilistic forecasts. The

247   forecast performance across climate regions is also associated with the choice of the ECMWF forecasts or the multi-model

248   forecasts (Table 4). The single model ECMWF forecasts performed better over northern climate regions than the multi-model

249   ensemble forecasts, while the multi-model did better than any single model forecast over the western regions. The performance

250   over the other regions was more variable among strategies. The performance of the ECMWF- UKMO forecasts was generally

251   better than that of the ECMWF-NCEP- UKMO forecasts (see Table 4, and Figs. 2 and 4). Unlike other performance metrics,

252   the coverage was mostly better for the ECMWF forecasts than for the multi-model forecasts.

3.1.3 Effect of the length of training period

254   The choice of an "optimum" training period is an important issue related to the operational use of post-processing techniques

255   for ETo forecasts. Here we compared the performance of different forecasts post-processed with NGR and AKD techniques

256   using 45 and 30 training days. The results suggest that the payoff from using 45 days is practically minimal. Table 5 shows

257   the percentage differences the forecasting performance of using 45 and 30 training days for post-processing. While there are

258   generally some minor improvements for using 45 days than 30 days, which tend to be higher at longer lead times than shorter

259   times, these improvements usually represent less than 3 percent of original statistics. The largest percentage difference,

260   accounting for the BSS at the middle tercile, actually represented a negligible gain in absolute terms since they were affected

261   by the close-to-zero range of the variable. The improvements were a bit higher for multi-model forecasts than for single model

262   forecasts.

3.1.4 Weighting coefficients

264   The weighting coefficients reflect both the performance of the ensemble models and the performance the post-processing

265   techniques relative to their counterparts. Figure 5 shows the mean b_i (Eq. 1) weighting coefficients of the NGR technique and

266   w_i (Eq. 7) weighting coefficient of the BMA techniques for each region and lead time for the post-processed ECMWF-NCEP-

267   UKMO, respectively. The coefficients for the NGR and BMA techniques exhibited some common patterns of variability across

268   regions and lead times. Both methods show that the weights of the ECMWF forecasts are at overall the highest, with a clear

269   maximum at medium lead times; the weights of the UKMO model are the highest at 1 and 2 days, but sharply decreases with

270   the lead time, while the weights of the the NCEP model are in general the lowest, although they consistently crease with

271   lead time. It explains well the most outstanding features of the performance assessments, in relation to the role of each model,

272   and the dependence on regions and lead times. Compared to the NGR method, the BMA method gives the UKMO forecasts

273   a higher relative weight, at the expense of the ECMWF forecast weights. For example, the weighting coefficients of the BMA

274   method over the west regions are consistently higher for the UKMO forecasts than for the ECMWF forecasts. It suggests that

275   the lower performance of the BMA post-processing relative to the NGR and the AKD methods may be related to a

276   misrepresentation of the model weights by the performance. This in turn may be caused by convergence problems during the

277   parameter optimization with the expectation-maximization algorithm (Vrugt et al., 2008).

278   We observed considerable similarities on the distribution of variance coefficients for the NGR method (Eq. 2) and the AKD

279   (Eq. 6) method after post-processing the ECMWF forecasts. The two methods also provide very similar adjustments on the

280   mean forecast because, unlike the BMA method, they independently bias correct the mean and optimize the spread-skill

281   relationship, (Bröcker and Smith, 2008). However, the computing speed using the NGR method is about 60 times faster than

282   using the AKD, which was perceived as the main drawback of the ▭ method. The BMA method is also more

283   computationally demanding than the NGR method but less than the AKD method. Considering the effectiveness,

284   computational efficiency and versatility of the NGR method, we applied this probabilistic technique to weekly ETo forecasts

285   based on single model and multi-model ensembles.

3.2 Assessing NGR me▭s for post-processing weekly ETo forecasts

286

3.2.1 Deterministic forecast assessments

287

288   As for the daily predictions, the bias, the RMSE and the correlation of the weekly forecasts post-processed with the NGR

289   method and the linear regression methods were similar (Fig. 6). However, while the RMSE of daily forecasts based on ECMWF

290   model varies between 12 and 20 % of the total ETo (Fig. 2), the RMSE for any of weekly forecasting strategies commonly

291   varies between 8 and 11%, which is lower than for daily forecasts, making it more useful for operational purpose. The post-

292   processed forecasts showed much lower RMSE and twice higher correlation than the predictions based on persistence, with

293   the weekly predictions based on ECMWF forecasts being generally better, followed by the predictions based on the UKMO

294   forecasts.

3.2.2 Probabilistic forecast assessments

295

296   Both the skill and the reliability of the weekly forecasts considerably improved through the NGR post-processing compared

297   with the bias correction post-processing (Table 6). The improvements were different among ETo forecast models. In most

298   cases, the better the forecasts performance, the lower the improvements are. The adjustments in the coverage ratio and the

299   Brier skill score were about 2.5 and 5 times larger for the UKMO and the NCEP forecasts, respectively, than for the ECMWF

300   forecasts. The bias corrected ECMWF forecasts are generally better than both the UKMO and NCEP forecasts post-processed

301   with the NGR method. We found that the post-processing of the NCEP forecasts with methods like the NGR is almost

302   mandatory to get reasonable probabilistic weekly forecasts ▭ example, the coverage ratio of the bias corrected forecasts on

303   the West region was only 29%, because of the considerable under-dispersion. However, it is notable that, once they were post-

304   processed with the NGR technique, they performed almost comparably to the UKMO forecasts post-processed with the same

305   metho ▭ ble 6 also shows that the multi-model ECMWF- UKMO weekly forecasts are commonly the best among all of

306   those post-processed using the NGR method, followed by the ECMWF and the ECMWF-NCEP- UKMO forecasts.

307   The improvements in the reliability came through substantial adjustments both in the ensemble spread and spread-skill

308   relationship of the raw forecasts (Fig. 7). The correlations between the standard deviation of the ensembles and the RMSE

309   were more than twice larger through the NGR post-processing than through the linear regression bias correction. The

310   adjustments seemed even slightly more effective than those resulting from the probabilistic post-processing of the daily

311   forecasts (Fig. 3), although at the expense of a greater loss of sharpness. The contrasts in the post-processing effectiveness are

312   probably associated with the differences in the training strategies.

313   In the case of the probabilistic forecast skill (Fig. 8), the improvements were larger for the middle tercile than for the other two

314   terciles, similarly as with daily forecasts. Unlike the bias corrected forecasts, any of the probabilistically post-processed

315   forecasts outperform climatology for practically any event and at any region. Maybe more importantly, the skill for the lower

316   and upper tercile events of the forecasts that have been post-processed with the NGR method is in most cases over 30% better

317   than the skill of climatology. In the coast regions, from the South to the Northwest the skill is commonly over 50% better,

318   similarly as for the daily forecasts. Finally, the improvements resulting from the use of multi-model forecasts compared to the

319   single model forecasts were generally small, except for the Southwest region.

320   **4. Discussion**

321   **4.1 Effects of probabilistic post-processing on ETo forecasting performance**

322   This study showed that NGR, AKD and BMA post-processing schemes considerably improved the probabilistic forecast

323   performance of the daily and weekly ETo forecasts compared with the simple bias correction method. While sharpness is a

324   wished quality of any forecast, the daily and weekly bias corrected ETo forecasts from NWP are spuriously sharp, which leads

325   to a poor consistency between the range of the ETo forecasts and the true values, and ultimately undermine the confidence on

326   those forecasts. They also experiment a poor consistency in that the variance of the ensembles are commonly insensitive to the

327   size of the forecast error. The probabilistic post-processed methods provided a much better reliability, with a coverage which

328   is close to the nominal value, and at a low cost on sharpness. Therefore, they lead to a much better agreement between the

329   forecasted probability of having an ETo event between certain thresholds and the proportions of times that the event occur (see

330   Gneitting et al., 2005). In the case of the weekly ETo forecasts, the rate of the improvements are considerably smaller for the

331   ECMWF forecasts, than for the UKMO, and especially the NCEP forecasts. The probabilistic post-processing of the weekly

332   NCEP forecasts seemed practically mandatory to produce reasonable predictions, but once implemented it provided

333   performance assessments almost comparable to those based on the UKMO forecasts. These results have important implications

334   for operational ETo forecasts, such as the U.S. national digital forecast database, one of the few operational products of its

335   type, which are based on the NCEP forecasts.

336   Unlike the probabilistic forecast metrics, the deterministic metrics (bias, RMSE and correlation) are low sensitive to the form

337   (deterministic or probabilistic) of post-processing. In particular, the RMSE and correlation seemed more affected by the choice

338   of the single or multi-model forecast strategy than the choice between the NGR, the AKD or the simple bias correction as post-

339   processing method. Whereas, RMSE and correlation provided by the BMA method are consistently worse at long lead times.

340   The daily errors under any post-processing were relatively large, but mostly random, and therefore tend to cancel out at weekly

341   scales. Therefore, while the RMSE varied between 12% and 20% of the daily totals, it represented between 8% and 11% of

342   the weekly totals. The RMSE for weekly ETo forecasts were in all cases more than 100% lower than for the persistence-based

343 ETo forecasts, and potentially more skillful than the forecasts that exploit the temporal autocorrelation of the ETo timeseries

344 (e.g., Landeras et al., 2009; Mohan and Arumugam, 2009).

4.2 Comparing the three probabilistic post-processing methods

346 The NGR and AKD based post-processing methods for the ECMWF forecasts produced comparable results, indicating that

347 the simple Gaussian predictive distribution from the NGR method represents well the uncertainty of the ETo predictions. The

348 methods led to similar distribution of the first two moments of the predictive probability function and similar performance

349 statistics (with the AKD based forecasts being just slightly better). However, the NGR method requires less computing time

350 and is more versatile since it can be applied to correct both single model and multi-model ensemble forecasts, while the AKD

351 method can only be applied to correct single model forecast. The NGR based predictive distribution function is also easier to

352 manipulate and interpret than the AKD based predictive distribution, which is given by an averaged sum of standard Gaussians.

353 The BMA method performed slightly less desirable compared to the NGR and AKD presumably due to issues with the

354 parameter identifiability. The implemented method uses the Expectation-Maximization (EM) algorithm to produce maximum

355 likelihood estimates of the fitting coefficients, which is susceptible to converge to local minima, especially when dealing with

356 multi-model forecasts with very different ensemble sizes (Vrugt et al., 2008). Archambeau et al. (2003) demonstrated that, in

357 presence of outliers or repeated values, this algorithm tends to identify local maximums of the likelihood of the parameters of

358 a Gaussian mixture model. Tian X. et al. (2012) found that adjusted BMA coefficient using both a quasi-Newtonian limited

359 memory algorithm and the Markov Chain Monte Carlo were more accurate than those fitted with the EM algorithm.

4.3 Multi-model ensemble versus single model forecasts

361 Daily multi-model ensemble forecasts performed better than daily ECMWF forecasts at short lead times (1-2 days) and over

362 the western and southern regions, while the ECMWF forecasts are better over the northeastern regions for longer lead days.

363 We observed similar patterns for the raw and simple bias corrected forecasts (Medina et al., 2018). When the effect of the

364 multi-model forecast is generally inconsistent at weekly scales, seemingly due to the variable impact of the forecasting strategy

365 with lead days. The observed behavior is associated with the performance of the ECMWF forecasts relative to the UKMO

366 forecasts. While the ECMWF forecasts are in general better than the UKMO and NCEP forecasts, they are much better over

367 the northeastern regions for medium lead times (4-6 days). The UKMO forecasts are in many cases the best at 1 and 2 lead

368 days, but tend to be the worst at the longest times (6-7 days), especially over these regions. The NCEP forecasts had a small

369 contribution with respect to the ECMWF and UKMO forecasts at short lead times. These forecasts are comparatively better

370 at longer lead times, but still keep a minor role with regard to the ECMWF forecasts.

371 When considering daily forecasts we adopted a length of the training period of 30 days and showed that by increasing the

372 length to 45 days the improvements were small (commonly lower than three percent). This seems a plausible range for future

373 works and represents an obvious advantage upon methods such as the analog forecast, which provide similar performance

374 (Tian and Martinez 2012 a, b, 2014) but require long training datasets. Gneiting et al. (2005) and Wilson (2007) found that

375   lengths between 30 and 40 days provided good and almost constant performance assessments of sea level pressure forecasts

376   post-processed with the NGR method, and temperatures forecasts post-processed with the BMA method, respectively.

4.4. Future outlook

378   It is worth noting that, while the ETo forecasts are produced for being used in agriculture, they were tested over USCRN

379   stations, which are not representative of agricultural settings. In real applications, the bias between the forecasts with no post-

380   processing and the measurements based on agricultural stations could be higher than the bias resolved in this study. A question

381   that should be addressed in the future studies is to what extent the improvements of the predictive distribution of the ETo

382   forecasts can be translated into a more reliable representation of the crop water use in agricultural lands and, ultimately, in

383   water savings and economic gains. Since the ETo estimations can have remarkable impacts on the soil moisture estimations

384   (Rodriguez-Iturbe et al., 1999), we envision that new studies relying on the combination of rainfall and ETo forecasts post-

385   processed with probabilistic methods will lead to considerable reductions on the uncertainty of soil moisture forecasts. New

386   attempts should also investigate the role of the emerging probabilistic post-processing techniques on ETo forecasts produced

387   from regional numerical weather prediction models, which have had improved spatial resolution and already been used in

388   different meteorological services (e.g., Baldauf et al. 2011; Seity et al. 2011; Hong and Dudhia, 2012; Bentzien and Friederichs,

389   2012).

5. Conclusions

391   This study for the first time evaluated probabilistic methods based on NGR, AKD, and BMA techniques for post-processing

392   daily and weekly ETo forecasts derived from single or multi-model numerical weather predictions. The different ETo

393   forecast strategies were compared against the simple linear regression bias correction method using both daily and weekly

394   forecasts, and also against persistence in the case of weekly forecasts. The probabilistic post-processing techniques largely

395   modified the spread of the original ETo forecasts, with very favorably impacts on the probabilistic forecast performance. They

396   corrected the notable under-dispersion and the poor consistency between the spread of the ETo forecasts and the dimension of

397   the errors, leading to better skill and reliability. The adjustments were crucial on the performance of the weekly NCEP

398   forecasts, followed by the weekly UKMO forecasts, whose bias corrected versions show a clear disadvantage compared with

399   the strategies that include the ECMWF forecasts.

400   The deterministic forecast performance based on the probabilistic methods were comparable to the linear regression bias

401   correction for both daily and weekly forecasts, and the skill is about 100% higher than those based on persistence in the case

402   of the weekly forecasts. The RMSE are between 12 and 20% for the daily totals and 8 and 11% for the weekly totals. The NGR

403   and AKD provided similar estimates of the first and second order moments of the predictive density distribution; they showed

404   similar effectiveness, but the NGR method exhibited higher stability and computational efficiency. Both NGR and AKD

405   post-processing methods outperformed the BMA method when considering daily forecasts at long lead times.

406 The multi-model forecasting provided benefits at daily scales compared to the ECMWF forecasting, while the benefits were

407 marginal at weekly scales. The multi-model ensemble forecasting seems a better choice when the UKMO forecasts are

408 comparable or slightly better than the ECMWF forecasts, such as at short (1-2 days) lead times and over the southern and

409 western regions. Post-processing single model forecast is a better choice than post-processing multi-model ensemble forecast

410 in the circumstances where the ECMWF forecasts perform considerably better than the UKMO and NCEP, such as at mid and

411 long lead times, especially over the northeastern regions. While we considered a length of the training period of 30 days for

412 daily post-processing, the increase of the training period to 45 days only led to minimal improvements. In conclusion, our

413 results suggest that the NGR post-processing of ETo forecasts generated from the ECMWF or ECMWF-UKMO predictions

414 is the most plausible strategy among those being evaluated, and is recommended for operational implementations.

Acknowledgement

416 This research was supported in part by the Alabama Agricultural Experiment Station and the Hatch program of the National

417 Institute of Food and Agriculture (NIFA), U.S. Department of Agriculture (USDA, Access No. 1012578), by the Auburn

418 University Intramural Grant Program, by the Auburn University Presidential Awards for Interdisciplinary Research, and by

419 the USDA-NIFA Agriculture and Food Research Initiative (AFRI) competitive grant.

Code/Data availability

421 Request for materials should be addressed to Di Tian.

Author contributions

423 HM and DT designed and conceptualized the research. HM implemented the design, performed data curation, analysis,

424 validation, visualization, and wrote the original draft. DT supervised the research, contributed by advice, and reviewed and

425 edited the manuscript.

Competing interests

427 The authors declare that they have no conflict of interest.

References

429 Allen, R. G., Pereira, L. S., Raes, D. and Smith, M.: Crop evapotranspiration-Guidelines for computing crop water

430 requirements-FAO, Irrigation and drainage paper 56, Fao, Rome, 300(9), p.D05109, 1998.

431   Archambeau, C., Lee, J. A. and Verleysen, M.: On Convergence Problems of the EM Algorithm for Finite Gaussian Mixtures,

432   In ESANN (Vol. 3, pp. 99-106), 2003.

433   Baldauf, M., Seifert, A., Förstner, J., Majewski, D., Raschendorfer, M. and Reinhardt, T.: Operational convective-scale

434   numerical weather prediction with the COSMO model: Description and sensitivities, Monthly Weather Review, 139(12),

435   pp.3887-3905, 2011.

436   Bauer, P., Thorpe, A. and Brunet, G.: The quiet revolution of numerical weather prediction, Nature, 525(7567): 47-55, 2015.

437   Bentzien, S. and Friederichs, P.: Generating and calibrating probabilistic quantitative precipitation forecasts from the high-

438   resolution NWP model COSMO-DE. Weather and Forecasting, 27(4), pp.988-1002, 2012.

439   Beran, R. and Hall, P.: Interpolated nonparametric prediction intervals and confidence intervals, Journal of the Royal Statistical

440   Society, Series B (Methodological), pp.643-652, 1993.

441   Bröcker, J. and Smith, L. A.: From ensemble forecasts to predictive distribution functions, Tellus A: Dynamic Meteorology

442   and Oceanography, 60(4), pp.663-678, 2008.

443   Buizza, R., Houtekamer, P. L., Pellerin, G., Toth, Z., Zhu, Y. and Wei, M.: A comparison of the ECMWF, MSC, and NCEP

444   global ensemble prediction systems, Monthly Weather Review, 133(5), pp.1076-1097, 2005.

445   Casella, G. and Berger, R. L. : Statistical inference (Vol. 2). Pacific Grove, CA: Duxbury, 2002.

446   Castro, F. X., Tudela, A. and Sebastià, M. T.: Modeling moisture content in shrubs to predict fire risk in Catalonia (Spain).

447   Agricultural and Forest Meteorology, 116(1-2), pp.49-59, 2003.

448   Chirico, G. B., Pelosi, A., De Michele, C., Bolognesi, S. F. and D'Urso, G.: Forecasting potential evapotranspiration by

449   combining numerical weather predictions and visible and near-infrared satellite images: an application in southern Italy, The

450   Journal of Agricultural Science, pp.1-9. https://doi.org/10.1017/S0021859618000084, 2018.

451   Fraley, C., Raftery, A. E. and Gneiting, T.: Calibrating multimodelmulti-model forecast ensembles with exchangeable and

452   missing members using Bayesian model averaging, Monthly Weather Review, 138(1), pp.190-202, 2010.

453   Fraley, C., Raftery, A. E., Sloughter, J. M., Gneiting T.: EnsembleBMA: Probabilistic Forecasting using Ensembles and

454   Bayesian Model Averaging. R package version 5.1.3. https://CRAN.R-project.org/package=ensembleBMA, 2016.

455   Glahn, H. R. and Lowry, D. A.: The use of model output statistics (MOS) in objective weather forecasting. J Appl Meteorol,

456   11(8): 1203-1211, 1972.

457   Glahn, H. R. and Ruth, D. P.: The new digital forecast database of the National Weather Service, Bulletin of the American

458   Meteorological Society, 84(2), pp.195-202, 2003.

459   Gneiting, T., Raftery, A. E., Westveld III, A. H. and Goldman, T.: Calibrated probabilistic forecasting using ensemble model

460   output statistics and minimum CRPS estimation, Monthly Weather Review, 133(5), 1098-1118., 2005.

461   Gneiting, T.: Calibration of medium-range weather forecasts, European Centre for Medium-Range Weather Forecasts,

462   Technical Memorandum No. 71, 2014.

463   Hagedorn, R., Buizza, R., Hamill, T. M., Leutbecher, M. and Palmer, T. N.: Comparing TIGGE multimodelmulti-model

464   forecasts with reforecast-calibrated ECMWF ensemble forecasts. Q J Roy Meteor Soc, 138(668): 1814-1827, 2012.

465   Hamill, T. M. and Colucci, S. J.: Verification of Eta–RSM short-range ensemble forecasts, Monthly Weather Review, 125(6),
466   pp.1312-1327, 1997.

467   Hamill, T. M. and Whitaker, J. S.: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and
468   application, Mon Weather Rev, 134(11): 3209-3229, 2006.

469   Hamill, T. M. et al.: Noaa's Second-Generation Global Medium-Range Ensemble Reforecast Dataset. B Am Meteorol Soc,
470   94(10): 1553-1565, 2013.

471   Hobbins, M., McEvoy, D. and Hain, C.: Evapotranspiration, evaporative demand, and drought, Drought and Water Crises:
472   Science, Technology, and Management Issues, pp.259-288, 2017.

473   Hong, S. Y. and Dudhia, J.: Next-generation numerical weather prediction: Bridging parameterization, explicit clouds, and
474   large eddies, Bulletin of the American Meteorological Society, 93(1), pp.ES6-ES9., 2012.

475   Ishak, A. M., Bray, M., Remesan, R. and Han, D.: Estimating reference evapotranspiration using numerical weather modelling,
476   Hydrological processes, 24(24), pp.3490-3509, 2010.

477   Klein, W. H. and Glahn, H. R.: Forecasting local weather by means of model output statistics, Bulletin of the American
478   Meteorological Society, 55(10), pp.1217-1227, 1974.

479   Landeras, G., Ortiz-Barredo, A. and López, J. J.: Forecasting weekly evapotranspiration with ARIMA and artificial neural
480   network models, Journal of irrigation and drainage engineering, 135(3), pp.323-334, 2009.

481   Leutbecher, M. and Palmer, T. N.: Ensemble forecasting, Journal of Computational Physics, 227(7), pp.3515-3539, 2008.

482   Mase, A. S. and Prokopy, L. S.: Unrealized potential: A review of perceptions and use of weather and climate information in
483   agricultural decision making, Weather, Climate, and Society, 6(1), pp.47-61, 2014.

484   Medina, H., Tian, D., Marin, F. R. and Chirico, G. B.: Comparing GEFS, ECMWF, and Postprocessing Methods for Ensemble
485   Precipitation Forecasts over Brazil, Journal of Hydrometeorology, 20(4), pp.773-790, 2019.

486   Medina, H., Tian, D., Srivastava, P., Pelosi, A. and Chirico, G. B.: Medium-range reference evapotranspiration forecasts for
487   the contiguous United States based on multimodelmulti-model numerical weather predictions, Journal of Hydrology, 562,
488   pp.502-517, 2018.

489   Mohan, S. and Arumugam, N.: Forecasting weekly reference crop evapotranspiration series, Hydrological sciences journal,
490   40(6), pp.689-702, 1995.

491   National Research Council of the National Academies: Completing the Forecast: Characterizing and Communicating
492   Uncertainty for Better Decisions Using Weather and Climate Forecasts, The National Academies Press, 124 pp, 2006:

493   Pelosi, A., Medina, H., Van den Bergh, J., Vannitsem, S., and Chirico, G. B.: Adaptive Kalman filtering for post-processing
494   ensemble numerical weather predictions, Mon Weather Rev, doi.org/10.1175/MWR-D-17-0084., 2017.

495   Pelosi, A., Medina, H., Villani, P., D'Urso, G. and Chirico, G. B.: Probabilistic forecasting of reference evapotranspiration
496   with a limited area ensemble prediction system, Agricultural water management, 178, pp.106-118, 2016.

497   Perera, K. C., Western, A. W., Nawarathna, B. and George, B.: Forecasting daily reference evapotranspiration for Australia
498   using numerical weather prediction outputs, Agr Forest Meteorol, 194: 50-63, 2014.

499   Prokopy, L. S., Haigh, T., Mase, A. S., Angel, J., Hart, C., Knutson, C., Lemos, M. C., Lo, Y. J., McGuire, J., Morton, L. W.
500   and Perron, J.: Agricultural advisors: a receptive audience for weather and climate information?, Weather, Climate, and
501   Society, 5(2), pp.162-167, 2013.

502   R Core Team: R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna,
503   Austria, http://www.R-project.org/, 2014.

504   Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M.: Using Bayesian model averaging to calibrate forecast
505   ensembles, Monthly Weather Review, 133(5), pp.1155-1174, 2005.

506   Rodriguez-Iturbe, I., Porporato, A., Ridolfi, L., Isham, V. and Coxi, D. R.: Probabilistic modelling of water balance at a point:
507   the role of climate, soil and vegetation, Proceedings of the Royal Society of London, Series A: Mathematical, Physical and
508   Engineering Sciences, 455(1990), pp.3789-3805, 1999.

509   Roulston, M. S. and Smith, L. A.: Combining dynamical and statistical ensembles, Tellus A: Dynamic Meteorology and
510   Oceanography, 55(1), pp.16-30, 2003.

511   Seity, Y., Brousseau, P., Malardel, S., Hello, G., Bénard, P., Bouttier, F., Lac, C. and Masson, V.: The AROME-France
512   convective-scale operational model, Monthly Weather Review, 139(3), pp.976-991, 2011.

513   Siegert, S.: SpecsVerification: Forecast Verification Routines for Ensemble Forecasts of Weather and Climate. R package
514   version 0.5-2. https://CRAN.R-project.org/package=SpecsVerificatio, 2017.

515   Silva, D., Meza, F. J. and Varas, E.: Estimating reference evapotranspiration (ETo) using numerical weather forecast data in
516   central Chile, Journal of hydrology, 382(1-4), pp.64-71, 2010.

517   Swinbank, R. et al.: The Tigge Project and Its Achievements. B Am Meteorol Soc, 97(1): 49-67, 2016.

518   Tian, D. and Martinez, C. J.: Comparison of two analog-based downscaling methods for regional reference evapotranspiration
519   forecasts, J Hydrol, 475: 350-364, 2012a

520   Tian, D. and Martinez, C. J.: Forecasting Reference Evapotranspiration Using Retrospective Forecast Analogs in the
521   Southeastern United States, J Hydrometeorol, 13(6): 1874-1892, 2012b

522   Tian, D. and Martinez, C. J.: The GEFS-based daily reference evapotranspiration (ETo) forecast and its implication for water
523   management in the southeastern United States. J Hydrometeorol, 15(3): 1152-1165, 2014.

524   Tian, X., Xie, Z., Wang, A. and Yang, X.: A new approach for Bayesian model averaging, Science China Earth Sciences,
525   55(8), 1336-1344, 2012.

526   Toth, Z., Talagrand, O., Candille, G. and Zhu, Y.: Probability and ensemble forecasts, Forecast Verification: A Practitioner's
527   Guide in Atmospheric Science, pp.137-163, 2003.

528   Vrugt, J. A., Diks, C. G. and Clark, M. P.: Ensemble Bayesian model averaging using Markov chain Monte Carlo sampling,
529   Environmental fluid mechanics, 8(5-6), pp.579-595, 2008.

530   Wang, X. and Bishop, C. H.: Improvement of ensemble reliability with a new dressing kernel, Quarterly Journal of the Royal
531   Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography, 131(607),
532   pp.965-986, 2005.

533   Wilks, D. S. and Hamill, T. M.: Comparison of ensemble-MOS methods using GFS reforecasts, Monthly Weather Review,

534   135(6), pp.2379-2390, 2007.

535   Wilks, D. S.: Statistical methods in the atmospheric sciences, (Vol 100), Academic press, 2011.

536   Wilson, L. J., Beauregard, S., Raftery, A. E. and Verret, R.: Calibrated surface temperature forecasts from the Canadian

537   ensemble prediction system using Bayesian model averaging, Monthly Weather Review, 135(4), pp.1364-1385, 2007.

538   Yuen, R., Baran, S., Fraley, C., Gneiting, T., Lerch, S., Scheuerer, M., and Thorarinsdottir, T.: ensembleMOS: Ensemble

539   Model Output Statistics. R package version 0.8.2. https://CRAN.R-project.org/package=ensembleMOS, 2018.

540   Zhao, T., Wang, Q. J. and Schepen, A.: A Bayesian modelling approach to forecasting short-term reference crop

541   evapotranspiration from GCM outputs, Agricultural and Forest Meteorology, 269, pp.88-101, 2019.

542

543

544

545

546

547

548

549

550

551

552

553

554

555

556

557
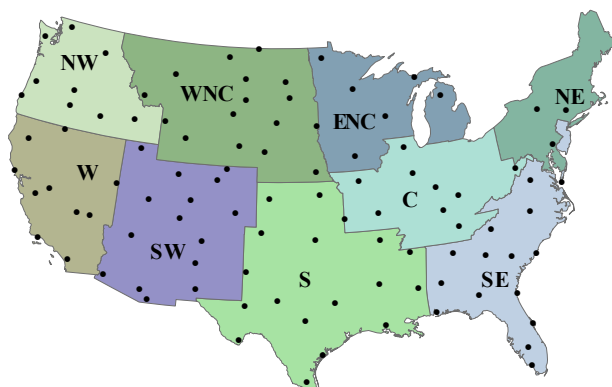
Figure 1. U.S. climate regions: NW (North West), WNC (West North Central), ENC (East North Central), NE (North East), C (Central), SE (South East), C (Central), S (South), SW (South West), W (West). The circles represent the sampled USCRN stations in the experiment.
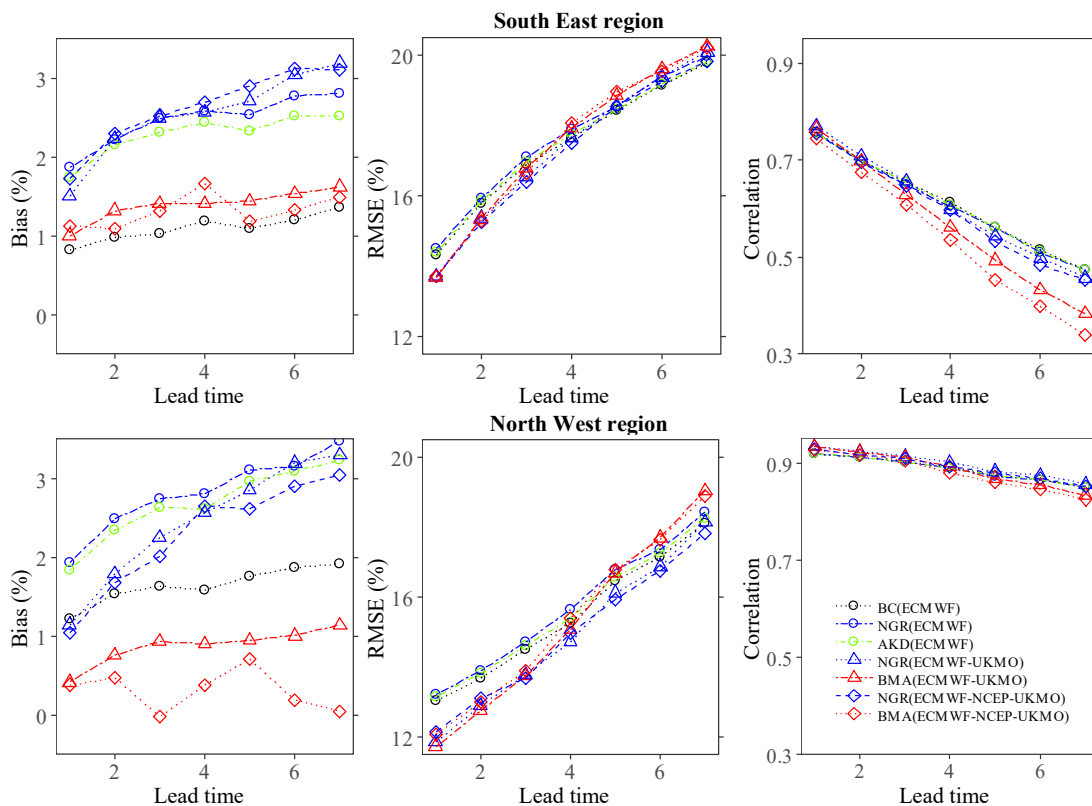
Figure 2. Relative ME, RMSE, and correlation for different lead times over the SE and NW regions.
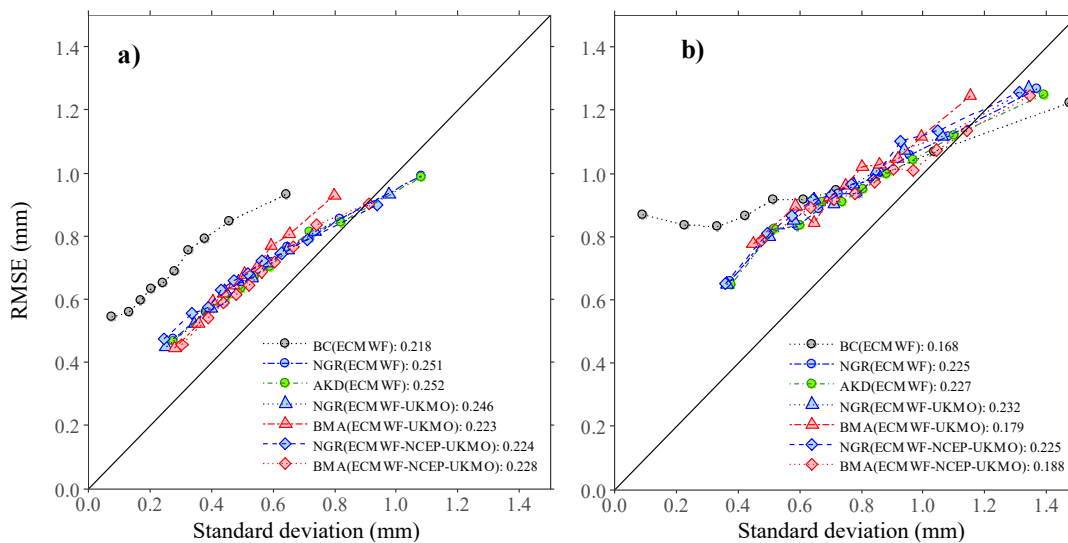
Figure 3. Binned spread-skill plots using all pairs of forecasts and observations at a) 1-day and b) 7-day lead. The correlation between the standard deviations and the absolute errors is reported after the colon. The solid line represents the 1:1 relationship.
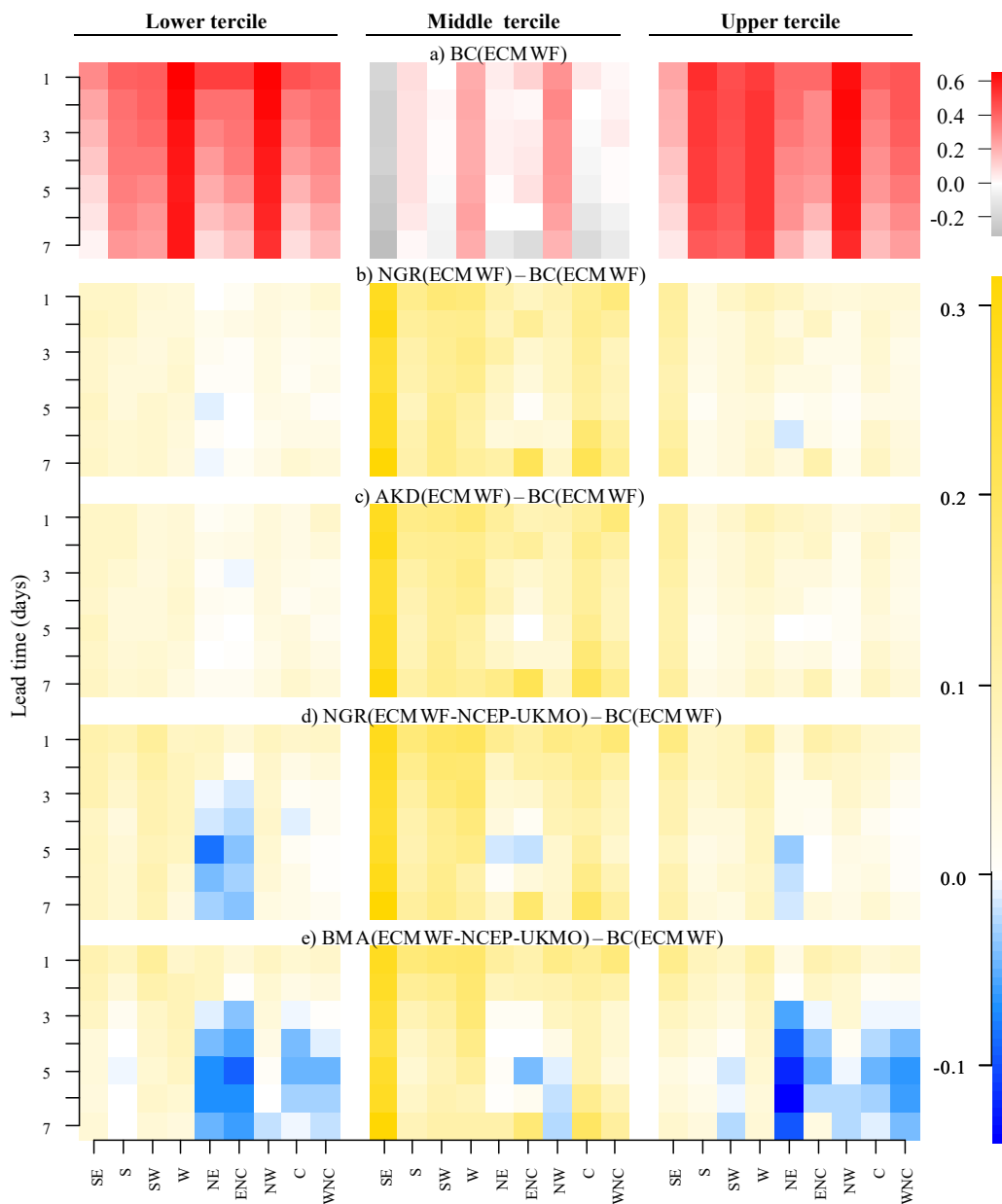
Figure 4. a) BSS of the ECMWF forecasts post-processed using simple bias correction (used as reference BSS values) and b-e) differences between the BSS of the ECMWF forecasts post-processed with the b) NGR and c) AKD methods and the ECMWF-NCEP-UKMO forecasts post-processed with the d) NGR and e) BMA methods and the reference BSS.
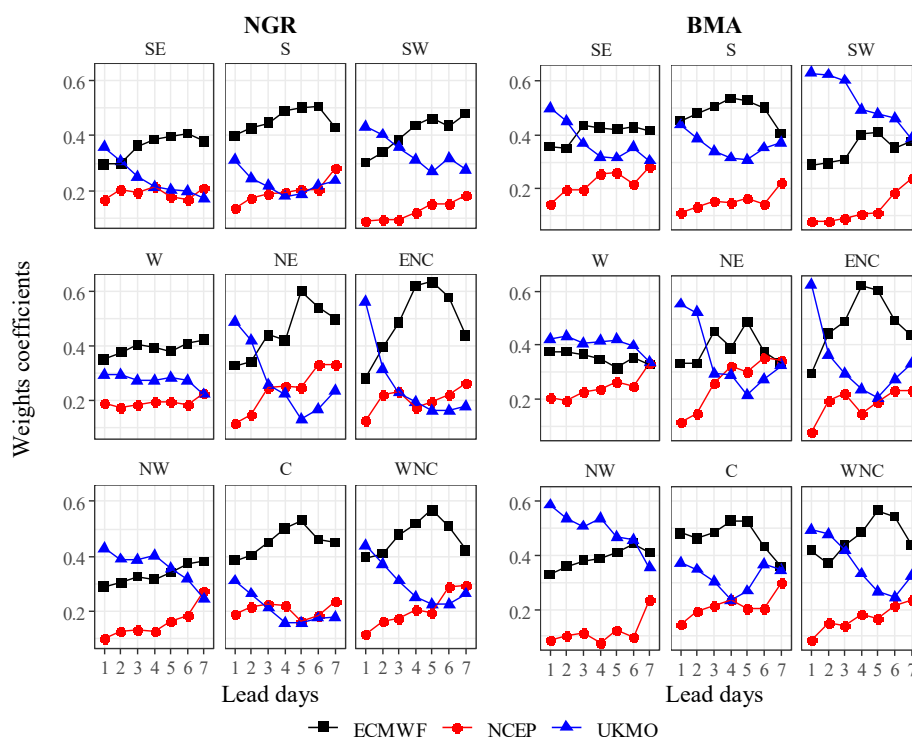
Figure 5. Regional mean weight coefficient *b* of the NGR technique (left panel) and the weight coefficient *w* of the BMA technique (right panel) for the post-processed ECMWF-NCEP-UKMO forecasts at different lead days.
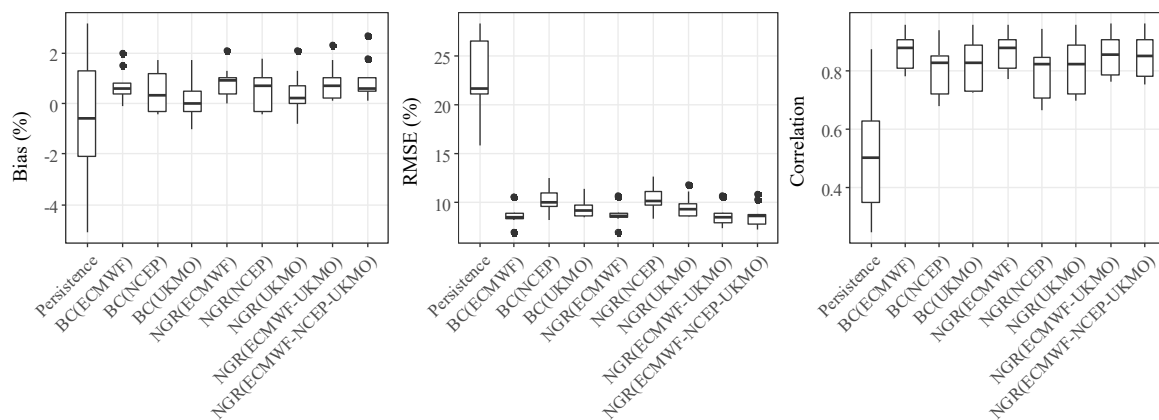
Figure 6. Relative bias, relative RMSE and correlation of weekly forecasts
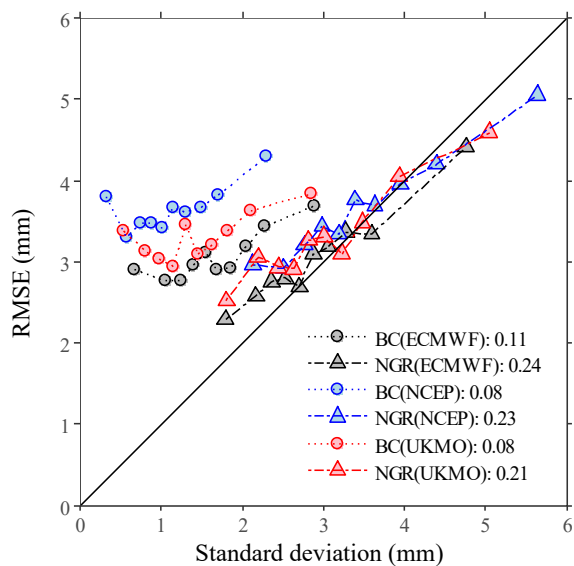
Figure 7. Binned spread-skill plot for the weekly forecasts using all pairs of forecasts and observations. The correlation between the standard deviations and the absolute errors is reported after the colon. The solid line represents the 1:1 relationship.
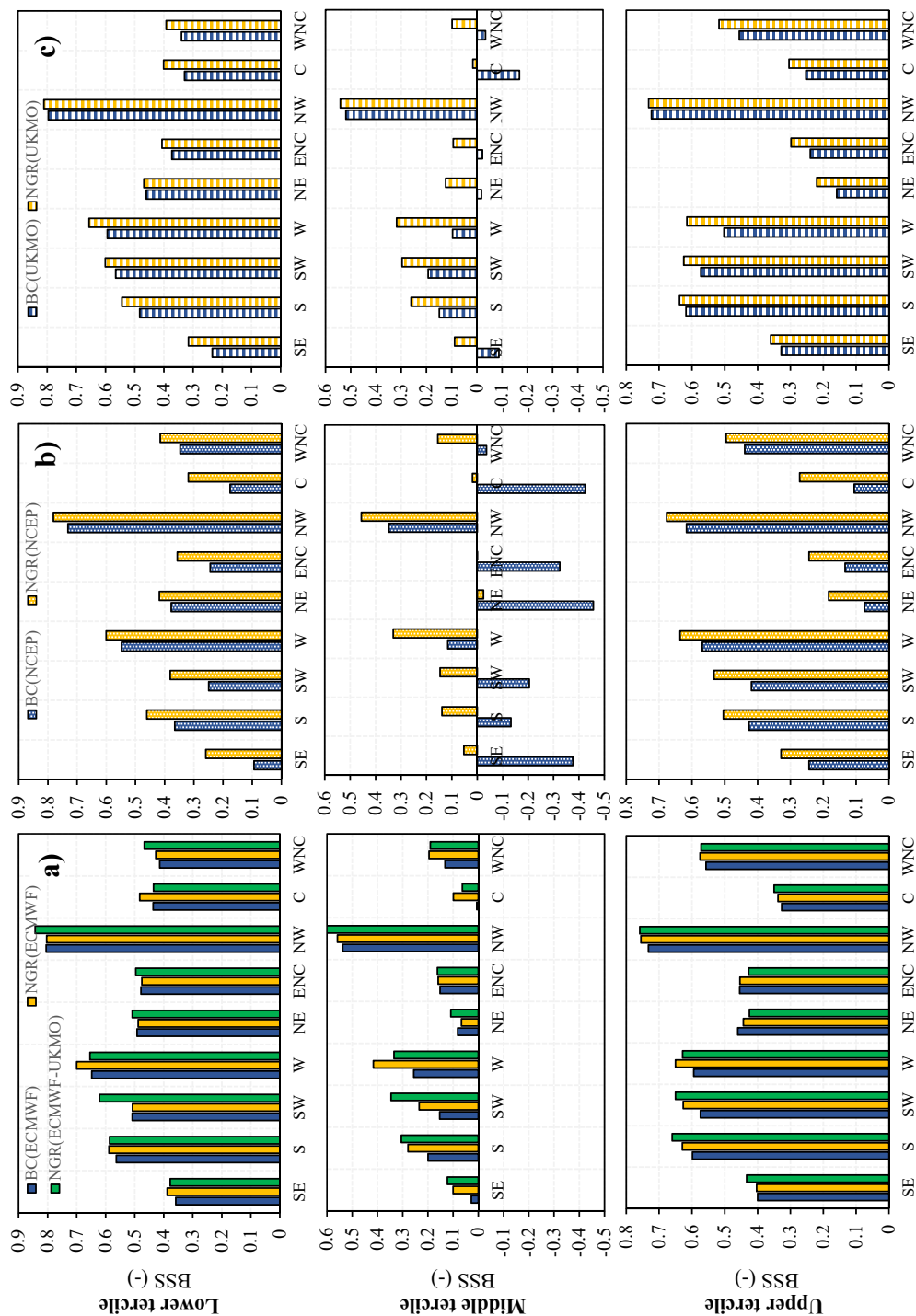
Figure 8. Comparison between BC and NGR based Brier Skill Scores considering a) ECMWF and ECMWF-UKMO forecasts, b) NCEP, and c) UKMO forecasts.

Table 1. Evaluated schemes for daily and weekly ETo forecasts, with different post-processing methods: BC (simple bias correction), NGR (nonhomogeneous Gaussian regression), AKD (affine kernel dressing), and BMA (Bayesian model averaging), and different model and ensemble schemes: ECMWF (European Centre for Medium-Range Weather Forecasts model), NCEP (National Centers for Environmental Prediction model), and UKMO (United Kingdom Meteorological office model), ECMWF-UKMO (ensemble of ECMWF and UKMO), ECMWF-NCEP-UKMO (ensemble of ECMWFMWF, NCEP, and UKMO).

	Persistence	BC			NGR					AKD	BMA		
		ECMWF	NCEP	UKMO	ECMWF	NCEP	UKMO	ECMWF-UKMO	ECMWF-NCEP-UKMO	ECMWF	ECMWF-UKMO	ECMWF-NCEP-UKMO
Daily		✓			✓			✓	✓	✓	✓	✓
Weekly	✓	✓	✓	✓	✓	✓	✓	✓	✓			

Table 2. Minimum, mean and maximum coverage ratios over all the climate regions and lead times for different methods. See the caption of Table 1 for explanations of the methods acronyms.

	BC ECMWF	NGR ECMWF	AKD ECMWF	NGR ECMWF-UKMO	BMA ECMWF-UKMO	NGR ECMWF-NCEP-UKMO	BMA ECMWF-NCEP-UKMO
Minimum coverage ratio	49.69	94.27	94.69	93.23	94.38	92.60	91.35
Mean coverage ratio	76.67	95.73	96.25	94.90	96.98	94.38	96.88
Maximum coverage ratio	93.13	98.02	98.33	97.29	99.38	96.56	99.58

Table 3. Spatial weighted average values of daily forecast metrics over all climate regions for different methods at lead days 1 and 7. See the caption of Table 1 for explanations of the methods acronyms. Numbers in bold indicate the best performance for each lead day.

	BC ECMWF		NGR ECMWF		AKF ECMWF		NGR ECMWF-UKMO		BMA ECMWF-UKMO		NGR ECMWF-NCEP-UKMO		BMA ECMWF-NCEP-UKMO	
	1 day	7 days	1 day	7 days	1 day	7 days	1 day	7 days	1 day	7 days	1 day	7 days	1 day	7 days
rBias (%)	0.822	1.203	1.695	2.682	1.626	2.419	1.327	2.735	0.632	0.939	1.394	2.778	**0.490**	**0.626**
rRMSE (%)	14.38	**19.64**	14.59	19.88	14.47	19.76	13.68	19.67	13.65	20.15	**13.59**	19.67	13.67	20.28
Bias (mm day^{-1})	0.038	0.057	0.080	0.128	0.077	0.115	0.063	0.131	0.029	0.046	0.067	0.134	0.005	0.006
RMSE (mm day^{-1})	0.708	0.950	0.718	0.961	0.716	0.958	0.682	0.965	0.681	0.990	0.681	0.971	0.685	1.002
Correlation	0.832	**0.652**	0.829	0.649	0.830	0.649	**0.843**	0.639	0.841	0.586	0.841	0.635	0.832	0.560
Coverage ratio	64.54	79.40	95.63	95.44	95.93	96.10	94.24	94.73	**96.51**	96.56	93.52	94.57	96.47	**97.24**
BSS_1st	0.442	0.232	0.492	0.279	0.492	**0.282**	**0.525**	0.274	0.519	0.240	0.521	0.271	0.513	0.225
BSS_2nd	0.042	-0.062	0.201	0.101	0.202	**0.101**	**0.224**	0.095	0.214	0.074	0.217	0.089	0.200	0.059
BSS_3nd	0.433	0.300	0.496	**0.359**	0.499	0.358	**0.519**	0.350	0.515	0.305	0.512	0.338	0.494	0.277

Table 4. Percentage differences (averaged over all lead times) of the ECMWF-UKMO and ECMWF-NCEP-UKMO forecast performance with the ECMWF forecast performance, after post-processing with the non-homogeneous Gaussian regression (NGR) method. See the caption of Table 1 for explanations of the forecast models acronyms.

| | Western climate regions | | | | | | Northern climate regions | | | | | |
| | SW | | W | | NW | | NE | | ENC | | WNC | |
	ECMWF-UKMO	ECMWF-NCEP-UKMO	ECMWF-UKMO	ECMWF-NCEP-UKMO	ECMWF-UKMO	ECMWF-NCEP-UKMO	ECMWF-UKMO	ECMWF-NCEP-UKMO	ECMWF-UKMO	ECMWF-NCEP-UKMO	ECMWF-UKMO	ECMWF-NCEP-UKMO
Bias	-26.753	-30.826	-9.111	9.421	-13.908	-18.800	-4.268	25.053	-2.149	-1.445	-10.119	0.761
RMSE	-4.682	-4.013	-3.455	-2.505	-3.973	-2.839	1.898	4.333	1.455	2.000	-1.313	-0.922
Correlation	1.760	0.627	0.947	0.707	1.19	0.607	-4.180	-4.600	-3.275	-3.137	-2.312	-2.062
Cov. ratio	-1.386	-2.094	-0.977	-1.194	-1.019	-1.144	-0.835	-1.656	-0.850	-0.986	-0.835	-1.402
BSS_1st	12.022	7.481	3.222	2.846	3.548	4.236	-11.999	-9.676	-9.643	-9.384	-3.680	-5.181
BSS_2nd	8.991	-6.504	5.792	9.044	4.984	3.961	-112.954	-93.092	-19.092	-13.642	-15.725	-27.949
BSS_3nd	2.295	-1.807	3.575	6.557	4.196	2.370	-9.105	-8.992	-6.420	-10.605	-4.595	-5.835

Table 5. Percentage differences (averaged over regions) of forecast performance of using 45 days training period with using 30 days training period for lead days 1 and 7. See the caption of Table 1 for explanations of the methods acronyms.

	NGR(ECMWF)		AKD(ECMWF)		NGR(ECMWF-UKMO)		NGR(ECMWF-NCEP-UKMO)	
	1 day	7 days	1 day	7 days	1 day	7 days	1 day	7 days
Bias	16.569	18.732	21.654	22.859	4.714	10.089	-0.496	7.070
RMSE	-0.701	-2.641	-1.007	-3.121	-0.404	-3.720	-0.045	-4.742
Correlation	-0.157	0.525	-0.141	0.605	-0.099	1.332	-0.467	0.741
Cov. Ratio	1.276	0.954	1.615	1.257	1.701	1.495	1.938	1.338
BSS_1st	-0.884	2.183	-1.164	2.761	-0.212	5.062	-2.600	6.277
BSS_2nd	-1.259	2.764	-1.283	5.680	3.614	8.959	-2.293	5.562
BSS_3nd	-0.382	-1.589	-0.904	-0.212	-1.340	2.632	-1.625	0.240

5   Table 6. Spatial weighted average values of weekly forecast metrics over all climate regions. See the caption of Table 1 for explanations of the methods acronyms.

	Persistence	BC			NGR				
		ECMWF	NCEP	UKMO	ECMWF	NCEP	UKMO	ECMWF-UKMO	ECMWF-NCEP-UKMO
rBias (%)	-0.288	0.683	0.296	0.097	0.846	0.496	0.305	0.764	0.814
rRMSE (%)	22.108	8.872	10.453	9.460	8.952	10.571	9.599	8.753	8.661
Bias (mm week^{-1})	-0.086	0.217	0.077	0.007	0.277	0.145	0.080	0.246	0.268
RMSE (mm week^{-1})	7.541	3.059	3.634	3.306	3.086	3.675	3.353	3.059	3.064
Correlation	0.530	0.872	0.806	0.835	0.870	0.801	0.829	0.863	0.856
Coverage (%)		78.40	48.07	62.92	99.29	98.58	98.13	97.74	97.40
BSS_1st		0.508	0.326	0.448	0.529	0.430	0.501	0.547	0.506
BSS_2nd		0.164	-0.147	0.069	0.238	0.150	0.204	0.255	0.225
BSS_3nd		0.528	0.371	0.468	0.553	0.461	0.515	0.558	0.550