Parametric post-processing of dual-resolution precipitation forecasts

Marianna Szabó,[a,b] Estíbaliz Gascón[c] and Sándor Baran,[a]

[a] *Faculty of Informatics, University of Debrecen, Hungary*

[b] *Doctoral School of Informatics, University of Debrecen, Hungary*

[c] *European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom*

Corresponding author: Sándor Baran, baran.sandor@inf.unideb.hu

ABSTRACT: Recently, all major weather centres issue ensemble forecasts, which differ both in ensemble size and spatial resolution – even while covering the same domain. These parameters directly determine both the forecast skill of the prediction and the computation cost. In the last few years, the plans of upgrading the configuration of the Integrated Forecast System of the European Centre for Medium-Range Weather Forecasts (ECMWF) from a single forecast with 9 km resolution and a 51-member ensemble with 18 km resolution induced an extensive study of the forecast skill of both raw and post-processed dual-resolution predictions comprising ensemble members of different horizontal resolutions.

We investigate the predictive performance of the censored shifted gamma (CSG) ensemble model output statistic (EMOS) approach for statistical post-processing with the help of dual-resolution 24h precipitation accumulation ensemble forecasts over Europe with various forecast horizons. We consider the operational 50-member ECMWF ensemble as high-resolution and extend it with a low-resolution (29-km grid) 200-member experimental forecast. The investigated dual-resolution combinations consist of subsets of these two forecast ensembles with equal computational cost, which is equivalent to the cost of the operational ensemble.

Our case study verifies that, compared with the raw ensemble combinations, EMOS post-processing results in a significant improvement in forecast skill and that skill is statistically indistinguishable between any of the analysed mixtures of dual-resolution combinations. Furthermore, the semi-locally trained CSG EMOS provides an efficient alternative to the state-of-the-art quantile mapping without requiring additional historical data.

1. Introduction

Capturing and modelling uncertainty is an essential need in any forecasting problem, and in weather prediction, it may result in enormous economic benefit. In the early '90s there was an important shift in weather forecasting practice from deterministic forecasts obtained using numerical weather prediction (NWP) models in the direction of probabilistic forecasting. The crucial step was the introduction of ensemble prediction systems (EPSs) into operational use in 1992 both at the European Centre for Medium-Range Weather Forecasts (ECMWF) and the U.S. National Meteorological Center. An EPS provides a range of forecasts corresponding to different runs of the NWP models, which are usually generated from random perturbations in the initial conditions and/or the stochastic physics parametrization. In the last decades, the ensemble method has become a widely used technique all over the world, opening the door for probabilistic forecasting (Gneiting and Raftery 2005).

Obviously, the larger the ensemble size, the higher the chance of correctly estimating forecast uncertainty (Machete and Smith 2016; Leutbecher 2018), while high spatial resolution is essential for detecting local phenomena. However, weather centres have a fixed amount of computational resources. As computational costs increase both with ensemble size and resolution, before introducing a new operational EPS configuration, a reasonable tradeoff on these key parameters should be made. At the moment the operational global 51-member medium-range (up to 15 days) ECMWF ensemble is generated at TCo639 resolution meaning an approximately 18 km grid spacing (Haiden et al. 2018), whereas for 2023 the plan is to reach the 9 km (TCo1279) resolution of the current deterministic forecast. The new configuration to be implemented in 2023 will comprise a 51-member medium-range TCo1279 forecast run twice a day and a 101-member extended-range (up to 46 days) forecast on a 36 km grid (TCo319) run only once daily. Though the change in configuration was driven by the separate benefits in skill for the medium-range ensemble originating from the resolution increase and the benefit in skill for the extended-range forecasts from the increase in ensemble size, this new setup will allow operational dual-resolution ensemble forecasts. Further, according to the ECMWF Strategy for 2021–2030[1], "ECMWF will continue to investigate a mixture of larger ensemble and increased vertical and horizontal resolution, and a blend of variational and ensemble methods across the Earth system components". In line with the

[1] https://www.ecmwf.int/sites/default/files/elibrary/2021/ecmwf-strategy-2021-2030-en.pdf [Accessed on 2 April 2023]

above goals, Leutbecher et al. (2020) investigated the forecast skill of dual-resolution ensembles by considering combinations of high- (TCo639) and low-resolution (TCo399, grid resolution $\approx$ 29 km; TCo255, grid resolution $\approx$ 45 km) temperature predictions for a given computation cost. The authors found that in the case of 2 m temperature, provided the ensemble size is large enough, combinations with roughly equal number of high- and low-resolution members exhibit the best predictive performance.

However, ensemble forecasts are often underdispersive, that is, the spread of the ensemble is too small to account for the full uncertainty, and may also be subject to systematic bias. This phenomenon has been observed with several operational EPSs (see e.g. Buizza et al. 2005) and can be resolved by some form of statistical post-processing (Buizza 2018). Over the last decades, various post-processing methods have been proposed for a large variety of weather variables; for a detailed overview of the most advanced methods see e.g. Wilks (2018) or Vannitsem and Coauthors (2021). Parametric approaches such as Bayesian model averaging (Raftery et al. 2005) or ensemble model output statistics (EMOS: Gneiting et al. 2005) provide full predictive distributions of future weather quantities. In contrast, non-parametric methods usually capture predictive distributions via estimating their quantiles (see e.g. Friederichs and Hense 2007; Bremnes 2019), whereas member-by-member post-processing uses e.g. linear regression to improve the raw ensemble (Van Schaeybroeck and Vannitsem 2015). Recently, approaches using artificial neural networks also become more and more popular as they provide more flexibility in modelling both in parametric (see e.g. Rasp and Lerch 2018; Scheuerer et al. 2020; Ghazvinian et al. 2022) and non-parametric contexts (Bremnes 2020).

Concerning dual-resolution ensemble forecasts, Baran et al. (2019) investigated whether statistical post-processing changed the conclusions regarding the optimal dual-resolution configuration found by Leutbecher et al. (2020). The authors discovered that EMOS calibration strongly reduced the differences in skill among the equal-cost configurations of single- and dual-resolution 2m temperature ensembles and the ranking of the different configurations could also change. In a parallel study, Gascón et al. (2019) studied the post-processing of ECMWF dual-resolution (TCo639 and TCo399) precipitation ensemble forecasts using a quantile mapping and objective weighting of sorted ensemble members approach (Hamill and Scheuerer 2018) and confirmed the superiority of the combination with an equal number of high- and low-resolution ensemble members. However,

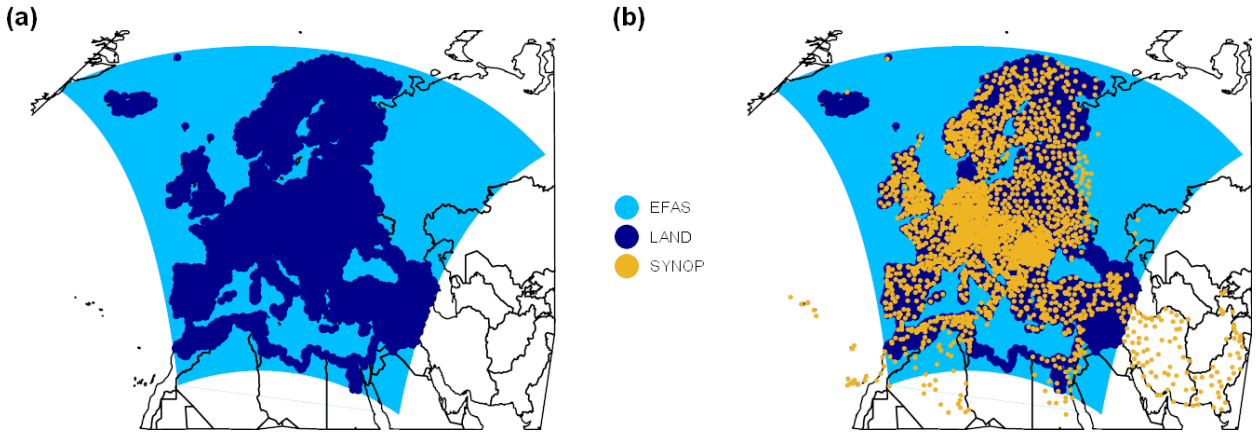


FIG. 1. Map of the (a) domain of the EFAS gridded data and the land subset; (b) SYNOP stations in the land subset of the EFAS gridded data.

the considered calibration method requires an extended set of historical data; in Gascón et al. (2019) forecast-analysis pairs for 20 years were utilized. In this paper, we propose the essentially simpler censored and shifted gamma (CSG) distribution-based EMOS approach of Baran and Nemoda (2016) to calibrate ECMWF dual-resolution forecasts and compare the results with the findings of Gascón et al. (2019). The advantage of this parametric model is that it requires no additional historical data and results in full predictive distributions.

The paper is organized as follows. Section 2 contains a detailed description of the precipitation accumulation dataset used in the study. In Section 3 the applied calibration approaches are reviewed and the parameter estimation methods and model verification tools are provided. The results are provided in Section 4, followed by a concluding Section 5.

2. Data

As established in the Introduction, the datasets used in this study are identical to the ones considered by Gascón et al. (2019). The weather variable at hand is 24h precipitation accumulation (from 0600 UTC to the same time the next day) and the dual-resolution ensemble consists of different combinations of forecasts of the ECMWF Integrated Forecast System (IFS) at resolutions TCo639 (high-resolution) and TCo399 (low-resolution). Note that the cost ratio between these two resolutions is 4:1, so in the different dual-resolution configurations four TCo399 members can be traded against a single TCo639 forecast. Both high- and low-resolution ensemble forecasts were

produced using the IFS model cycle 41r2, which was the operational version of the model during the verification period. The initial conditions and the stochastic representation of model uncertainties were identical for both ensemble resolutions. Further details can be found in Leutbecher (2018).

The first dataset consists of 24h gridded accumulated precipitation analyses of the European Flood Awareness System (EFAS: Ntegeka et al. 2013) for 1996 – 2016 covering Europe and some of the surrounding countries (see Figure 1a). Data from 2016 serve as validation data for the investigated post-processing approaches, whereas analyses from the preceding years (1996 – 2015) are required for training the quantile mapping-based methods. Note that whether model training is performed using all EFAS grid points corresponding to the land subset (5 km grid spacing, 363534 grid points), for validation purposes only data of 2370 grid points corresponding to SYNOP stations are considered (Figure 1b).

Post-processing is applied to 24h precipitation accumulation forecasts of the ECMWF IFS for June-July-August (JJA) 2016 with forecast horizons of up to 10 days. All forecasts are initialized at 0000 UTC and in order to match the accumulation period of the EFAS analyses, lead times of 6h, 30h, ..., 246h are considered. Similar to Gascón et al. (2019); Baran et al. (2019) and Leutbecher et al. (2020), 50 perturbed members of the operational TCo639 ensemble and forecasts from the 200-member TCo399 experiment are analyzed. The investigated dual-resolution mixtures (M_H, M_L) of M_H high-resolution and M_L low-resolution members are

$$(50,0),\ (40,40),\ (20,120),\ (10,160),\ (0,200),$$

all having the same computational cost corresponding to the available HPC resources of the ECMWF at the moment the forecasts were generated.

Finally, quantile mapping-based approaches are trained with the help of 11-member gridded reforecasts for JJA of the period 1996 – 2016 with forecast horizons matching the lead times of the dual-resolution combinations generated both at TCo639 and TCo399 resolutions. Note that both the dual-resolution forecasts and the reforecast data were interpolated to the EFAS 5 km grid using a nearest-neighbour technique. For a more detailed description of the above datasets, please refer to Gascón et al. (2019) and the references therein.

3. Statistical post-processing

In terms of computational costs and model complexity, EMOS is one of the most efficient post-processing approaches (see e.g. Vannitsem and Coauthors 2021, Fig. 1) showing excellent performance for a large variety of weather quantities. It fits a single parametric law to the ensemble forecast with parameters depending on the ensemble members. EMOS models for different weather variables usually differ in the applied parametric law and/or in the link functions connecting the ensemble forecasts to the parameters of the predictive distribution.

a. Censored and shifted gamma EMOS model

In contrast to continuous weather quantities, such as temperature or wind speed that can be modelled with a normal (Gneiting et al. 2005) or a truncated normal distribution (Thorarinsdottir and Gneiting 2010), respectively, precipitation accumulation has a special discrete-continuous nature. Hence, one can only consider non-negative predictive distributions that assign positive mass to the zero precipitation event. A popular choice is to consider a continuous distribution that can take both positive and negative values and censor it at zero from the left (Scheuerer 2014; Scheuerer and Hamill 2015), which approach is also the core idea of the CSG distribution-based EMOS model suggested by Baran and Nemoda (2016).

Let $G_{\kappa,\theta}$ denote the cumulative distribution function (CDF) of a gamma distribution $\Gamma(\kappa,\theta)$ with shape $\kappa > 0$ and scale $\theta > 0$ defined by probability density function (PDF)

$$g_{\kappa,\theta}(x) := \begin{cases} \frac{x^{\kappa-1}\mathrm{e}^{-x/\theta}}{\theta^{\kappa}\Gamma(\kappa)}, & x > 0, \\ 0, & \text{otherwise}, \end{cases}$$

where $\Gamma(\kappa)$ stands for the value of the gamma function at κ. Since there is a one-to-one correspondence between the parameters κ and θ and the mean $\mu > 0$ and standard deviation $\sigma > 0$ of the corresponding gamma distribution via equations

$$\kappa = \frac{\mu^2}{\sigma^2} \qquad \text{and} \qquad \theta = \frac{\sigma^2}{\mu},$$

$\Gamma(\kappa,\theta)$ can be characterized by these two moments as well. After extending the support of the gamma distribution to negative values with the help of a shift parameter $\delta > 0$, one can introduce

a shifted gamma distribution, left censored at zero $\Gamma^0(\kappa,\theta,\delta)$ with shape κ, scale θ and shift δ by CDF

$$G^0_{\kappa,\theta,\delta}(x) := \begin{cases} G_{\kappa,\theta}(x+\delta), & x \geq 0, \\ 0, & x < 0. \end{cases}$$

In what follows, let $f_1, f_2, \ldots, f_K$ denote the precipitation accumulation ensemble forecast for a given location, time and lead time, and denote by $\overline{f}$ the ensemble mean. In the CSG EMOS model mean μ and variance σ^2 of the underlying gamma distribution are linked to the ensemble members as

$$\mu = a^2 + b_1^2 f_1 + b_2^2 f_2 + \cdots + b_K^2 f_K \qquad \text{and} \qquad \sigma^2 = c^2 + d^2 \overline{f}, \tag{1}$$

whereas shift parameter $\delta > 0$ is independent of the ensemble forecast. Here a, $b_1, \ldots, b_K$, c, d and δ are unknown parameters to be estimated.

However, the 50 perturbed members of the operational ECMWF IFS are considered statistically indistinguishable and, in this way, exchangeable. The same applies to the investigated 200-member low-resolution ensemble as well. In this way the investigated dual-resolution forecasts consist of groups of exchangeable ensemble members, which should be taken into account in the modelling process. In the following, if the M-member ensemble is divided into K groups of exchangeable forecasts, where the kth group contains $M_k \geq 1$ ensemble members ($\sum_{k=1}^{M} M_k = M$), the notation $\overline{f}_k$ will be used for the corresponding group mean. In this situation, ensemble members within a given exchangeable group should share the same parameters (Wilks 2018) and link functions in (1) should be replaced by

$$\mu = a^2 + b_1^2 \overline{f}_1 + b_2^2 \overline{f}_2 + \cdots + b_K^2 \overline{f}_K \qquad \text{and} \qquad \sigma^2 = c^2 + d^2 \overline{f}. \tag{2}$$

b. Parameter estimation

Following the optimum score estimation principle of Gneiting and Raftery (2007), mean parameters $a, b_1, \ldots, b_K$, variance parameters c, d, and shift parameter $\delta > 0$ of the CSG EMOS model specified either by (1) or by (2) can be estimated by optimizing the mean value of a proper scoring rule over an appropriate set of training data. Scoring rules measure predictive performance by assigning numerical values to forecast-observation pairs and in atmospheric sciences, one of

the most popular choices is the continuous ranked probability score (CRPS: Wilks 2019, Section 9.5.1), as it assesses simultaneously both calibration and sharpness of the probabilistic forecast. For a prediction represented by a CDF $F(y)$ and a real value x, the CRPS is defined as

$$\mathrm{CRPS}(F,x) \coloneqq \int_{-\infty}^{\infty} \left[F(y) - \mathbb{I}_{\{y \geq x\}}\right]^2 \mathrm{d}y, \tag{3}$$

with $\mathbb{I}_H$ denoting the indicator function of a set H. Note that CRPS can be reported in the same units as the observation and it is a negatively oriented score where smaller values mean better forecast skill. Further, for the CSG distribution, CRPS can be expressed in a closed form (Scheuerer and Hamill 2015) allowing for efficient optimization in the parameter estimation process.

The next step in EMOS modelling is the appropriate choice of training data. The most common approach is the use of rolling training periods where ensemble forecasts and corresponding validation observations from the preceding n calendar days are considered. The spatial selection of the training data is also an important issue, where the traditional approaches are local and global (regional) modelling (Thorarinsdottir and Gneiting 2010). In the former case, EMOS model parameters for a given location are estimated using training data of that particular spot. In the latter, all available forecast cases in the training data are considered resulting in a single set of parameters for the whole ensemble domain. In general, local EMOS models outperform their regional counterparts; however, in order to avoid numerical problems in parameter estimation, they require rather long training periods (Hemri et al. 2014). In contrast, regional modelling can be performed with much shorter training periods (see e.g. Baran and Baran 2021), though it is usually unsuitable for large and heterogeneous ensemble domains like the one at hand (Figure 1a). The advantages of the above selection methods can be combined using semi-local approaches, where one either augments the training data for a given location with data of locations with similar characteristics or splits the ensemble domain into more homogeneous subdomains and within each subdomain performs a regional modelling (Hamill et al. 2008; Lerch and Baran 2017). The data augmentation technique of Hamill and Scheuerer (2018) is applied in the reference quantile mapping approaches described in Section 3c, whereas in EMOS modelling we consider the clustering-based semi-local method of Lerch and Baran (2017). For a given date of the verification period, to each EFAS grid point of the land subset, one assigns a feature vector depending both on the grid point climatology and the forecast errors of the raw ensemble during the training period. Then based on the corre-

sponding feature vectors the grid points are grouped into clusters using k-means clustering. When the training period rolls ahead, grid points are regrouped dynamically.

c. Quantile mapping

In the case study of Section 4, the reference post-processing approach is quantile mapping and its weighted version investigated by Gascón et al. (2019). The main idea behind the quantile mapping is that using climatological CDFs $F_f(y)$ and $F_o(y)$ of forecasts and observations, respectively, a raw forecast f is adjusted to match the distribution of the observation x. The adjusted forecast $\widetilde{f}$ is given by

$$\widetilde{f} \coloneqq F_o^{-1}\big(F_f(f)\big),$$

and for a forecast ensemble the adjustment should be made separately for each member.

In our case $F_f(y)$ and $F_o(y)$ are estimated from historical data of calendar years 1996 – 2015 using the control member of the 11-member ECMWF reforecasts and the EFAS analysis, respectively. For each grid point and each date of the verification period (JJA 2016) climatological CDFs are developed from 9000 sample values corresponding to the 9 closest Julian dates to the given date from the whole 20-year period for 50 supplemental similar locations chosen according to suggestions by Hamill and Scheuerer (2018).

For the weighted version of quantile mapping first each member of the ECMWF reforecast for 1996 – 2015 has to be adjusted separately. For a given calendar year, climatological CDFs are calculated in the same way as before utilizing the matching reforecasts and corresponding analyses of 9 neighbouring dates from the remaining 19 years for 50 similar supplemental locations (for the details see Gascón et al. 2019, Section 2.4.1).

The 11-member quantile-mapped reforecasts for the period from 1996 to 2015 are then applied to derive the 11-bin closest-member histograms. For a given forecast horizon, date and grid point, the 11-member adjusted reforecasts are sorted and compared with the corresponding analysed precipitation amount. Then for various quantile-mapped ensemble mean values representing e.g. light (0.01 – 2 mm), moderate (2 – 6 mm) or heavy precipitation ($\geq$ 6 mm), the histograms of ranks of the nearest adjusted members over all grid points are considered. By fitting a beta distribution to an 11-bin closest-member histogram one can generate weights either for the operational TCo639 or for the experimental TCo399 ensemble forecasts (controls included, resulting in 51 and 201 bins,

respectively). For a detailed description of weighted quantile mapping we again refer to Hamill and Scheuerer (2018) and Gascón et al. (2019). Dual-resolution combinations are then formed from the weighted members of TCo639 and TCo399 forecasts.

d. Verification scores

The predictive performance of various probabilistic forecasts is quantified with the help of the mean CRPS over all forecast cases in the verification period. Besides this verification measure, we also consider the Brier score (BS: Wilks 2019, Section 9.4.2) for the dichotomous event that the observed precipitation accumulation x is above a particular threshold y. Given again a predictive CDF $F(y)$ representing a probabilistic forecast, the BS is defined as

$$\mathrm{BS}(F,x;y) \coloneqq \big(F(y) - \mathbb{I}_{\{y \geq x\}}\big)^2. \tag{4}$$

In line with the reference study of Gascón et al. (2019), in Section 4 we report the mean BS of 24h accumulated precipitation for thresholds 0.1, 5 and 10 mm. Note that BS is again a negatively oriented score and the CRPS is the integral of the BS over all thresholds. Further, for a probabilistic forecast provided in the form of a forecast ensemble, both in (3) and in (4) the predictive CDF F should be replaced by the empirical one.

One can gain a better insight into the smaller differences in the predictive performance of the competing forecasts by examining the continuous ranked probability skill scores (CRPSS: see e.g. Gneiting and Raftery 2007) and Brier skill scores (BSS: see e.g. Friederichs and Thorarinsdottir 2012) quantifying improvement in CRPS and BS of a forecast F over a reference forecast F_{ref}, respectively. If $\overline{\mathrm{CRPS}}$, $\overline{\mathrm{BS}}$ and $\overline{\mathrm{CRPS}}_{ref}$, $\overline{\mathrm{BS}}_{ref}$ denote the mean score values over the verification data corresponding to F and F_{ref}, respectively, CRPSS and BSS are defined as

$$\mathrm{CRPSS} \coloneqq 1 - \frac{\overline{\mathrm{CRPS}}}{\overline{\mathrm{CRPS}}_{ref}} \qquad \text{and} \qquad \mathrm{BSS} \coloneqq 1 - \frac{\overline{\mathrm{BS}}}{\overline{\mathrm{BS}}_{ref}}.$$

The calibration of probabilities of a dichotomous event of exceeding a given threshold calculated from the various competing forecasts is compared with the help of reliability diagrams (Wilks 2019, Section 9.4.4) depicting the graph of the observed relative frequencies of the event against the corresponding binned forecast probabilities. In the case of proper calibration, this graph should

lie on the main diagonal of the unit square. We follow the suggestions of Bröcker and Smith (2007) and plot the observed relative frequency of a bin against the mean of the corresponding probabilities and we also add inset plots showing the frequencies of the bins on log 10 scales.

Finally, the statistical significance of the differences between the verification scores is assessed with two different methods. On the one hand, we report confidence intervals for the mean score values and skill scores calculated from 2000 block bootstrap samples based on the stationary bootstrap scheme with mean block length according to Politis and Romano (1994). On the other hand, we apply the Diebold–Mariano (DM) test for equal predictive performance (Diebold and Mariano 1995), which is able to account for temporal dependencies. In simultaneous testing for the different locations, we also handle spatial dependencies by applying a Benjamini-Hochberg algorithm (Benjamini and Hochberg 1995) to control the false discovery rate at a 5 % level of significance (see e.g. Wilks 2016).

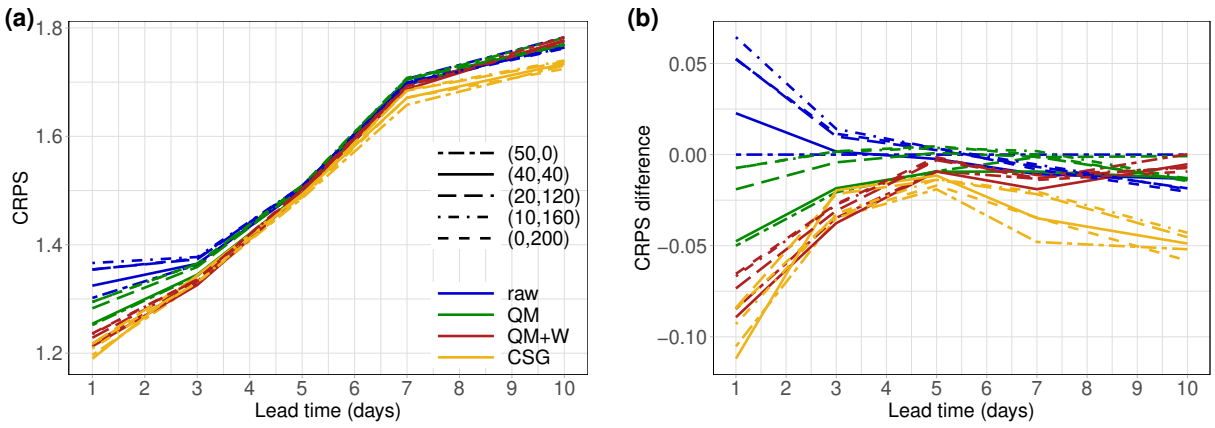
4. Results

As mentioned in the Introduction, in this study we investigate the effect of CSG EMOS post-processing (see Section 3a) on various combinations of TCo639 and TCo399 24h precipitation accumulation ensemble forecasts, including the pure high- and pure low-resolution ensemble. In the following analysis, the notation $f_{H,1}, f_{H,2}, \ldots, f_{H,M_H}$ is used for the ensemble members at TCo639 resolution and $f_{L,1}, f_{L,2}, \ldots, f_{L,M_L}$ for the TCo399 members of the dual-resolution forecast for a given location, time point and lead time. As we consider only forecasts obtained using perturbed initial conditions for both resolutions, where all members are equally likely, ensemble members at a given resolution can be considered as exchangeable. Hence, link functions (2) of the CSG EMOS model reduce to

$$\mu = a^2 + b_H^2 \overline{f}_H + b_L^2 \overline{f}_L \qquad \text{and} \qquad \sigma^2 = c^2 + d^2 \overline{f},$$

where $\overline{f}_H$ and $\overline{f}_L$ denote the mean of high- and low-resolution members, respectively. Model parameters are estimated by minimizing the mean CRPS over the training data where we fix $b_L = 0$ for the pure high-resolution $(M_L = 0)$ and $b_H = 0$ for the pure low-resolution $(M_H = 0)$ case.

Due to the large number of zeros both in predicted and observed precipitation accumulation, statistical post-processing of this weather variable requires far more training data than e.g. temperature



303 FIG. 2. (a) CRPS of raw and post-processed forecasts; (b) difference in CRPS from the raw (50,0) combination
304 as function of the forecast horizon.

285 or wind speed. For local EMOS models, Hemri et al. (2014) suggest using data of almost 5 years
286 (1816 calendar days), whereas our ECMWF dual-resolution forecasts cover just the 97-day time
287 interval between 1 June 2016 and 5 September 2016. The extension and heterogeneity of the EFAS
288 domain (Figure 1a) make regional modelling unreliable as well, so as mentioned in Section 3b, a
289 clustering-based semi-local approach is applied. Similar to Lerch and Baran (2017), the clustering
290 of grid points is performed using 24-dimensional feature vectors assigned to each location, where
291 half of the features are obtained by taking equidistant quantiles of the climatological CDF over the
292 training period, whereas the other half consists of the same equidistant quantiles of the empirical
293 distribution of the forecast error of the ensemble mean. After a detailed data analysis where several
294 combinations of the training period length and the number of clusters had been tested, we decided
295 to estimate the parameters of the CSG EMOS model over 8000 clusters using a 30-day rolling
296 training period. This configuration results in an average of 1363 forecast-observation pairs for each
297 estimation task (5 or 6 parameters to be estimated) and leaves 52 calendar days for verification
298 purposes (period 11 July 2016 – 31 August 2016). As noted before, CSG EMOS post-processed
299 predictions are validated using data from just 2370 SYNOP stations (Figure 1b), allowing a di-
300 rect comparison with quantile-mapped (QM) and weighted quantile-mapped (QM+W) forecasts
301 of Gascón et al. (2019), and to be fully in line with this work, we report the various verification
302 scores only for forecast horizons 1, 3, 5, 7 and 10 days.

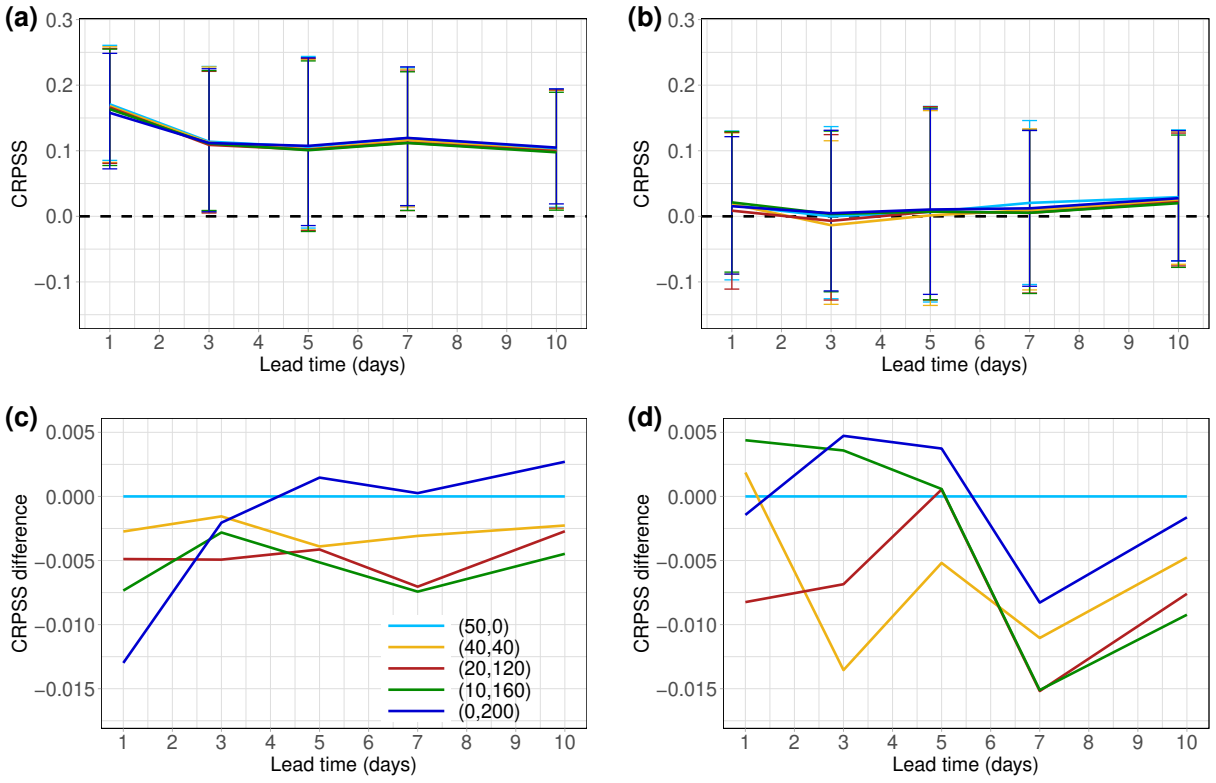


FIG. 3. CRPSS of the CSG EMOS model for different dual-resolution configurations (a) with respect to the raw (50,0) combination; (b) with respect to the corresponding QM+W forecast with 95 % confidence intervals. Panels (c) and (d) provide the differences in CRPSS from the curves in (a) and (b), respectively, corresponding to mixture (50,0).

Figure 2 shows the mean CRPS of raw and post-processed forecasts of the investigated dual-resolution combinations. Through day 5 all post-processed forecast combinations outperform all raw dual-resolution forecasts, the largest differences appear at day 1, whereas the smallest gain appears at day 5. For longer lead times the advantage of QM and QM+W forecasts disappears and the CSG EMOS models result in the lowest mean CRPS.

A slightly better insight into the differences between the various forecasts can be obtained from Figure 3, where the skill scores of the dual-resolution CSG EMOS models with respect to the raw pure high-resolution (50,0) forecast (Figures 3a,c) and with respect to the corresponding QM+W forecast (Figures 3b,d) are plotted. The difference in skill between CSG EMOS models corresponding to various dual-resolution combinations is negligible, which is in line with the findings of Baran et al. (2019), and they significantly outperform the raw high-resolution forecasts

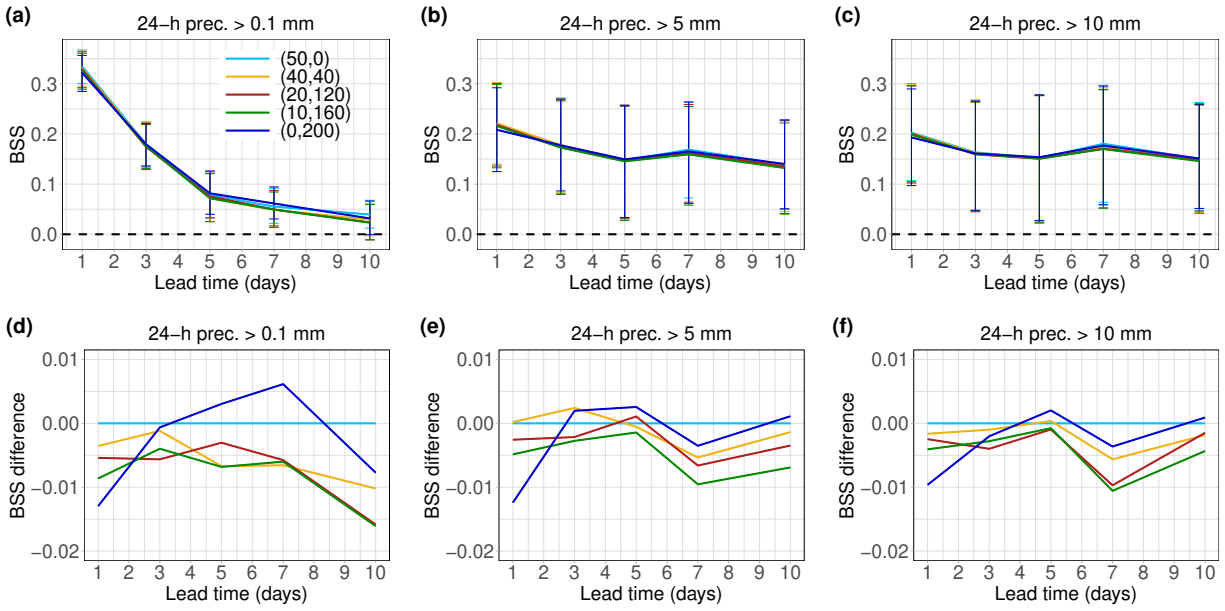


FIG. 4. BSS of the CSG EMOS model for different dual-resolution configurations with respect to the raw (50,0) configuration with 95 % confidence intervals for thresholds (a) 0.1 mm; (b) 5 mm; (c) 10 mm. Panels (d), (e) and (f) provide the differences in BSS from the curves in (a), (b) and (c), respectively, corresponding to mixture (50,0).

at a 5 % level of significance for all investigated lead times except day 5, where the lower bounds of the corresponding 95 % error bars are all slightly below zero. Further, according to Figure 3b, the essentially simpler CSG EMOS approach is completely able to catch up with the QM+W forecasts for all dual-resolution configurations and all lead times.

The analysis of Brier skill scores with thresholds of 0.1 mm, 5 mm and 10 mm leads us to rather similar conclusions. CSG EMOS forecasts outperform the ECMWF high-resolution (50,0) precipitation accumulation forecast for all investigated lead times for all dual-resolution combinations and all three thresholds (Figure 4); however, for 0.1 mm at day 10 the difference is significant on a 5 % level only for the EMOS models based either on the pure high-, or on the pure low-resolution ensemble. Further, Figure 5 again confirms that there is no dual-resolution combination and forecast horizon where the difference in skill between the matching CSG EMOS and QM+W forecasts is significant on a 5 % level.

The simultaneous DM tests for all considered stations also confirm, that there are no real differences in skill between the CSG EMOS models corresponding to various dual-resolution

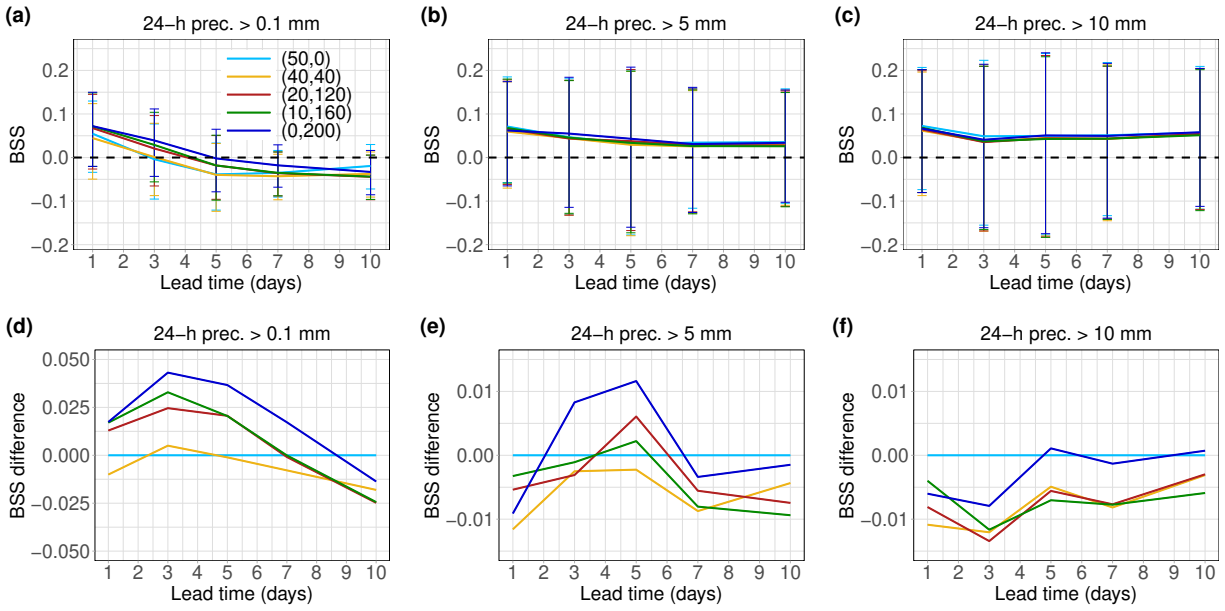


FIG. 5. Brier Skill scores with respect to the corresponding QM+W forecasts for each dual-resolution combination with 95 % confidence intervals for all 3 thresholds. Panels (d), (e) and (f) provide the differences in BSS from the curves in (a), (b) and (c), respectively, corresponding to mixture (50,0).

mixtures. At days 1, 3, 5 and 10 practically there are no stations where the difference in mean CRPS between any pairs of combinations is significant at a 5 % level, whereas at day 7 only mixtures (50,0) and (40,40) differ significantly at 1.47 % of the locations. Up to day 7, this is also the only mixture, where for all three thresholds there are stations with significantly different mean BS; however, their proportions vary just between 4.41 % and 6.85 % for 0.1 mm, 3.4 % and 10.72 % for 5 mm and 2.13 % and 11.68 % for 10 mm. At day 10 the situation changes, as for all pairs of mixtures and all thresholds there are locations with significantly different mean BS; though their proportions just barely exceed 9 %.

Further, Figure 6 displays the reliability diagrams for 0.1, 5 and 10 mm thresholds of raw (50,0) and (40,40) combinations and corresponding CSG EMOS forecasts for days 1, 5 and 10. At day 1 CSG EMOS models definitely outperform the raw forecasts, especially for the 0.1 mm threshold, where the fit to the reference line is almost perfect. For longer lead times the clear advantage of post-processing is preserved only for the lowest threshold where the diagrams are based on 36.5 % (day 5) and 34.9 % (day 10) of the observations, whereas for 5 mm and 10 mm these proportions are 11.6 % and 10.8 % and 5.5 % and 6.2 %. Further, as the corresponding inset histograms indicate,

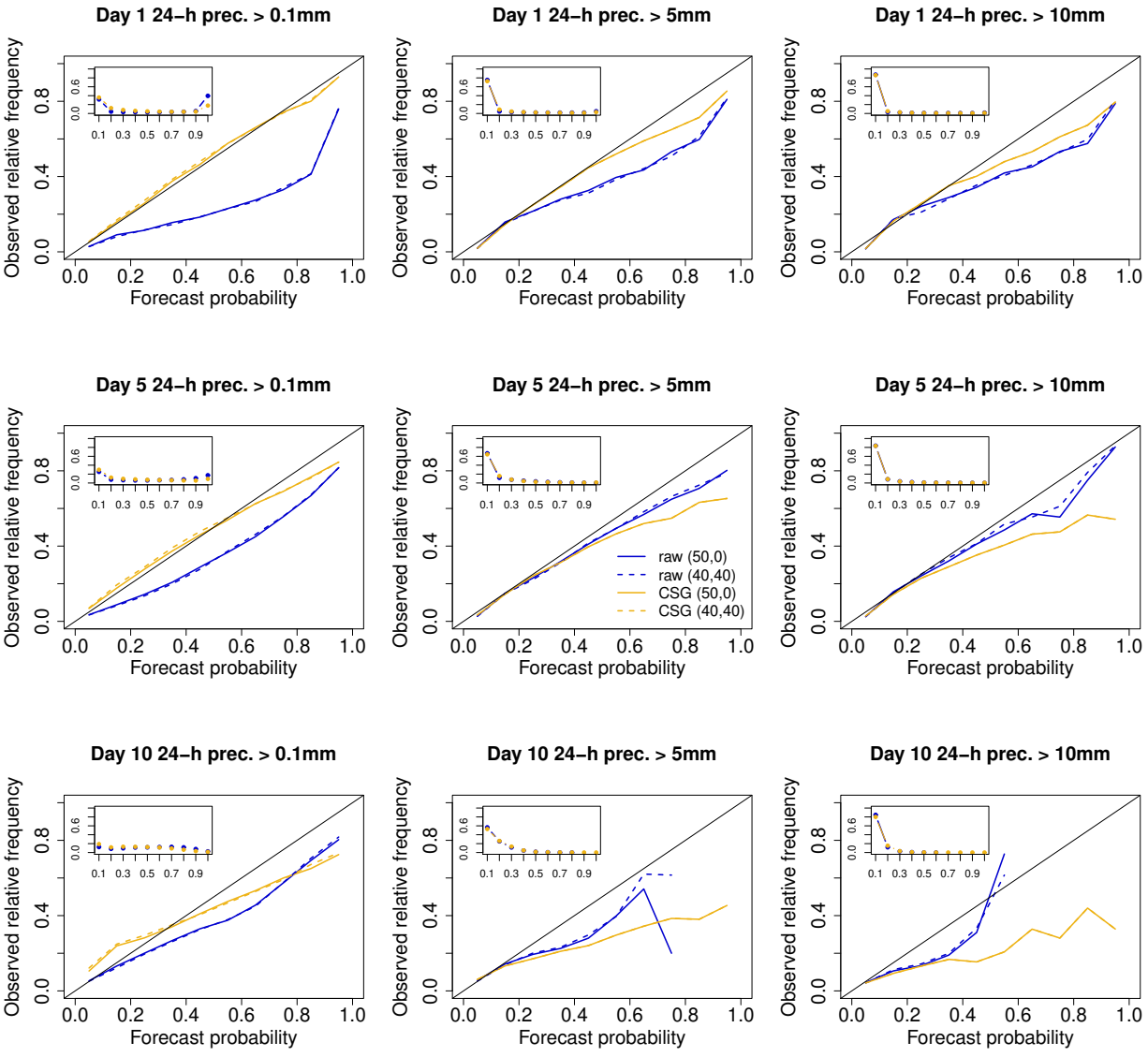


FIG. 6. Reliability diagrams for 0.1, 5 and 10 mm thresholds of raw (50,0) and (40,40) combinations and corresponding CSG EMOS forecasts for days 1, 5 and 10. The inset curves display the relative frequency of cases within the respective bins for the (50,0) mixture.

the distribution of forecast cases is rather biased with very low frequencies at the upper bins. This shortage of data might explain the hectic behaviour of the reliability diagrams for 5 mm and 10 mm thresholds at day 10.

5. Conclusions

The predictive performance of the censored shifted gamma EMOS approach of Baran and Nemoda (2016) for statistical post-processing is investigated with the help of various dual-resolution 24h precipitation accumulation ensemble forecasts over Europe. All considered dual-resolution combinations have equal computational costs, which are equivalent to the cost of the operational 50-member ECMWF ensemble. As reference post-processing approaches we consider the quantile mapping and weighted quantile mapping of Hamill and Scheuerer (2018). All calibration methods are trained using forecast-analysis pairs at EFAS grid points and validated on data of grid points corresponding to SYNOP stations.

Compared with the raw ensemble combinations, semi-local EMOS post-processing results in a significant improvement for all studied lead times both in terms of the mean CRPS and the mean BS for various thresholds. Moreover, in contrast to the raw ensemble where up to day 5 the mixture of 40 high- and 40 low-resolution forecasts significantly outperforms the other combinations (Gascón et al. 2019), there are no significant differences between the skill of CSG EMOS forecasts corresponding to the various mixtures. Further, in terms of the mean CRPS, CSG EMOS forecasts outperform the reference QM and QM+W predictions for all lead times; however, none of the differences are significant. The same is true for the differences between CSG EMOS and QM+W in terms of the Brier scores. These results indicate that the semi-local CSG EMOS method trained merely using data from a 30-day rolling training period is fully able to catch up with the essentially more complex quantile mapping based on historical data of 20 years.

The introduction of the new cycle at the ECMWF from 2023, where the current operational setup of a single TCo1279 and 51 TCo639 forecasts will be replaced by 51 forecasts at TCo1279 resolution and 101 forecasts at TCo319 resolution immediately provides new avenues of further research on the calibration of dual-resolution predictions. Another possible direction is the investigation of the skill of machine learning-based parametric post-processing approaches in the dual-resolution context, focusing on methods that, similar to EMOS, require short training data, see e.g. Baran and Baran (2021, 2022); Ghazvinian et al. (2022).

Acknowledgments. Sándor Baran and Marianna Szabó were supported by the Hungarian National Research, Development and Innovation Office under Grant No. K142849. They are also grateful to the ECMWF for supporting their research stay in Reading. Further, the authors are indebted to

Martin Leutbecher for his help in connection with the new ECMWF IFS configuration. Last but not least, the authors are grateful to the reviewers and the editor for their valuable comments.

Data availability statement. The data used in this study are proprietary, and the authors are not allowed to share it. However, it may be obtained from the European Centre for Medium-Range Weather Forecasts directly for research purposes.

References

Baran, S., and Á. Baran, 2021: Calibration of wind speed ensemble forecasts for power generation. *Időjárás*, **125**, 609–624, https://doi.org/10.28974/idojaras.2021.4.4.

Baran, S., and Á. Baran, 2022: A two-step machine learning approach to statistical post-processing of weather forecasts for power generation. *arXiv:2207.07589*, https://doi.org/10.48550/arXiv.2207.07589.

Baran, S., M. Leutbecher, M. Szabó, and Z. Ben Bouallègue, 2019: Statistical post-processing of dual-resolution ensemble forecasts. *Q. J. R. Meteorol. Soc.*, **145**, 1705–1720, https://doi.org/10.1002/qj.3521.

Baran, S., and D. Nemoda, 2016: Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting. *Environmetrics*, **27**, 280–292, https://doi.org/10.1002/env.2391.

Benjamini, Y., and Y. Hochberg, 1995: Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.*, **57**, 289–300, https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Bremnes, J. B., 2019: Constrained quantile regression splines for ensemble postprocessing. *Mon. Wea. Rev.*, **147**, 1769–1780, https://doi.org/10.1175/MWR-D-18-0420.1.

Bremnes, J. B., 2020: Ensemble postprocessing using quantile function regression based on neural networks and Bernstein polynomials. *Mon. Wea. Rev.*, **148**, 403–414, https://doi.org/10.1175/MWR-D-19-0227.1.

Bröcker, J., and L. A. Smith, 2007: Increasing the reliability of reliability diagrams. *Wea. Forecasting*, **22**, 651–661, https://doi.org/10.1175/WAF993.1.

Buizza, R., 2018: Ensemble forecasting and the need for calibration. In Vannitsem, S., Wilks, D. S., Messner, J. W. (eds.). *Statistical Postprocessing of Ensemble Forecasts*, Elsevier, Amsterdam, 15–48, https://doi.org/10.1016/B978-0-12-812372-0.00002-9.

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

Diebold, F. X., and R. S. Mariano, 1995: Comparing predictive accuracy. *J. Bus. Econ. Stat.*, **13**, 253–263, https://doi.org/10.1080/07350015.1995.10524599.

Friederichs, P., and A. Hense, 2007: Statistical downscaling of extreme precipitation events using censored quantile regression. *Mon. Wea. Rev.*, **135**, 2365–2378, https://doi.org/10.1175/MWR3403.1.

Friederichs, P., and T. L. Thorarinsdottir, 2012: Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics*, **23**, 579–594, https://doi.org/10.1002/env.2176.

Gascón, E., D. Lavers, T. M. Hamill, D. S. Richardson, Z. Ben Bouallègue, M. Leutbecher, and F. Pappenberger, 2019: Statistical postprocessing of dual-resolution ensemble precipitation forecasts across Europe. *Q. J. R. Meteorol. Soc.*, **145**, 3218–3235, https://doi.org/10.1002/qj.3615.

Ghazvinian, M., Y. Zhang, T. M. Hamill, D.-J. Seo, and N. Fernando, 2022: Improving probabilistic quantitative precipitation forecasts using short training data through artificial neural networks. *J. Hydrometeor.*, **23**, 1365–1382, https://doi.org/10.1175/JHM-D-22-0021.1.

Gneiting, T., and A. E. Raftery, 2005: Weather forecasting with ensemble methods. *Science*, **310**, 248–249, https://doi.org/10.1126/science.1115255.

Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction and estimation. *J. Am. Stat. Assoc.*, **102**, 359–378, https://doi.org/10.1198/016214506000001437.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, https://doi.org/10.1175/MWR2904.1.

Haiden, T., M. Janousek, J. Bidlot, R. Buizza, L. Ferranti, F. Prates, and F. Vitart, 2018: Evaluation of ECMWF forecasts, including the 2018 upgrade. *ECMWF Technical Memorandum*, no. 831, https://www.ecmwf.int/sites/default/files/elibrary/2018/18746-evaluation-ecmwf-forecasts-including-2018-upgrade.pdf.

Hamill, T. M., R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: precipitation. *Mon. Wea. Rev.*, **136**, 2620––2632, https://doi.org/10.1175/2007MWR2411.1.

Hamill, T. M., and M. Scheuerer, 2018: Probabilistic precipitation forecast postprocessing using quantile mapping and rank-weighted best-member dressing. *Mon. Wea. Rev.*, **146**, 4079—4098, https://doi.org/10.1175/MWR-D-18-0147.1.

Hemri, S., M. Scheuerer, F. Pappenberger, K. Bogner, and T. Haiden, 2014: Trends in the predictive performance of raw ensemble weather forecasts. *Geophys. Res. Lett.*, **41**, 9197–9205, https://doi.org/10.1002/2014GL062472.

Lerch, S., and S. Baran, 2017: Similarity-based semi-local estimation of EMOS models. *J. R. Stat. Soc. Ser. C Appl. Stat.*, **66**, 29–51, https://doi.org/10.1111/rssc.12153.

Leutbecher, M., , and Z. Ben Bouallègue, 2020: On the probabilistic skill of dual-resolution ensemble forecasts. *Q. J. R. Meteorol. Soc.*, **146**, 707–723, https://doi.org/10.1002/qj.3704.

Leutbecher, M., 2018: Ensemble size: How suboptimal is less than infinity. *Q. J. R. Meteorol. Soc.*, **145**, 107–128, https://doi.org/10.1002/qj.3387.

Machete, R. L., and L. A. Smith, 2016: Demonstrating the value of larger ensembles in forecasting physical systems. *Tellus A*, **68**, 28 393, https://doi.org/10.3402/tellusa.v68.28393.

Ntegeka, V., P. Salomon, G. Gomes, H. Sint, V. Lorini, M. Zambrano-Bigiarini, and J. Thielen, 2013: *EFAS-Meteo: a European daily high-resolution gridded meteorological data set for 1990—2011.*, EU, Ispra: Joint Research Centre, Technical Report JRC86388.

Politis, D. N., and J. P. Romano, 1994: The stationary bootstrap. *J. Am. Stat. Assoc.*, **89**, 1303–1313, https://doi.org/10.1080/01621459.1994.10476870.

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

Rasp, S., and S. Lerch, 2018: Neural networks for postprocessing ensemble weather forecasts. *Mon. Wea. Rev.*, **146**, 3885–3900, https://doi.org/10.1175/MWR-D-18-0187.1.

Scheuerer, M., 2014: Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Q. J. R. Meteorol. Soc.*, **140**, 1086–1096, https://doi.org/10.1002/qj.2183.

Scheuerer, M., and T. M. Hamill, 2015: Statistical post-processing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Wea. Rev.*, **143**, 4578—-4596, https://doi.org/10.1175/MWR-D-15-0061.1.

Scheuerer, M., M. B. Switanek, R. P. Worsnop, and T. M. Hamill, 2020: Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over California. *Mon. Wea. Rev.*, **148**, 3489–3506, https://doi.org/10.1175/MWR-D-20-0096.1.

Thorarinsdottir, T. L., and T. Gneiting, 2010: Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *J. R. Stat. Soc. Ser. A Stat. Soc.*, **173**, 371–388, https://doi.org/10.1111/j.1467-985X.2009.00616.x.

Van Schaeybroeck, B., and S. Vannitsem, 2015: Ensemble post-processing using member-by-member approaches: Theoretical aspects. *Q. J. R. Meteorol. Soc.*, **141**, 807–818, https://doi.org/10.1002/qj.2397.

Vannitsem, S., and Coauthors, 2021: Statistical postprocessing for weather forecasts – review, challenges and avenues in a big data world. *Bull. Am. Meteorol. Soc.*, **102**, E681–E699, https://doi.org/10.1175/BAMS-D-19-0308.1.

Wilks, D. S., 2016: "The stippling shows statistically significant grid points": How research results are routinely overstated and overinterpreted, and what to do about it. *Bull. Am. Meteorol. Soc.*, **97**, 2263–2273, https://doi.org/10.1175/BAMS-D-15-00267.1.

Wilks, D. S., 2018: Univariate ensemble forecasting. In Vannitsem, S., Wilks, D. S., Messner, J. W. (eds.). *Statistical Postprocessing of Ensemble Forecasts*, Elsevier, Amsterdam, 49–89, https://doi.org/10.1016/B978-0-12-812372-0.00003-0.

Wilks, D. S., 2019: *Statistical Methods in the Atmospheric Sciences. 4th ed.*, Elsevier, Amsterdam.