Hydrol. Earth Syst. Sci., 24, 1011–1030, 2020
https://doi.org/10.5194/hess-24-1011-2020

Hydrology and Earth System Sciences
Open Access
EGU

Comparison of probabilistic post-processing approaches for improving numerical weather prediction-based daily and weekly reference evapotranspiration forecasts

Hanoi Medina and Di Tian

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

Correspondence: Di Tian (tiandi@auburn.edu)

Received: 15 July 2019 – Discussion started: 12 August 2019
Revised: 2 December 2019 – Accepted: 22 January 2020 – Published: 3 March 2020

Abstract. Reference evapotranspiration (ET_0) forecasts play an important role in agricultural, environmental, and water management. This study evaluated probabilistic post-processing approaches, including the nonhomogeneous Gaussian regression (NGR), affine kernel dressing (AKD), and Bayesian model averaging (BMA) techniques, for improving daily and weekly ET_0 forecasting based on single or multiple numerical weather predictions (NWPs) from the THORPEX Interactive Grand Global Ensemble (TIGGE), which includes the European Centre for Medium-Range Weather Forecasts (ECMWF), the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS), and the United Kingdom Meteorological Office (UKMO) forecasts. The approaches were examined for the forecasting of summer ET_0 at 101 US Regional Climate Reference Network stations distributed all over the contiguous United States (CONUS). We found that the NGR, AKD, and BMA methods greatly improved the skill and reliability of the ET_0 forecasts compared with a linear regression bias correction method, due to the considerable adjustments in the spread of ensemble forecasts. The methods were especially effective when applied over the raw NCEP forecasts, followed by the raw UKMO forecasts, because of their low skill compared with that of the raw ECMWF forecasts. The post-processed weekly forecasts had much lower rRMSE values (between 8 % and 11 %) than the persistence-based weekly forecasts (22 %) and the post-processed daily forecasts (between 13 % and 20 %). Compared with the single-model ensemble, ET_0 forecasts based on ECMWF multi-model ensemble ET_0 forecasts showed higher skill at shorter lead times (1 or 2 d) and over the southern and western regions of the US. The improvement was higher at a daily timescale than at a weekly timescale. The NGR and AKD methods showed the best performance; however, unlike the AKD method, the NGR method can post-process multi-model forecasts and is easier to interpret than the other methods. In summary, this study demonstrated that the three probabilistic approaches generally outperform conventional procedures based on the simple bias correction of single-model forecasts, with the NGR post-processing of the ECMWF and ECMWF–UKMO forecasts providing the most cost-effective ET_0 forecasting.

1 Introduction

Reference crop evapotranspiration (ET_0) represents the weather-driven component of the water transfer from plants and soils to the atmosphere. It plays a fundamental role in estimating mass and energy balance over the land surface as well as in agronomic, forestry, and water resource management. In particular, ET_0 forecasting is important for aiding water management decision-making (such as irrigation scheduling, reservoir operation, and so on) under uncertainty by identifying the range of future plausible water stress and demand (Pelosi et al., 2016; Chirico et al., 2018). While ET_0 forecasts have mostly been focused on the daily timescale (e.g., Perera et al., 2014; Medina et al., 2018), weekly ET_0 forecasts are also important for users. Studies show that both daily and weekly forecasts have increasing influence on the decision makers in agriculture (Prokopy et al., 2013; Mase and Prokopy, 2014) and water resource management (Hobbins et al., 2017). For example, irrigation is commonly scheduled considering both daily and weekly forecasts, while weekly evapotranspiration forecasts are useful for planning water allocation from reservoirs, especially in cases of shortages. Weekly ET_0 anomalies can also provide warnings regarding wildfires (Castro et al., 2003) and evolving flash drought conditions (Hobbins et al., 2017).

However, ET_0 forecasting is highly uncertain due to the chaotic nature of weather systems. In addition, ET_0 estimation requires full sets of meteorological data which are usually not easy to obtain. Due to the improvement of numerical weather predictions (NWPs), studies have recently emerged that forecast ET_0 using outputs from NWPs over different regions of the world (Silva et al., 2010; Tian and Martinez, 2012a, b, 2014; Perera et al., 2014; Pelosi et al., 2016; Chirico et al., 2018; Medina et al., 2018). Operationally, experimental ET_0 forecast products are being developed, such as the Forecast Reference EvapoTranspiration (FRET) product (https://digital.weather.gov/, last access: 26 February 2020), as part of the US National Weather Service (NWS) National Digital Forecast Database (NDFD; Glahn and Ruth, 2003) and the Australian Bureau of Meteorology's Water and Land website (http://www.bom.gov.au/watl, last access: 26 February 2020), which provide current and forecasted ET_0 at the continental scale.

The improved performance of NWPs in recent years is largely due to the improvement of physical, statistical representations of the major processes in the models as well as the use of ensemble forecasting (Hamill et al., 2013; Bauer et al., 2015). Nevertheless, NWP forecasts still commonly show systematic inconsistencies with measurements, which are often caused by inherent errors in the NWPs or local land–atmospheric variability that is not well resolved in the models. Post-processing methods, which are defined as any form of adjustment to the model outputs in order to get better predictions (e.g., Hagedorn et al., 2012), are highly recommended to attenuate, or even eliminate, these inconsistencies (Wilks, 2006). Until a few years ago, most post-processing applications only considered single-model predictions (i.e., predictions generated by a single NWP model) and addressed errors in the mean of the forecast distribution while ignoring those in the forecast variance (Gneiting, 2014). These procedures regularly adopted some form of model output statistics (MOS; Glahn and Lowry, 1972; Klein and Glahn, 1974) method, focusing on correcting current ensemble forecasts based on the bias in the historical forecasts.

As no forecast is complete without an accurate description of its uncertainty (National Research Council of the National Academies, 2006), the dispersion of the forecast ensemble often misrepresents the true density distribution of the forecast uncertainty (Krzysztofowicz, 2001; Smith, 2001; Hansen, 2002). The ensemble forecasts are, for example, commonly under-dispersed (e.g., Buizza et al., 2005; Leutbecher and Palmer, 2008), which cases the probabilistic predictions to be overconfident (Wilks, 2011). Therefore, another generation of probabilistic techniques has been proposed to also address dispersion errors in the ensembles (Hamill and Colucci, 1997; Buizza et al., 2005; Pelosi et al., 2017), in some cases via the manipulation of multi-model weather forecasts.

Nonhomogeneous Gaussian regression (NGR; Gneiting et al., 2005), Bayesian model averaging (BMA; Raftery et al., 2005; Fraley et al., 2010), extended logistic regression (ELR; Wilks et al., 2009; Whan and Schmeits, 2018), quantile mapping (Verkade et al., 2013), and the family of kernel dressing (Roulston and Smith, 2003; Wang and Bishop, 2005), such as the affine kernel dressing (AKD; Brocker and Smith, 2008), are state-of-the-art probabilistic techniques (Gneiting, 2014). However, ELR has been reported to fall short with respect to using the information contained in the ensemble spread in efficient way (Messner et al., 2014), whereas the quantile mapping method has been found to degrade rather than improve the forecast performance under some circumstances (Madadgar et al., 2014). NGR, AKD, and BMA are sometimes considered as variants of dressing methods (Brocker and Smith, 2008), as they produce a continuous forecast probability distribution function (pdf) based on the original ensemble. This property makes them particularly useful for decision-making (Gneiting, 2014) compared with methods that provide post-processed ensembles. Another common advantage is that they perform equally well with relatively short training datasets (Geiting et al., 2005; Raftery et al., 2005; Wilks and Hamill, 2007). A limitation of NGR (compared with the AKD and BMA methods) is that the resulting forecast pdf is invariably Gaussian, whereas a limitation of AKD is that it only considers single-model ensembles. Instead, the NGR and AKD methods provide more flexible mechanisms for simultaneous adjustments in the forecast mean and spread–skill (Brocker and Smith, 2008).

Studies have suggested that the post-processing of NWP-based ET_0 forecasts are crucial for informing decision-making (e.g., Ishak et al., 2010). Medina et al. (2018) compared single- and multi-model NWP-based ensemble ET_0 forecasts, and the results showed that the performance of the multi-model ensemble ET_0 forecasts was considerably improved via a simple bias correction post-processing and that the bias-corrected multi-model ensemble forecasts were generally better than the single-model ensemble forecasts. In reality, while most applications for ET_0 forecasting have involved some form of post-processing, these have been often limited to simple MOS procedures of single-model ensembles (e.g., Silva et al., 2010; Perera et al., 2014). The poor treatment of uncertainty and variability is considered to be a main issue affecting users' perceptions and adoptions of weather forecasts (Mase and Prokopy, 2014). The appropriate representation of the second and higher moments of the ET_0 forecast probability density is especially important to predict extreme values, as shown by Williams et al. (2014). Therefore, the use of probabilistic post-processing techniques, such as NGR, AKD, and BMA, may greatly enhance the overall performance of the ET_0 forecasts compared with simple MOS procedures.

Only a few studies have considered probabilistic methods for the post-processing of ET_0 forecasts; these include the works of Tian and Martinez (2012a, b, 2014) and, more recently, Zhao et al. (2019). The former authors showed the analog forecast (AF) method to be useful for the post-processing of ET_0 forecasts based on Global Forecast System (GFS, Hamill et al., 2006) and Global Ensemble Forecast System (GEFS, Hamill et al., 2013) reforecasts. Tian and Martinez (2014) found that water deficit forecasts produced with the post-processed ET_0 forecasts had higher accuracy than those produced with climatology. Zhao et al. (2019), in contrast, improved the skill and the reliability of the Australian BoM model using a Bayesian joint probability (BJP) post-processing approach, which is based on the parametric modeling of the joint probability distribution between forecast ensemble means and observations. However, a main disadvantage of the BJP method compared with the aforementioned state-of-the-art probabilistic approaches is that, while the probabilistic approaches transform the spread of the ensembles, they rely on the mean of retrospective reforecasts, thereby neglecting information about their dispersion. The AF approach has the disadvantages of requiring long time series of retrospective forecasts and possibly being unsuitable for extreme events forecasting (e.g., Medina et al., 2019). The use of new ET_0 forecasting strategies relying on the post-processing of single- and multi-model ensemble forecasts with the NGR, AKD, and the BMA probabilistic techniques provide good opportunities to improve the predictions.

In this paper, we address several scientific questions that have not been adequately studied in previous literature, including the following:

- How effective are state-of-the-art probabilistic post-processing methods compared with the traditional MOS bias correction methods for post-processing ET_0 forecasts?
- Is it worth implementing probabilistic post-processing for multi-model rather than single-model ensemble forecasting?

For the first time, this work aims to evaluate and compare multiple strategies for post-processing both daily and weekly ET_0 forecasts using the NGR, AKD, and BMA approaches. The study represents a major step forward with respect to Medina et al. (2018), who evaluated the performance of raw and linear regression bias-corrected daily ET_0 forecasts produced with single- and multi-model ensemble forecasts. It provides a broad characterization of the performance for different probabilistic post-processing strategies but also diagnoses the causes of better and worse performance.

2 Methods and datasets

2.1 The probabilistic methods

The NGR, AKD, and BMA techniques follow a common strategy: they yield a predictive probability density function (pdf) of the post-processed forecasts y given the raw forecasts x and some fitting parameters θ ($p(y|x,\theta)$). The parameters θ are fitted using a training dataset of ensemble forecasts and observations, as in the MOS techniques. A brief description of each technique is given in the following.

2.1.1 Nonhomogeneous Gaussian regression

Nonhomogeneous Gaussian regression (Gneiting et al., 2005) produces a Gaussian predictive (pdf) based on the current ensemble (of typically multi-model) forecasts. If x_{ij} denotes the jth ($j=1,\ldots,m_i$) ensemble forecast member of model i ($i=1,\ldots,n$), then $p(y|x,\theta)\sim\mathcal{N}(\mu,v)$; here, the mean

$$\mu = a+\sum\nolimits_{i=1}^{n} b_i\bar{x}_i \tag{1}$$

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

$$v=c+dS^2 \tag{2}$$

is a linear function of the ensemble variance S^2. The fitting parameters a, b_i, c, and d are determined by minimizing the continuous rank probability score (CRPS) using the training set of forecasts and observations. Notice that parameters a, c, and d are indistinguishable among members; therefore, b_i can be seen as a weighting parameter that reflects the better or worse performance of one model compared with the others. The NGR technique is implemented in R (R Core Team) using the ensembleMOS package (Yuen et al., 2018).

2.1.2 Affine kernel dressing

The affine kernel dressing method (Bröcker and Smith, 2008) only considers single-model ensemble forecasts. It estimates $p(y|x,\theta)$ using a mixture of normally distributed variables

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

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

$$z_j=ax_j+r_1+r_2\bar{x} \tag{4}$$

and

$$\sigma^2=h_\mathrm{s}^2\left(s_1+s_2u(\mathbf{z})\right). \tag{5}$$

Here, h_s is the Silverman's factor (Bröcker and Smith, 2008)l; $u(\mathbf{z})$ is the variance of $\mathbf{z}$; and a, r_1, r_2, s_1, and s_2

are fitting parameters obtained by minimizing the mean ignorance score. For clarity, we use the same nomenclature for the parameters as 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 (Brocker and Smith, 2008)

$$v = h_s^2 s_1 + a^2 \left(1 + h_s^2 s_2\right) S^2 = c^* + d^* S^2. \quad (6)$$

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

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

2.1.3 Bayesian model averaging

The BMA method (Raftery et al., 2005; Fraley et al., 2010) also produces a mixture of normally distributed variables (as in the AKD method), but they are based on multi-model ensemble forecasts. In this case, the predictive pdf is given by a weighted sum of component pdfs, $g_i(y|x_{i,j};\theta_i)$, with one for each member, as follows:

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

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

$$\sum\nolimits_{i=1}^{n} m_i w_i = 1.$$

In this study, the component pdfs are assumed normal, as in the affine kernel dressing method. Estimates of w_is and θ_is are produced by maximizing the likelihood function using an expectation-maximization algorithm (Casella and Berger, 2002). The BMA technique is implemented using the ensembleBMA R package (Fraley et al., 2016).

2.2 Measurement and forecast datasets

ET$_0$ observations and forecasts were computed using the FAO-56 PM equation (Allen et al., 1998), with daily meteorological data as inputs. They covered the same period: between May and August from 2014 to 2016. The observations used daily measurements of minimum and maximum temperature, minimum and maximum relative humidity, wind speed, and surface incoming solar radiation from 101 US Climate Reference Network (USCRN) weather stations. The USCRN stations are distributed over nine climatologically consistent regions in CONUS (Fig. 1). The ET$_0$ forecasts used daily maximum and minimum temperature, solar radiation, wind speed, and dew point temperature reforecasts from the European Centre for Medium-Range Weather Forecasts model (ECMWF) outputs, United Kingdom Meteorological Office (UKMO) model outputs, and the National Centers for Environmental Prediction (NCEP) model from the THORPEX Interactive Grand Global Ensemble (TIGGE; Swinbank et al., 2016) database at each of these stations, considering a maximum lead time of 7 d. We used the same models as Medina et al. (2018) for comparison purposes, and because they are considered to be among the most skillful globally (e.g., Hagedorn et al., 2012). The forecasts were interpolated to the same 0.5° × 0.5° grid using the TIGGE data portal. The weekly forecasts accounted for the sum of the daily predictions generated on a specific day of each week, and the weekly observations considered the sum of the daily observations over the corresponding forecasting days; thus, the weekly observations were independent of one another. In this 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 forecast errors before post-processing, it could also affect the multivariate dependence of the weather variables. Hagedorn et al. (2012) showed that post-processing can not only address the discrepancies related to the model's spatial resolution, but it can also serve as a means of downscaling the forecasts.

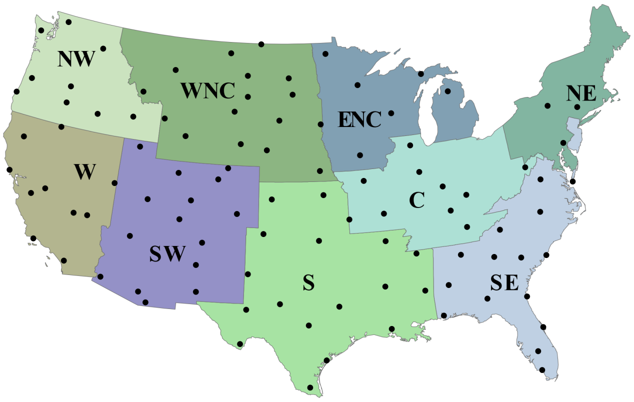

Figure 1. US climate regions: NW (Northwest), WNC (West north central), ENC (East north central), NE (Northeast), C (Central), SE (Southeast), S (South), SW (Southwest), and W (West). The circles represent the sampled USCRN stations in the experiment.

2.3 Post-processing schemes

2.3.1 Training and verification periods

The training data for the daily post-processing comprised the pairs of daily forecasts and corresponding observations from 30 d prior to the forecast initial day, as in Medina et al. (2018). Instead, the training data for the weekly post-processing included all of the other pairs of weekly forecasts and observations available for the forecast location, similar to a leave-one-out cross-validation framework. In the study, both the daily and weekly forecasts were verified for events between June and August from 2014 to 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, AKD, and BMA methods considering both daily and weekly forecasts. Here, the current forecast bias is estimated as a linear function of the forecast mean, and the members of the ensemble are shifted accordingly. The function is calibrated using the forecast mean and the actual biases based on the same training periods as for the other post-processing methods. Persistence is also used as a baseline approach for weekly forecasts, considering its applicability in productive systems. In this case, the ET_0 for a current week is estimated as the observed ET_0 during the previous week.

2.3.3 Forecasting experiments

Table 1 summarizes the daily and weekly NWP-based ET_0 forecasting experiments based on different post-processing methods and model combinations. The analyses of the daily forecasts put more emphasis on the differences among the post-processing methods. They include an examination of the effect of the duration of the training period on the forecast assessments as well as the regression weights from the tested post-processing methods. In contrast, the weekly forecasts put more emphasis on the differences among the several single- and multi-model ET_0 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 ET_0 forecasts. For consistency purposes, the metrics of the tested methods were assessed using 50 random samples, i.e., the same number of samples as the number of members in the bias-corrected ECMWF forecasts. The deterministic ET_0 forecast was produced by taking the average of the ensemble members. The deterministic forecast performance was assessed using the bias or mean error (ME) and the relative ME (rME), the root-mean-square error (RMSE) and the relative RMSE (rRMSE), and the correlation (ρ), which are common measures of agreement in many studies. The absolute bias and relative bias were calculated and reported.

The ME and rME were computed as follows:

$$\mathrm{ME}=\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\bar{f_i}-o_i\right) \tag{8}$$

$$\mathrm{rME}=\frac{\sum_{i=1}^{n}\left(\bar{f_i}-o_i\right)}{n\bar{o}}, \tag{9}$$

where $\bar{f_i}$ represents the average ensemble forecast for the event i ($i=1\ldots n$), o_i is the corresponding observation, and $\bar{o}$ is the mean observed data.

Table 1. Evaluated schemes for daily and weekly ET_0 ensemble forecasts using different post-processing methods, including BC (simple bias correction), NGR (nonhomogeneous Gaussian regression), AKD (affine kernel dressing), and BMA (Bayesian model averaging), and different model and ensemble schemes, including ECMWF (European Centre for Medium-Range Weather Forecasts model), NCEP (National Centers for Environmental Prediction model), and UKMO (United Kingdom Meteorological office model) ensemble forecasts as well as ECMWF–UKMO (ensembles of ECMWF and UKMO) and ECMWF–NCEP–UKMO (ensembles of ECMWF, NCEP, and UKMO) ensemble forecasts.

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

The RMSE and the rRMSE were computed as

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\bar{f}_i-o_i\right)^2} \tag{10}$$

$$\mathrm{rRMSE}=\frac{\sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\bar{f}_i-o_i\right)^2}}{\bar{o}}. \tag{11}$$

The correlation was obtained as follows:

$$\rho=\frac{\sum\nolimits_{i=1}^{n}\left(\bar{f}_i-\overline{\overline{f}}\right)(o_i-\bar{o})}{s_{\bar{f}}s_{\mathrm{o}}}, \tag{12}$$

where $\overline{\overline{f}}$ is the mean of the average ensemble forecast, and $s_{\bar{f}}$ and s_{o} are the standard deviation of the average forecasts and the observations, respectively.

The probabilistic forecast performance was assessed using a range histogram, the spread–skill relationship (see Wilks, 2011), and the forecast coverage as measures of the forecast reliability; the Brier skill score (BSS) as a measure of the skill; and the continuous rank probability score (CRPS) to provide an overall view of the performance (Hersbach, 2000), as the latter is simultaneously sensitive to both errors in location and spread.

Here, reliability refers to statistical consistency (as in Toth et al., 2003), which is met when the observations are statistically indistinguishable from the forecast ensembles (Wilks, 2011). To obtain the rank histogram, we get the rank of the observation when merged into the ordered ensemble of ET_0 forecasts and then plot the rank's histogram. The spread–skill relationships are represented as binned-type plots (e.g., Pelosi et al., 2017), accounting for the mean of the ensemble standard deviation deciles (as an indication of the ensemble spread) against the mean RMSE of the forecasts in each decile over the verification period. The plots include the correlation between these two quantities. Calibrated ensembles should show a 1:1 relationship between the standard deviations and the RMSE. If the forecasts are unbiased and the spread is small compared with the RMSE, the ensembles tend to be under-dispersive. The inverse of the spread provides an indication of sharpness, which is the level of "compactness" of the ensemble (Wilks, 2011).

In addition to the spread–skill relationship, we also report the ratio between the observed and nominal coverage (hereinafter referred to as the coverage ratio). The coverage of a $(1-\alpha)100\,\%$, $\alpha\in(0,\,1)$, central prediction interval is the fraction of the observations from the verification dataset that lie between the $\alpha/2$ and $1-\alpha/2$ quantiles of the predictive distribution. It is empirically assessed by considering the observations that lie between the extreme values of the ensembles. The nominal or theoretical coverage of a calibrated predictive distribution is $(1-\alpha)100\,\%$. A calibrated forecast of m ensemble members provides a nominal coverage of about a $(m-1)/(m+1)100\,\%$ central prediction interval (e.g., Beran and Hall, 1993). For example, an ensemble of 50 members provides a 96 % central prediction interval. The ratio between the observed and nominal coverages provides a quantitative indicator of the quality of the forecast dispersion under unbiasedness: a ratio lower or higher than 1 suggests that the forecasts tend to be under-dispersive or over-dispersive, respectively.

The BSS is computed as follows:

$$\mathrm{BSS}=1-\frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{clim}}}, \tag{13}$$

where BS is the Brier score of the forecast and is calculated as

$$\mathrm{BS}=\frac{1}{N}\sum_{i=1}^{N}(p_i-o_i)^2. \tag{14}$$

Here, p is the forecast probability p of the event (which is estimated based on the ensemble), and o is equal to 1 if the event occurs and 0 otherwise.

$\mathrm{BS}_{\mathrm{clim}}$ in Eq. (8) represents the Brier score of the sample climatology, computed as follows (Wilks, 2010):

$$\mathrm{BS}_{\mathrm{clim}}=\bar{o}\,(1-\bar{o})\,, \tag{15}$$

where $\bar{o}$ is the sample climatology computed as the mean of the binary observations o_i in the verification dataset.

In this study, we compute the BSS associated with the tercile events of the ET_0 forecasts (upper or first, middle or second, and lower or third terciles). Therefore, the sample climatology is equal to $0.3\bar{3}$ and $\mathrm{BS}_{\mathrm{clim}}$ is equal to $0.2\bar{2}$.

The CRPS was computed as follows:

$$\mathrm{CRPS}=\frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\left(F_i^{\mathrm{f}}(h)-F_i^{\mathrm{o}}(h)\right)^2\mathrm{d}h, \tag{16}$$

where F^{f} and F^{o} are the cumulative distribution function of the forecast and the observations, respectively, and h represents the threshold value; $F_i^{\mathrm{o}}(h)=H(h-o_i)$, where H is the Heaviside function, which is 0 for $h<o_i$ and 1 for $h\geq o_i$.

3 Results

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

3.1.1 Deterministic forecast performance

Figure 2 shows the rME and rRMSE as well as the correlation of the forecasts post-processed using different approaches over the southeast (SE) and northwest (NW) regions. These regions are representative of the eastern and western zones, which tended to provide the worst and best rRMSE values and correlations, respectively. In general, the probabilistic post-processing methods add no additional skill to the deterministic forecast performance compared with the simple bias correction. While the rRMSE values are relatively high, the rME values are very low, which indicates that the errors are mostly random. The BMA and the simple linear regression methods provided lower bias values than the NGR and AKD methods. However, the BMA method provided higher rRMSE values and lower correlations than the other three methods at long lead times. The rRMSE values and the correlations tended to be more variable among lead times and regions than among post-processing methods, whereas the opposite was found for the rME values. In addition, the changes in rRMSE and correlation values with lead time tended to be larger over the eastern regions.

3.1.2 Probabilistic forecast performance

Figure 3 shows the spread–skill relationship and the rank histograms using all pairs of forecasts and observations for lead times of 1 and 7 d. The spread–skill relationship shows that the probabilistic post-processing methods considerably improved the reliability of the ET_0 forecasts compared with the linear regression bias correction. The former methods tend to correct evident shortcomings in the ensemble raw forecasts which are unresolved by simple post-processing, i.e., the considerable under-dispersion at short lead times, and the poor consistency between the ensemble spread and the RMSEs at longer lead times. The adjustments had a low cost in terms of sharpness, judging by the range of ensemble spreads for the different line plots, but seemed slightly insufficient. The correlations between the ensemble standard deviation and the RMSE were fairly low, suggesting a limited predictive ability of the spread (Wilks, 2011). Nonetheless, they were consistently higher for probabilistic post-processing methods (compared with the linear regression method) and at short lead times (compared with long lead times). The rank histograms in Fig. 3 show that the probabilistic methods provided a better calibration than the linear regression approach at lead times of both 1 and 7 d, but the improvements were considerably larger at 1 d. At the short lead time, the three methods slightly over-forecasted ET_0, suggesting that departures from the predictive mean have a negative skew, but, in general, they were fairly confident. In this case, all of the methods provided almost the same result. At the long lead time, there was also an overestimation and then a positive bias, in addition to a slight U-shaped pattern; this was associated with some under-dispersion in the range of the low and medium observations, which is coherent with the spread–skill relationships. These issues are more pronounced when using the BMA method and less pronounced when using the AKD methods. Scheuerer and Büermann (2014) reported similar issues when post-processing ensemble forecasts of temperatures using the NGR method and a version of the BMA method. Conversely, the calibration was affected little by the choice of a single- or multi-model strategy for a given post-processing method. Nevertheless, the probabilistic methods provided a coverage ratio close to 100 % that was independent of the lead time (see Table 2) and the region (not shown). The simple bias correction method instead provided coverage ratios that were much lower and more variable among regions (see Table 2) and lead times.

The NGR and AKD methods provided a better Brier skill score (BSS) than the BC method for the three categories of ET_0 values, with improvements being higher for the middle tercile than for the lower and upper terciles (Fig. 4). The BMA-based skill scores tended to decrease with lead time. In the western regions (SW, W, and NW) and at short lead times, the multi-model ensemble forecasts post-processed using NGR were the most skillful; in the other cases, the ECMWF forecasts post-processed using the NGR and AKD methods tended to be best. The differences in the BSS among regions were larger at longer lead times because the skill decreased more sharply over the eastern regions. This issue is somewhat addressed by the NGR and AKD methods based on the ECMWF.

3.1.3 Summary of average performance for daily forecast

Table 2 shows the average performance for the lead times of 1 and 7 d by weighting the values of each metric according to the number of stations in each region. The ECMWF–UKMO forecasts post-processed using the NGR method were best at short lead times (1–2 d), whereas the ECMWF forecasts post-processed using the AKD and NGR methods were the first and second best at longer lead times. The BMA method performed well at short lead times but poorly at long times, whereas the simple bias correction method performed well for deterministic forecasts but poorly for the probabilistic forecasts. The forecast performance across climate regions is also associated with the choice of the ECMWF ensemble forecasts or the multi-model ensemble forecasts (Table A1 in the Appendix). The single-model ECMWF forecasts performed better over northern climate regions than the multi-model ensemble forecasts, whereas the multi-model showed better performance than any single-model forecast over the western regions. The performance over the other regions was more variable among strategies. The performance of the ECMWF–UKMO forecasts was generally better than that of the ECMWF–NCEP–UKMO forecasts (see Table A1, and Figs. 2 and 4). Unlike other performance metrics, the coverage was mostly better for the ECMWF ensemble forecasts than for the multi-model ensemble forecasts. Our CRPS values are comparable with those reported by Osnabrugge (2019) based on the ECMWF ensemble forecasts of potential evapotranspiration over the Rhine Basin in Europe.

3.1.4 Effect of the training period length

The choice of an "optimum" training period is an important issue related to the operational use of post-processing techniques for ET_0 forecasts. Here, we compared the perfor-

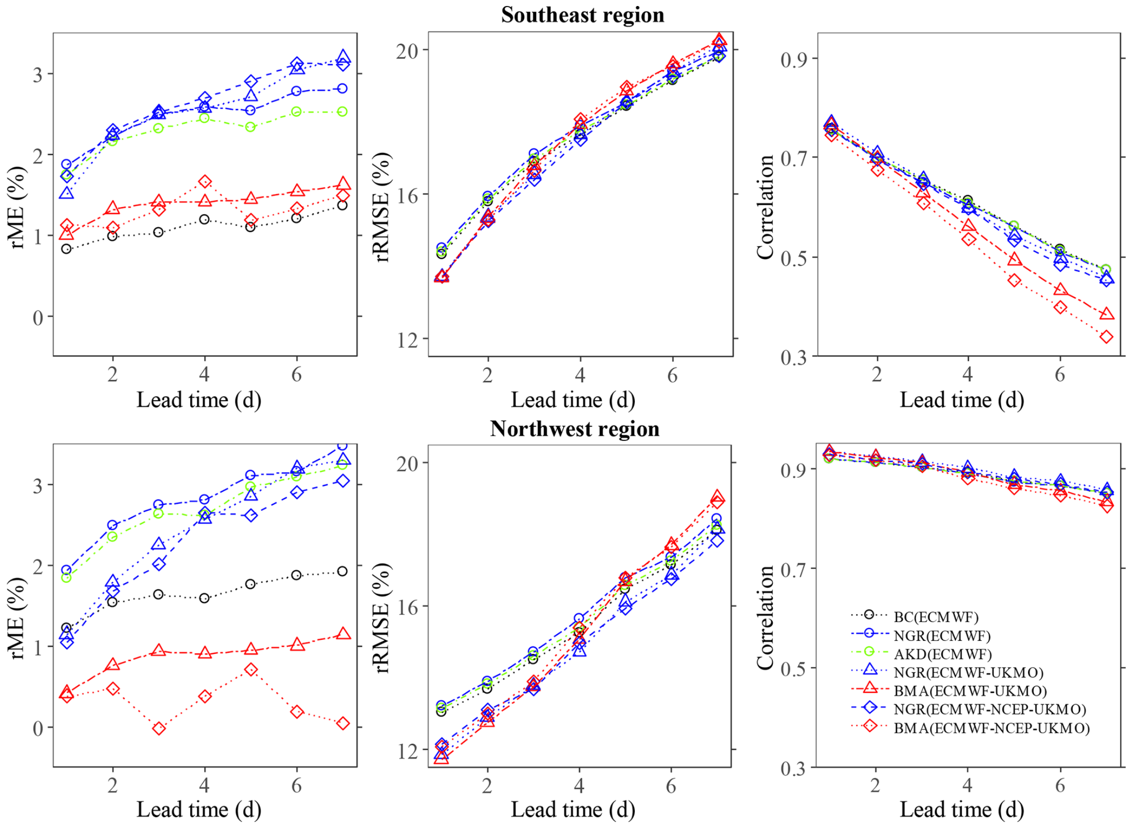

Figure 2. Relative mean error (rME), relative root-mean-square error (rRMSE), and correlation values considering daily forecasts for different lead times over the southeast and northwest regions.

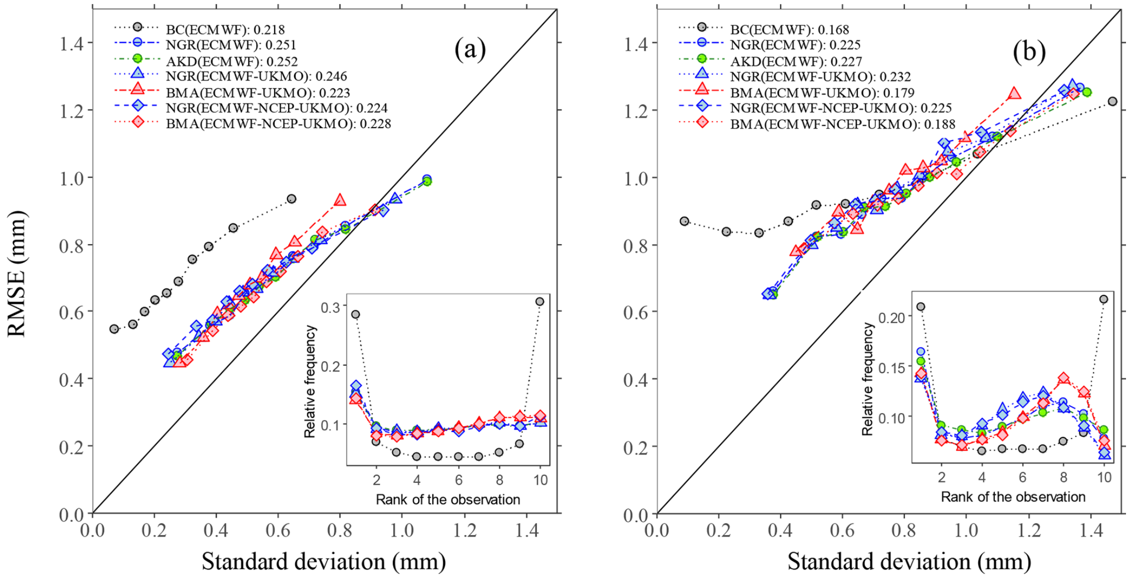

Figure 3. Binned spread–skill plots accounting for the mean of the ensemble standard deviation deciles against the mean RMSE of the forecasts in each decile over the verification period based on all pairs of forecasts and observations at lead times of **(a)** 1 d and **(b)** 7 d. The inset panels show the corresponding rank histograms. The correlation between the standard deviations and the absolute errors is included in the legend. The solid line represents the 1 : 1 relationship.

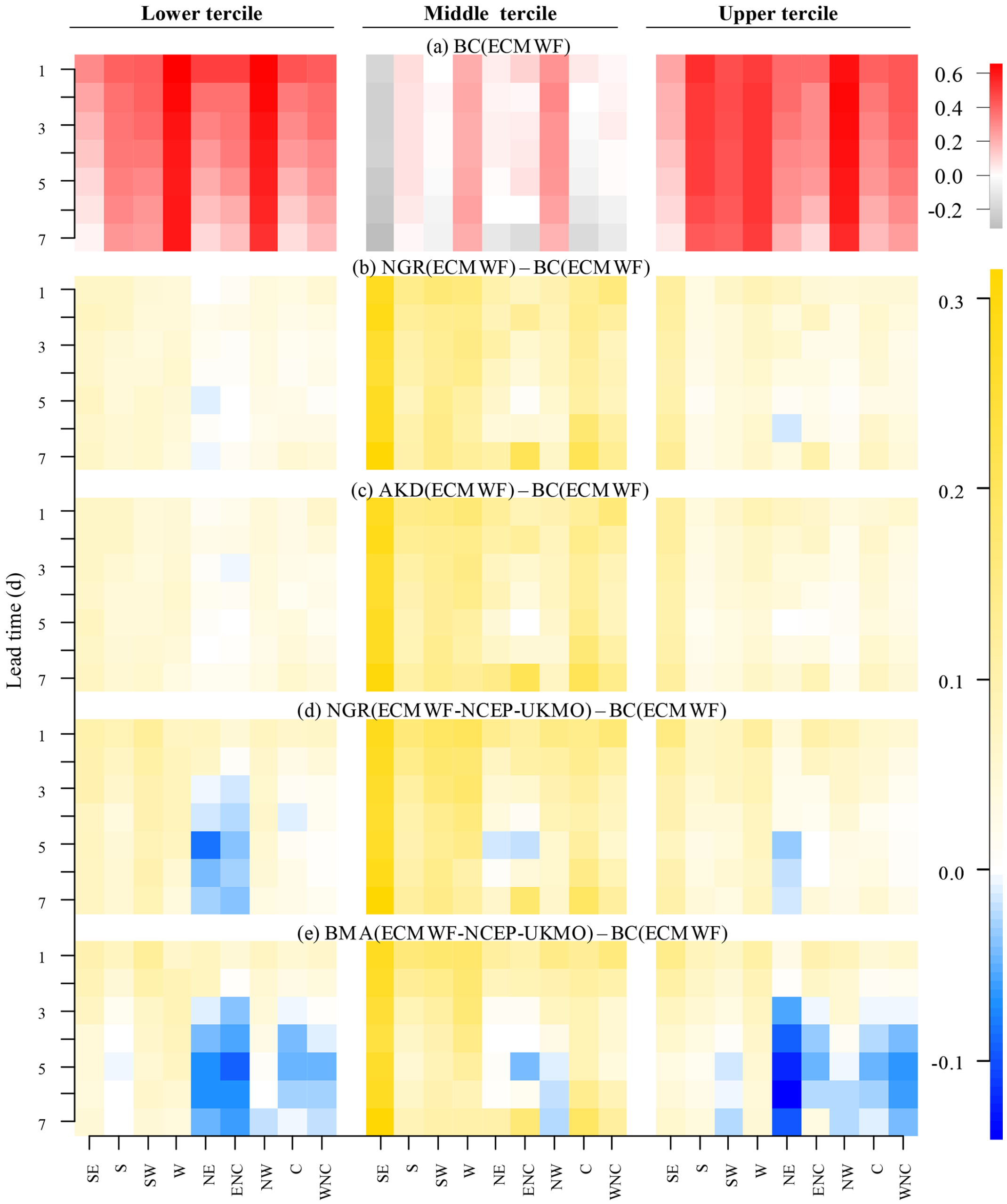

Figure 4. (a) The BSS for every region and lead time of the daily ECMWF forecasts post-processed using simple bias correction (utilized as reference BSS values); **(b–e)** differences between the BSS of the daily ECMWF forecasts post-processed using the **(b)** NGR and **(c)** AKD methods; and the daily ECMWF–NCEP–UKMO forecasts post-processed using the **(d)** NGR and **(e)** BMA methods and the reference BSS.

mance of different forecasts post-processed using the NGR and AKD techniques with training times of 45 and 30 d. The results suggest that the payoff from using 45 d is minimal. Table A2 in the Appendix shows the percentage differences in the forecasting performance when using training times of 45 and 30 d for post-processing. While there are generally some minor improvements when using 45 d over 30 d (which tend to be higher at longer lead times than at shorter times), these improvements usually represent less than 3 % of original statistics. The largest percentage difference, accounting for the BSS in the middle tercile, actually represented a negligible gain in absolute terms as they were affected by the close-to-zero range of the variable. The improvements were slightly higher for multi-model ensemble forecasts than for single-model forecasts. Notice that, while testing two different periods may be limited to evaluating the methods' sensitivity to the training period, the periods comprised a range for which methods such as the NGR and BMA have been reported to provide stable results (Gneiting et al., 2005; Raftery et al., 2005).

3.1.5 Weighting coefficients

The weighting coefficients reflect both the performance of the ensemble models and the performance of the post-processing techniques relative to their counterparts. Figure 5 shows the mean b_i (Eq. 1) weighting coefficients of the NGR technique and the w_i (Eq. 7) weighting coefficient of the BMA techniques for each region and lead time for the post-processed the post-processed ECMWF-NCEP-UKMO, respectively. The coefficients for the NGR and BMA techniques exhibited some common patterns of variability across regions and lead times. Both methods show that the weights of the ECMWF forecasts are highest overall, with a clear maximum at medium lead times. The weights of the UKMO model are the highest at 1 and 2 d but sharply decrease with lead time, whereas the weights of the NCEP model are generally the lowest, although they consistently increase with lead time (most likely due to the stronger decrease in performance with lead time by the other two models). This explains the most outstanding features of the performance assessments well in relation to the role of each model and the dependence among regions and lead times. Compared with the NGR method, the BMA method gives the UKMO forecasts a higher relative weight, although it is at the expense of the ECMWF forecast weights. For example, the weighting coefficients of the BMA method over the western regions are consistently higher for the UKMO forecasts than for the ECMWF forecasts. This suggests that the lower performance of the BMA post-processing relative to the NGR and AKD methods may be related to a misrepresentation of the model weights on the performance. This, in turn, may be caused by convergence problems during the parameter optimization with the expectation-maximization algorithm (Vrugt et al., 2008).

We observed considerable similarities in the distribution of the variance coefficients for the NGR method (Eq. 2) and the AKD (Eq. 6) method after post-processing the ECMWF forecasts. The two methods also provide very similar adjustments in the mean forecast because, unlike the BMA method, they independently bias correct the mean and optimize the spread–skill relationship (Bröcker and Smith, 2008). However, in the experiment, the NGR method was about 60 times faster than the AKD method. The BMA method was also faster than the AKD method, but it was still considerably slower than the NGR method. Considering the effectiveness of the NGR method, and its versatility with respect to post-processing both single- and multi-model ensemble forecasts, we applied this probabilistic technique to weekly ET_0 forecasts based on single- and multi-model ensembles.

Table 2. Spatially weighted average values of daily forecast metrics over all climate regions for different methods at lead times of 1 and 7 d. See the caption of Table 1 for explanations of the acronyms. The numbers in bold indicate the best performance for each lead time.

	BC ECMWF		NGR ECMWF		AKD ECMWF		NGR ECMWF–UKMO		BMA ECMWF–UKMO		NGR ECMWF–NCEP–UKMO		BMA ECMWF–NCEP–UKMO	
	1 d	7 d	1 d	7 d	1 d	7 d	1 d	7 d	1 d	7 d	1 d	7 d	1 d	7 d
rME (%)	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
ME (mm d^{-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 d^{-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**
CRPS (mm)	0.432	0.555	0.395	0.526	0.394	**0.525**	**0.374**	0.529	0.374	0.547	0.375	0.534	0.377	0.557
BSS_first	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_second	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_third	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

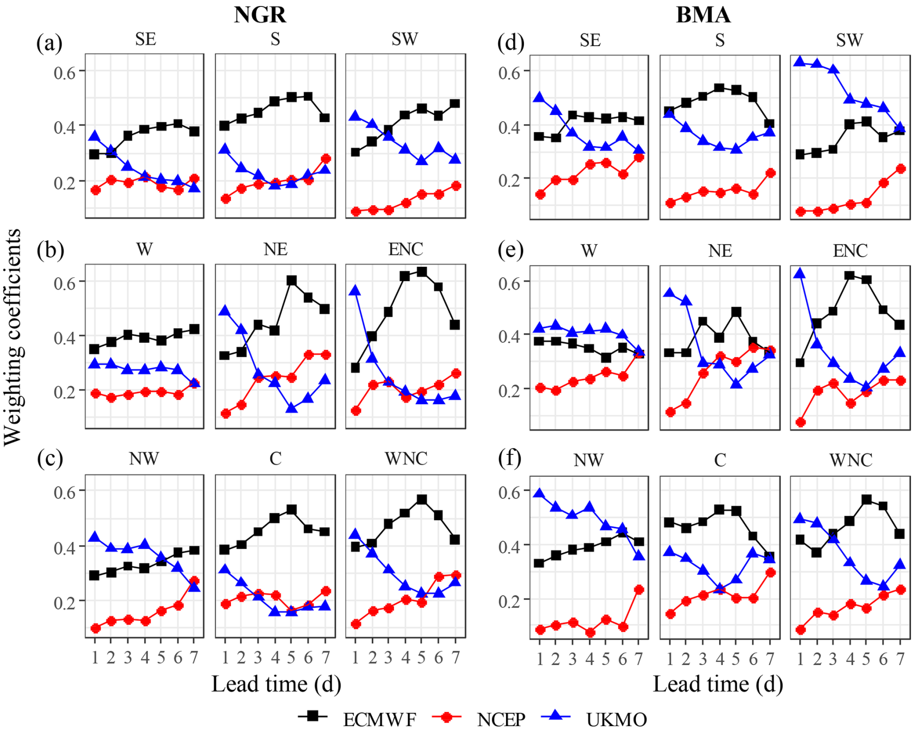

Figure 5. Regional mean weighting coefficient b of the NGR technique **(a–c)** and the weighting coefficient w of the BMA technique **(d–e)** for the post-processed daily ECMWF-NCEP-UKMO forecasts at different lead times.

3.2 Assessing the NGR method for post-processing weekly ET$_0$ forecasts

3.2.1 Deterministic forecast assessments

As for the daily predictions, the bias, the RMSE, and the correlation of the weekly forecasts post-processed with the NGR method and the linear regression methods were similar (Fig. 6). However, while the RMSE of daily forecasts based on ECMWF model varied between 12 % and 20 % of the total ET$_0$ (Fig. 2), the RMSE for any of weekly forecasting strategies commonly varied between 8 % and 11 %, which is lower than for daily forecasts; this made the latter more useful for operational purpose. The post-processed forecasts showed much lower RMSE values as well as correlation values that were twice as high as the predictions based on persistence, with the weekly predictions based on ECMWF forecasts generally being better, followed by the predictions based on the UKMO forecasts.

3.2.2 Probabilistic forecast assessments

Both the skill and the reliability of the weekly forecasts considerably improved when using NGR post-processing compared with bias correction post-processing (Table 3). The improvements were different among ET$_0$ forecast models. In most cases, the better the forecasts performance, the lower the improvements were. The adjustments in the coverage ratio and the Brier skill score were about 2.5 and 5 times larger for the UKMO and NCEP forecasts, respectively, than for the ECMWF forecasts. The bias-corrected ECMWF forecasts were generally better than both the UKMO and NCEP forecasts post-processed with the NGR method. We found that post-processing the NCEP forecasts with methods like NGR is almost mandatory in order to get reasonable probabilistic weekly forecasts of ET$_0$. For example, the coverage ratio of the bias-corrected forecasts in the west region was only 29 % due to the considerable under-dispersion. However, it is notable that, once the forecasts were post-processed using the NGR technique, they performed almost as well as the UKMO forecasts post-processed using the same method, increasing the coverage ratio to 98.4 %. Table 3 also shows that the multi-model ECMWF–UKMO weekly forecasts are commonly the best among those post-processed using the NGR method, followed by the ECMWF and the ECMWF–NCEP–UKMO forecasts.

The improvements in the reliability occurred due to substantial adjustments in both the ensemble spread and the spread–skill relationship of the raw forecasts (Fig. 7). The correlations between the standard deviation of the ensembles and the RMSEs were more than twice as high with NGR post-processing than with linear regression bias correction. These adjustments seemed even slightly more effective than adjustments resulting from the probabilistic post-processing of the daily forecasts (Fig. 3), although at the ex-

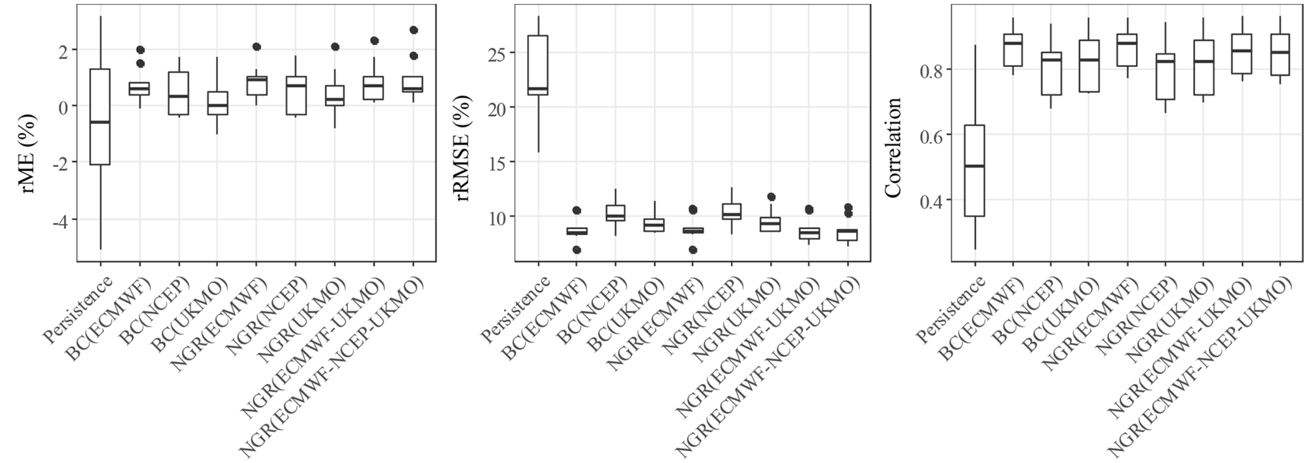

Figure 6. Whisker plot showing the 2.5th, 25th, 50th, 75th, and 97.5th percentiles of the distribution of the rME, rRMSE, and correlation values of weekly forecasts across different regions.

Table 3. Spatial weighted average values of weekly forecast metrics over all climate regions. See the caption of Table 1 for explanations of the acronyms. The best performance is shown in bold.

	Persistence	BC			NGR				
		ECMWF	NCEP	UKMO	ECMWF	NCEP	UKMO	ECMWF–UKMO	ECMWF–NCEP–UKMO
rME (%)	−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**
ME (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 ratio (%)		78.40	48.07	62.92	**99.29**	98.58	98.13	97.74	97.40
CRPS (mm)		1.836	2.406	2.072	1.727	2.071	1.884	**1.708**	1.715
BSS_first		0.508	0.326	0.448	0.529	0.430	0.501	**0.547**	0.506
BSS_second		0.164	−0.147	0.069	0.238	0.150	0.204	**0.255**	0.225
BSS_third		0.528	0.371	0.468	0.553	0.461	0.515	**0.558**	0.550

pense of a greater loss in sharpness. These contrasts in the post-processing effectiveness are probably associated with the differences in the training strategies.

In the case of the probabilistic forecast skill (Fig. 8), the improvements were larger for the middle tercile than for the other two terciles (similar to the daily forecasts). Unlike the bias-corrected forecasts, any of the probabilistically post-processed forecasts outperformed climatology for practically any tercile and in any region. Maybe more importantly, the Brier scores for the lower- and upper-tercile events of the forecasts that had been post-processed using the NGR method were over 30 % better than the scores of climatology in most cases. In the coastal regions, from the south to the northwest, the score was commonly over 50 % better, which was similar to the daily forecasts. Finally, the improvements resulting from the use of multi-model ensemble forecasts compared with single-model ensemble forecasts were generally small, except for in the southwest region.

4 Discussion

4.1 Effects of probabilistic post-processing on ET_0 forecasting performance

This study showed that NGR, AKD, and BMA post-processing schemes considerably improved the probabilistic forecast performance (coverage ratio, calibration, spread–skill, BSS, and CRPS) of the daily and weekly ET_0 forecasts compared with the simple (i.e., linear regression based on ensemble mean) bias correction method. While sharpness is a desired quality for any forecast, the daily and weekly bias-corrected ET_0 forecasts from NWP are spuriously sharp; this leads to poor consistency between the range of the ET_0 forecasts and the true values and ultimately undermines the confidence in those forecasts. The forecasts also exhibit a poor consistency in that the variance of the ensembles are commonly insensitive to the size of the forecast error. The probabilistic post-processed methods provided a much better reliability (with a coverage that was close to the nominal value) at a low cost with respect to sharpness. Therefore, these meth-

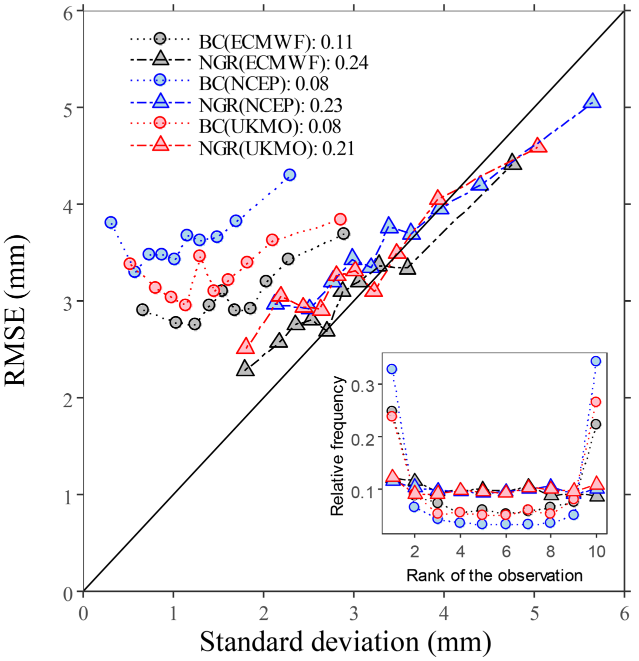

Figure 7. Binned spread–skill plots for the weekly forecasts accounting for the mean of the ensemble standard deviation deciles against the mean RMSEs of the forecasts in each decile over the verification period using all pairs of forecasts and observations. The inset panel shows the corresponding rank histograms. The correlation between the standard deviations and the absolute errors is included in the legend. The solid line represents the 1 : 1 relationship.

ods lead to a much better agreement between the forecasted probability of an ET_0 event occurring between certain thresholds and the proportion of times that the event occurs (see Gneiting et al., 2005).

In the case of the weekly ET_0 forecasts, the rate of improvement is considerably smaller for the ECMWF forecasts than for the UKMO and, especially, the NCEP forecasts. This seems to be largely due to the better performance of the ECMWF raw forecasts compared with the other forecasting systems. The probabilistic post-processing of the weekly NCEP forecasts seemed practically mandatory to produce reasonable predictions, but, once implemented, it provided performance assessments that were almost comparable to those based on the UKMO forecasts. These results have important implications for operational ET_0 forecasts that are based on the NCEP forecasts, such as the US National Digital Forecast Database (one of the few operational products of its type).

Unlike the probabilistic forecast metrics, the deterministic metrics (the ME, RMSE, and correlation of the ensemble mean) had a low sensitivity to the form (deterministic or probabilistic) of post-processing. In particular, the RMSE and correlation seemed more affected by the choice of the single- or multi-model ensemble forecast strategy than the choice between the NGR, AKD or simple bias correction post-processing method. However, the RMSE and correlation provided by the BMA method were consistently worse at long lead times. The daily errors using any post-processing method were relatively large, although mostly random, and, therefore, tended to cancel out at weekly scales. Thus, while the RMSE varied between 12 % and 20 % of the daily totals, it represented between 8 % and 11 % of the weekly totals. The RMSE for weekly ET_0 forecasts was more than 100 % lower than for the persistence-based ET_0 forecasts in all cases and was potentially more skillful than the forecasts that exploited the temporal persistence of the ET_0 time series (e.g., Landeras et al., 2009; Mohan and Arumugam, 2009).

4.2 Comparing the three probabilistic post-processing methods

The NGR- and AKD-based post-processing methods for the ECMWF forecasts produced comparable results, indicating that the simple Gaussian predictive distribution from the NGR method represents the uncertainty of the ET_0 predictions fairly well. The methods led to a similar distribution of the first two moments of the predictive probability function and similar performance statistics (with the AKD-based forecasts being only slightly better). However, the NGR method is more versatile as it can be applied to correct both single-model and multi-model ensemble forecasts, whereas the AKD method can only be applied to correct single-model forecasts. The NGR-based predictive distribution function is also easier to interpret than the AKD-based predictive distribution, which is given by an averaged sum of standard Gaussians.

The BMA method showed slightly less desirable performance compared with the NGR and AKD methods, which was presumably due to issues with the parameter identifiability. The implemented method uses the expectation-maximization (EM) algorithm to produce maximum likelihood estimates of the fitting coefficients; this algorithm is susceptible to converging to local minima, especially when dealing with multi-model ensemble forecasts with very different ensemble sizes (Vrugt et al., 2008). Archambeau et al. (2003) demonstrated that this algorithm also tends to identify local maximums of the likelihood of the parameters of a Gaussian mixture model in the presence of outliers or repeated values. Tian et al. (2012) found that adjusted BMA coefficients using both a limited-memory quasi-Newtonian algorithm and the Markov chain Monte Carlo were more accurate than those fitted with the EM algorithm; thus, this is a procedure that is worth testing in future studies.

4.3 Multi-model ensemble versus single-model ensemble forecasts

Daily multi-model ensemble forecasts performed better (in terms of the ME, RMSE, correlation, CRPS, and BSS) than daily ECMWF forecasts at short lead times (1–2 d) and over

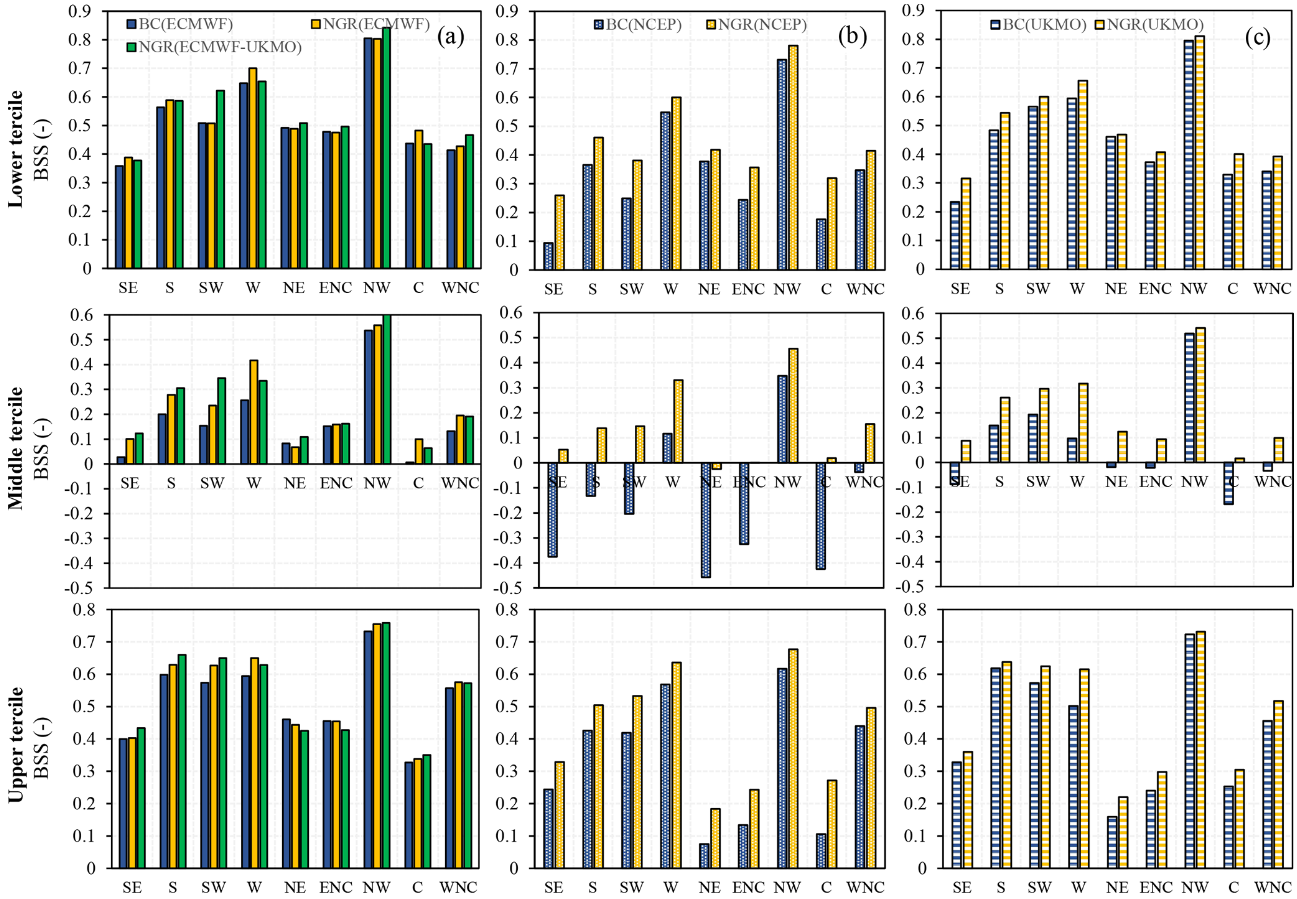

Figure 8. Comparison between BC- and NGR-based Brier skill scores considering **(a)** ECMWF and ECMWF–UKMO forecasts, **(b)** NCEP, and **(c)** UKMO forecasts across the different climate regions.

the western and southern regions, whereas the ECMWF forecasts are better over the northeastern regions for longer lead times. For other region–lead time combinations, the performance of single-model ensemble and multi-model ensemble forecasts did not differ much. We observed similar patterns for the raw and simple bias-corrected forecasts (Medina et al., 2018). However, the weekly multi-model ensemble forecast where only consistently better than the weekly single-model forecasts in the southwest region, seemingly because the weekly forecasts logically involve both short and long lead time assessments, and the effectiveness of the multi-models is degraded for long lead times. The observed behavior is associated with the performance of the ECMWF forecasts relative to the UKMO forecasts. While the ECMWF forecasts are generally better than the UKMO and NCEP forecasts, they are much better over the northeastern regions for medium lead times (4–6 d). In many cases, the UKMO forecasts are the best at lead times of 1–2 d, but they tend to be the worst at the longest times (6–7 d), especially over the abovementioned regions. The NCEP forecasts had a small contribution with respect to the ECMWF and UKMO forecasts at short lead times. These forecasts are comparatively better at longer lead times, but they still maintain a minor role regarding the ECMWF forecasts.

When considering daily forecasts, we adopted a 30 d training period length and showed that improvements were small (commonly lower than 3 %) when increasing the training period length to 45 d. This seems a plausible range for future works and represents an obvious advantage over methods such as the analog forecast, which provide similar performance (Tian and Martinez, 2012a, b, 2014) but require long training datasets. Gneiting et al. (2005) and Wilson (2007) found that lengths between 30 and 40 d provided good and almost constant performance assessments of sea level pressure forecasts post-processed using the NGR method and temperatures forecasts post-processed using the BMA method, respectively.

4.4 Post-processing the individual inputs versus post-processing ET_0

While we considered the post-processing of ET_0 ensembles produced with raw NWP forecasts in this study, it is possible that better predictions may be obtained by post-processing the forcing variables such as temperature, radiation and wind speed first, and then computing the ET_0. The NGR method has been shown to be successful for post-processing surface temperatures (e.g., Wilks and Hamill, 2007) that have a fairly Gaussian distribution. For example, Hagedorn (2008) and Hagedorn et al. (2008) showed gains in the lead time of between 2 and 4 d, with the gains being larger over areas where the raw forecast showed poor skill. Kann et al. (2009, 2011) used the NGR method to improve short-range ensemble forecasts of 2 m temperature. Recently, Scheuerer and Büermann (2014) provided a generalization of the original approach of Gneiting et al. (2005) that produces spatially calibrated probabilistic temperature forecasts. The wind speed forecasts have been commonly post-processed using of quantile regression method (e.g., Bremnes, 2004; Pinson et al., 2007; Møller et al., 2008). Even more recently, Sloughter et al. (2010) extended the original BMA method of Raftery et al. (2005) for wind speed by considering a gamma distribution for modeling the distribution of every member of the ensemble, which considerably improved the CRPS, the absolute errors, and the coverage. Vanvyve et al. (2015) and Zhang et al. (2015), in comparison, used the analog method following the methodology of Delle Monache (2013). Accurate solar radiation forecasting is particularly challenging because it requires a detailed representation of the cloud fields (Verzijlbergh et al., 2015), which are usually not resolved well by the NWP models. Davò et al. (2016) used artificial neural networks (ANN) and analog method approaches for the post-processing of both wind speed and solar radiation ensemble forecasts, which outperformed a simple bias correction approach. However, the post-processing of meteorological forecasts for producing ET_0 ensemble forecasts may require the consideration of the multivariate dependence among the forcing variables, which is often difficult (e.g., Wilks, 2015). Kang et al. (2010) found that post-processing of streamflow forecasts provided more accurate predictions than post-processing the forcing alone, whereas Vekade et al. (2013) showed that the improvements in precipitation and temperature via post-processing hardly benefited streamflow forecasts. Lewis et al. (2014) showed that the performance of the ET_0 forecasts can largely surpass the performance of individual input variables. Therefore, it is unclear if any benefit is obtained by using the post-processed inputs (instead of the raw forecasts) to construct ET_0 forecasts.

4.5 Future outlook

It is worth noting that, while the ET_0 forecasts are produced for use in agriculture, they have been tested over USCRN stations, which are not representative of agricultural settings. In real applications, the bias between the forecasts with no post-processing and the measurements based on agricultural stations could be higher than the bias resolved in this study. A question that should be addressed in the future studies is the extent to which the improvements in the predictive distribution of the ET_0 forecasts can be translated into a more reliable representation of the crop water use in agricultural lands and, ultimately, in water savings and economic gains. As ET_0 estimations can have remarkable impacts on soil moisture estimations (Rodriguez-Iturbe et al., 1999), we envision that new studies relying on a combination of rainfall and ET_0 forecasts post-processed using probabilistic methods will lead to considerable reductions on the uncertainty of soil moisture forecasts. New attempts should also investigate the role of the state-of-the-art probabilistic post-processing techniques on ET_0 forecasts produced from regional numerical weather prediction models, which have had improved spatial resolution and have already been used by different meteorological services (e.g., Baldauf et al., 2011; Seity et al., 2011; Hong and Dudhia, 2012; Bentzien and Friederichs, 2012).

5 Conclusions

To our knowledge, this study is the first work evaluating probabilistic methods based on NGR, AKD, and BMA techniques for post-processing daily and weekly ET_0 forecasts derived from single- or multi-model ensemble numerical weather predictions. The different ET_0 post-processing methods were compared against the simple linear regression bias correction method using both daily and weekly forecasts as well as against persistence in the case of weekly forecasts. The probabilistic post-processing techniques largely modified the spread of the original ET_0 forecasts, with very favorable impacts on the probabilistic forecast performance. They corrected the notable under-dispersion and the poor consistency between the spread of the ET_0 forecasts and the dimension of the errors, leading to a better BSS, reliability (both the coverage ratio and spread–skill relationship), and CRPS. The adjustments were crucial for the performance of the weekly NCEP forecasts and the weekly UKMO forecasts, whose bias-corrected versions showed a clear disadvantage compared with simply post-processed ECMWF forecasts.

The deterministic performance based on the NGR, AKD, and BMA methods was comparable to the performance based on the linear regression bias correction for both daily and weekly forecasts, and the skill was about 100 % higher than that based on persistence in the case of the weekly forecasts. The rRMSE values were between 12 % and 20 % for the daily totals and 8 % and 11 % for the weekly totals. The NGR and AKD methods provided similar estimates of the first- and second-order moments of the predictive density distribution; they showed similar effectiveness, but the NGR method had the advantage of being able to post-process both single- and multi-model ensemble forecasts. Both the NGR and AKD post-processing methods outperformed the BMA method when considering daily forecasts at long lead times.

Multi-model ensemble forecasting provided benefits at daily scales compared with the ECMWF ensemble forecasting, while the benefits were marginal at weekly scales.

The multi-model ensemble forecasting seems a better choice when the UKMO forecasts are comparable or slightly better than the ECMWF forecasts, such as at short lead times (1–2 d) and over the southern and western regions. Post-processing single-model forecasts is a better choice than post-processing multi-model ensemble forecasts in circumstances where the ECMWF forecasts perform considerably better than the UKMO and NCEP forecasts, such as at medium and long lead times, especially over the northeastern regions. While we considered a 30 d training period length for daily post-processing, the increase of the training period to 45 d only led to minimal improvements. In conclusion, our results suggest that the NGR post-processing of ET_0 forecasts generated from the ECMWF or ECMWF–UKMO predictions is the most plausible strategy among those evaluated, and it is recommended for operational implementations; this is due to the fact that the accuracy and reliability requirements for practical applications have not been discussed.

Appendix A

Table A1. 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 using the nonhomogeneous Gaussian regression (NGR) method. See the caption of Table 1 for an explanation of the forecast model 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
ME	−26.75	−30.83	−9.11	9.42	−13.91	−18.80	−4.27	25.05	−2.15	−1.45	−10.12	0.76
RMSE	−4.68	−4.01	−3.46	−2.51	−3.97	−2.84	1.90	4.33	1.46	2.00	−1.31	−0.92
Correlation	1.76	0.63	0.95	0.71	1.20	0.61	−4.18	−4.60	−3.28	−3.14	−2.31	−2.06
Coverage ratio	−1.39	−2.09	−0.98	−1.19	−1.02	−1.14	−0.84	−1.66	−0.85	−0.99	−0.84	−1.40
CRPS	−4.84	−3.89	−3.42	−1.99	−3.90	−2.81	1.41	4.02	1.58	2.45	−1.00	−0.27
BSS_first	12.02	7.48	3.22	2.85	3.55	4.24	−12.00	−9.68	−9.64	−9.38	−3.68	−5.18
BSS_second	8.99	−6.50	5.79	9.04	4.98	3.96	−112.95	−93.09	−19.09	−13.64	−15.73	−27.95
BSS_third	2.30	−1.81	3.58	6.56	4.20	2.37	−9.11	−8.99	−6.42	−10.61	−4.60	−5.84

Table A2. Percentage differences (averaged over regions) of forecast performance using a 45 d training period compared with using a 30 d training period for lead times of 1 and 7 d. See the caption of Table 1 for an explanation of the acronyms.

	NGR(ECMWF)		AKD(ECMWF)		NGR(ECMWF–UKMO)		NGR(ECMWF–NCEP–UKMO)	
	1 d	7 d	1 d	7 d	1 d	7 d	1 d	7 d
ME	16.57	18.73	21.65	22.86	4.71	10.09	−0.50	7.07
RMSE	−0.70	−2.64	−1.01	−3.12	−0.40	−3.72	−0.05	−4.74
Correlation	−0.16	0.53	−0.14	0.61	−0.10	1.33	−0.47	0.74
Coverage Ratio	1.28	0.95	1.62	1.26	1.70	1.50	1.94	1.34
CRPS (mm)	−0.77	−3.00	−1.22	−3.51	−0.92	−3.89	−0.01	−4.53
BSS_first	−0.88	2.18	−1.16	2.76	−0.21	5.06	−2.60	6.28
BSS_second	−1.26	2.76	−1.28	5.68	3.61	8.96	−2.29	5.56
BSS_third	−0.38	−1.59	−0.90	−0.21	−1.34	2.63	−1.63	0.24

Code and data availability. All of the data and the R codes used in this study are posted in a public data repository at https://doi.org/10.17605/OSF.IO/NG6WA (Medina and Tian, 2020).

Author contributions. HM and DT designed and conceptualized the research. HM implemented the design; performed data curation, analysis, validation, and visualization; and wrote the original draft of the paper. DT supervised the research, contributed advice, and reviewed and edited the paper.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgement. We are grateful to the National Centers for Environmental Information (NCEI), National Oceanic and Atmospheric Administration (NOAA), for hosting and providing public access to the US Climate Reference Network (USCRN) database. We are also grateful to the THORPEX Interactive Grand Global Ensemble (TIGGE) program, for providing public access to the ensemble forecast data from the NWP considered in this study, and to the ECMWF for hosting these datasets. The authors thank the reviewers for their very helpful comments.

Financial support. This research was supported by the Alabama Agricultural Experiment Station and the Hatch program of the National Institute of Food and Agriculture (NIFA), US Department of Agriculture (USDA; access no. 1012578); the Auburn University Intramural Grant Program; the Auburn University Presidential Awards for Interdisciplinary Research; and the USDA–NIFA Agriculture and Food Research Initiative (AFRI; grant no. G00012690).

Review statement. This paper was edited by Albrecht Weerts and reviewed by two anonymous referees.

References

Allen, R. G., Pereira, L. S., Raes, D., and Smith, M.: Crop evapotranspiration-Guidelines for computing crop water requirements-FAO, Irrigation and drainage paper 56, FAO, Rome, 300, p. D05109, 1998.

Archambeau, C., Lee, J. A., and Verleysen, M.: On Convergence Problems of the EM Algorithm for Finite Gaussian Mixtures, in: ESANN'2003 proceedings – European Symposium on Artificial Neural Networks, 23–25 April 2003, Bruges, Belgium, 99–106, ISBN 2-930307-03-X, 2003.

Baldauf, M., Seifert, A., Förstner, J., Majewski, D., Raschendorfer, M. and Reinhardt, T.: Operational convective-scale numerical weather prediction with the COSMO model: Description and sensitivities, Mon. Weather Rev., 139, 3887–3905, 2011.

Bauer, P., Thorpe, A., and Brunet, G.: The quiet revolution of numerical weather prediction, Nature, 525, 47–55, 2015.

Bentzien, S. and Friederichs, P.: Generating and calibrating probabilistic quantitative precipitation forecasts from the high-resolution NWP model COSMO-DE, Weather Forecast., 27, 988–1002, 2012.

Beran, R. and Hall, P.: Interpolated nonparametric prediction intervals and confidence intervals, J. Roy. Stat. Soc. B, 55, 643–652, 1993.

Bremnes, J. B.: Probabilistic Wind Power Forecasts Using Local Quantile Regression, Wind Energy, 7, 47–54, 2004.

Bröcker, J. and Smith, L. A.: From ensemble forecasts to predictive distribution functions, Tellus A, 60, 663–678, 2008.

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

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

Castro, F. X., Tudela, A., and Sebastià, M. T.: Modeling moisture content in shrubs to predict fire risk in Catalonia (Spain), Agr. Forest Meteorol., 116, 49–59, 2003.

Chirico, G. B., Pelosi, A., De Michele, C., Bolognesi, S. F., and D'Urso, G.: Forecasting potential evapotranspiration by combining numerical weather predictions and visible and near-infrared satellite images: an application in southern Italy, J. Agric. Sci., 156, 702–710, https://doi.org/10.1017/S0021859618000084, 2018.

Davò, F., Alessandrini, S., Sperati, S., Delle Monache, L., Airoldi, D., and Vespucci, M. T.: Post-processing techniques and principal component analysis for regional wind power and solar irradiance forecasting, Solar Energy, 134, 327–338, 2016.

Delle Monache, L., Eckel, F. A., Rife, D. L., Nagarajan, B., and Searight, K.: Probabilistic weather prediction with an analog ensemble, Mon. Weather Rev., 141, 3498–3516, 2013.

Fraley, C., Raftery, A. E., and Gneiting, T.: Calibrating multimodelmulti-model forecast ensembles with exchangeable and missing members using Bayesian model averaging, Mon. Weather Rev., 138, 190–202, 2010.

Fraley, C., Raftery, A. E., Sloughter, J. M., and Gneiting T.: EnsembleBMA: Probabilistic Forecasting using Ensembles and Bayesian Model Averaging, R package version 5.1.3, available at: https://CRAN.R-project.org/package=ensembleBMA (last access: 27 February 2020), 2016.

Glahn, H. R. and Lowry, D. A.: The use of model output statistics (MOS) in objective weather forecasting, J. Appl. Meteorol., 11, 1203–1211, 1972.

Glahn, H. R. and Ruth, D. P.: The new digital forecast database of the National Weather Service, B. Am. Meteorol. Soc., 84, 195–202, 2003.

Gneiting, T.: Calibration of medium-range weather forecasts, European Centre for Medium-Range Weather Forecasts, Technical Memorandum No. 719, Reading, UK, 30 pp., 2014.

Gneiting, T., Raftery, A. E., Westveld III, A. H., and Goldman, T.: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, Mon. Weather Rev., 133, 1098–1118, 2005.

Hagedorn, R.: Using the ECMWF reforecast data set to calibrate EPS forecasts, ECMWF Newslett., 117, 8–13, 2008.

Hagedorn, R., Hamill, T. M., and Whitaker, J. S.: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts.

Part I: Two-meter temperatures, Mon. Weather Rev., 136, 2608–2619, 2008.

Hagedorn, R., Buizza, R., Hamill, T. M., Leutbecher, M., and Palmer, T. N.: Comparing TIGGE multimodelmulti-model forecasts with reforecast-calibrated ECMWF ensemble forecasts, Q. J. Roy. Meteorol. Soc., 138, 1814–1827, 2012.

Hamill, T. M. and Colucci, S. J.: Verification of Eta–RSM short-range ensemble forecasts, Mon. Weather Rev., 125, 1312–1327, 1997.

Hamill, T. M. and Whitaker, J. S.: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application, Mon. Weather Rev., 134, 3209–3229, 2006.

Hamill, T. M., Bates, G. T., Whitaker, J. S., Murray, D. R., Fiorino, M., Galarneau Jr., T. J., Zhu, Y., and Lapenta, W.: Noaa's Second-Generation Global Medium-Range Ensemble Reforecast Dataset, B. Am. Meteorol. Soc., 94, 1553–1565, 2013.

Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems, Weather Forecast., 15, 559–570, 2000.

Hobbins, M., McEvoy, D., and Hain, C.: Evapotranspiration, evaporative demand, and drought, in: Drought and Water Crises: Science, Technology, and Management Issues, edited by: Wilhite, D. and Pulwarty, R., CRC Press, Boca Raton, USA, pp. 259–288, 2017.

Hong, S. Y. and Dudhia, J.: Next-generation numerical weather prediction: Bridging parameterization, explicit clouds, and large eddies, B. Am. Meteorol. Soc., 93, ES6–ES9, 2012.

Ishak, A. M., Bray, M., Remesan, R., and Han, D.: Estimating reference evapotranspiration using numerical weather modelling, Hydrol. Process., 24, 3490–3509, 2010.

Kang, T. H., Kim, Y. O., and Hong, I. P.: Comparison of pre- and post-processors for ensemble streamflow prediction, Atmos. Sci. Lett., 11, 153–159, 2010.

Kann, A., Wittmann, C., Wang, Y., and Ma, X.: Calibrating 2-m temperature of limited-area ensemble forecasts using high-resolution analysis, Mon. Weather Rev., 137, 3373–3387, 2009.

Kann, A., Haiden, T., and Wittmann, C.: Combining 2-m temperature nowcasting and short-range ensemble forecasting, Nonlinear Proc. Geoph., 18, 903–910, 2011.

Klein, W. H. and Glahn, H. R.: Forecasting local weather by means of model output statistics, B. Am. Meteorol. Soc., 55, 1217–1227, 1974.

Landeras, G., Ortiz-Barredo, A., and López, J. J.: Forecasting weekly evapotranspiration with ARIMA and artificial neural network models, J. Irrig. Drain. Eng., 135, 323–334, 2009.

Leutbecher, M. and Palmer, T. N.: Ensemble forecasting, J. Comput. Phys., 227, 3515–3539, 2008.

Madadgar, S., Moradkhani, H., and Garen, D.: Towards improved post-processing of hydrologic forecast ensembles, Hydrol. Process., 28, 104–122, 2014.

Mase, A. S. and Prokopy, L. S.: Unrealized potential: A review of perceptions and use of weather and climate information in agricultural decision making, Weather Clim. Soc., 6, 47–61, 2014.

Medina, H. and Tian, D.: Post-processed reference crop evapotranspiration forecasts, https://doi.org/10.17605/OSF.IO/NG6WA, 2020.

Medina, H., Tian, D., Srivastava, P., Pelosi, A., and Chirico, G. B.: Medium-range reference evapotranspiration forecasts for the contiguous United States based on multimodelmulti-model numerical weather predictions, J. Hydrol., 562, 502–517, 2018.

Medina, H., Tian, D., Marin, F. R., and Chirico, G. B.: Comparing GEFS, ECMWF, and Postprocessing Methods for Ensemble Precipitation Forecasts over Brazil, J. Hydrometeorol., 20, 773–790, 2019.

Messner, J. W., Mayr, G. J., Zeileis, A., and Wilks, D. S.: Heteroscedastic Extended Logistic Regression for Postprocessing of Ensemble Guidance, Mon. Weather Rev., 142, 448–456, https://doi.org/10.1175/MWR-D-13-00271.1, 2014.

Mohan, S. and Arumugam, N.: Forecasting weekly reference crop evapotranspiration series, Hydrol. Sci. J., 40, 689–702, 1995.

Møller, J. K., Nielsen, H. A., and Madsen, H.: Time-Adaptive Quantile Regression, Comput. Stat. Data Anal., 52, 1292–1303, 2008.

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

Pelosi, A., Medina, H., Villani, P., D'Urso, G., and Chirico, G. B.: Probabilistic forecasting of reference evapotranspiration with a limited area ensemble prediction system, Agr. Water Manage., 178, 106–118, 2016.

Pelosi, A., Medina, H., Van den Bergh, J., Vannitsem, S., and Chirico, G. B.: Adaptive Kalman filtering for post-processing ensemble numerical weather predictions, Mon. Weather Rev., 145, 4837–4854, https://doi.org/10.1175/MWR-D-17-0084.1, 2017.

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

Pinson, P. and Madsen, H.: Ensemble-Based Probabilistic Forecasting at Horns Rev, Wind Energy, 12, 137–155, 2009.

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., and Perron, J.: Agricultural advisors: a receptive audience for weather and climate information?, Weather Clim. Soc., 5, 162–167, 2013.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using Bayesian model averaging to calibrate forecast ensembles, Mon. Weather Rev., 133, 1155–1174, 2005.

R Core Team: R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, available at: http://www.R-project.org/ (last access: 27 February 2020), 2014.

Rodriguez-Iturbe, I., Porporato, A., Ridolfi, L., Isham, V., and Coxi, D. R.: Probabilistic modelling of water balance at a point: the role of climate, soil and vegetation, P. Roy. Soc. Lond. A, 455, 3789–3805, 1999.

Roulston, M. S. and Smith, L. A.: Combining dynamical and statistical ensembles, Tellus A, 55, 16–30, 2003.

Scheuerer, M. and Büermann, L.: Spatially adaptive post-processing of ensemble forecasts for temperature, J. Roy. Stat. Soc. C, 63, 405–422, 2014.

Seity, Y., Brousseau, P., Malardel, S., Hello, G., Bénard, P., Bouttier, F., Lac, C., and Masson, V.: The AROME-France convective-scale operational model, Mon. Weather Rev., 139, 976–991, 2011.

Siegert, S.: SpecsVerification: Forecast Verification Routines for Ensemble Forecasts of Weather and Climate, R package version 0.5-2, available at: https://cran.r-project.org/web/packages/SpecsVerification/ (last access: 27 February 2020), 2017.

Silva, D., Meza, F. J., and Varas, E.: Estimating reference evapotranspiration (ET_0) using numerical weather forecast data in central Chile, J. Hydrol., 382, 64–71, 2010.

Sloughter, J. M., Gneiting, T., and Raftery, A. E.: Probabilistic wind speed forecasting using ensembles and Bayesian model averaging, J. Am. Stat. Assoc., 105, 25–35, 2010.

Swinbank, R., Kyouda, M., Buchanan, P., Froude, L., Hamill, T. M., Hewson, T. D., Keller, J. H., Matsueda, M., Methven, J., Pappenberger, F., and Scheuerer, M.: The Tigge Project and Its Achievements, B. Am. Meteorol. Soc., 97, 49–67, 2016.

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

Tian, D. and Martinez, C. J.: Forecasting Reference Evapotranspiration Using Retrospective Forecast Analogs in the Southeastern United States, J. Hydrometeorol., 13, 1874–1892, 2012b.

Tian, D. and Martinez, C. J.: The GEFS-based daily reference evapotranspiration (ET_0) forecast and its implication for water management in the southeastern United States, J. Hydrometeorol., 15, 1152–1165, 2014.

Tian, X., Xie, Z., Wang, A., and Yang, X.: A new approach for Bayesian model averaging, Sci. China Earth Sci., 55, 1336–1344, 2012.

Toth, Z., Talagrand, O., Candille, G., and Zhu, Y.: Probability and ensemble forecasts, Forecast Verification: A Practitioner's Guide in Atmospheric Science, John Wiley & Sons Ltd., England, 137–163, 2003.

van Osnabrugge, B., Uijlenhoet, R., and Weerts, A.: Contribution of potential evaporation forecasts to 10-day streamflow forecast skill for the Rhine River, Hydrol. Earth Syst. Sci., 23, 1453–1467, https://doi.org/10.5194/hess-23-1453-2019, 2019.

Vanvyve, E., Delle Monache, L., Monaghan, A. J., and Pinto, J. O.: Wind resource estimates with an analog ensemble approach, Renew. Energ., 74, 761–773, 2015.

Verkade, J. S., Brown, J. D., Reggiani, P., and Weerts, A. H.: Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales, J. Hydrol., 501, 73–91, https://doi.org/10.1016/j.jhydrol.2013.07.039, 2013.

Verzijlbergh, R. A., Heijnen, P. W., de Roode, S. R., Los, A., and Jonker, H. J.: Improved model output statistics of numerical weather prediction based irradiance forecasts for solar power applications, Solar Energy, 118, 634–645, 2015.

Vrugt, J. A., Diks, C. G., and Clark, M. P.: Ensemble Bayesian model averaging using Markov chain Monte Carlo sampling, Environ. Fluid Mech., 8, 579–595, 2008.

Wang, X. and Bishop, C. H.: Improvement of ensemble reliability with a new dressing kernel, Q. J. Roy. Meteorol. Soc., 131, 965–986, 2005.

Whan, K. and Schmeits, M: Comparing Area Probability Forecasts of (Extreme) Local Precipitation Using Parametric and Machine Learning Statistical Postprocessing Methods, Mon. Weather Rev., 146, 3651–3673, https://doi.org/10.1175/MWR-D-17-0290.1, 2018.

Wilks, D. S.: Comparison of ensemble-MOS methods in the Lorenz'96 setting, Meteorol. Appl., 13, 243–256, 2006.

Wilks, D. S.: Extending logistic regression to provide full probability distribution MOS forecasts, Meteorol. Appl., 16, 361–368, 2009.

Wilks, D. S.: Sampling distributions of the Brier score and Brier skill score under serial dependence, Q. J. Roy. Meteor. Soc., 136, 2109–2118, 2010.

Wilks, D. S.: Multivariate ensemble Model Output Statistics using empirical copulas, Q. J. Roy. Meteor. Soc., 141, 945–952, 2015.

Wilks, D. S. and Hamill, T. M.: Comparison of ensemble-MOS methods using GFS reforecasts, Mon. Weather Rev., 135, 2379–2390, 2007.

Williams, R. M., Ferro, C. A. T., and Kwasniok, F.: A comparison of ensemble post-processing methods for extreme events, Q. J. Roy. Meteor. Soc., 140, 1112–1120, 2014.

Wilson, L. J., Beauregard, S., Raftery, A. E., and Verret, R.: Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian model averaging, Mon. Weather Rev., 135, 1364–1385, 2007.

Yuen, R., Baran, S., Fraley, C., Gneiting, T., Lerch, S., Scheuerer, M., and Thorarinsdottir, T.: ensembleMOS: Ensemble Model Output Statistics, R package version 0.8.2, available at: https://CRAN.R-project.org/package=ensembleMOS (last access: 27 February 2020) 2018.

Zhang, J., Draxl, C., Hopson, T., Delle Monache, L., Vanvyve, E., and Hodge, B. M.: Comparison of numerical weather prediction based deterministic and probabilistic wind resource assessment methods, Appl. Energy, 156, 528–541, 2015.

Zhao, T., Wang, Q. J., and Schepen, A.: A Bayesian modelling approach to forecasting short-term reference crop evapotranspiration from GCM outputs, Agr. Forest Meteorol., 269, 88–101, 2019.