Received: 22 June 2022 | Revised: 11 January 2023 | Accepted: 27 January 2023

DOI: 10.1002/qj.4436

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

Comparison of multivariate post-processing methods using global ECMWF ensemble forecasts

Mária Lakatos[1,2] | **Sebastian Lerch**[3] | **Stephan Hemri**[4] | **Sándor Baran**[1]

[1]Faculty of Informatics, University of Debrecen, Debrecen, Hungary

[2]Doctoral School of Informatics, University of Debrecen, Debrecen, Hungary

[3]Institute of Economics, Karlsruhe Institute of Technology, Karlsruhe, Germany

[4]Department of Mathematics, University of Zurich, Zurich, Switzerland

Correspondence
Sándor Baran, Faculty of Informatics, University of Debrecen, Kassai út 26, H-4028 Debrecen, Hungary.
Email: baran.sandor@inf.unideb.hu

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: MO-3394/1-1; Nemzeti Kutatási Fejlesztési és Innovációs Hivatal, Grant/Award Number: NN125679; Vector Stiftung

Abstract

An influential step in weather forecasting was the introduction of ensemble forecasts in operational use owing to their capability to account for the uncertainties in the future state of the atmosphere. However, ensemble weather forecasts are often underdispersive and might also contain bias, which calls for some form of post-processing. A popular approach to calibration is the ensemble model output statistics approach resulting in a full predictive distribution for a given weather variable. However, this form of univariate post-processing may ignore the prevailing spatial and/or temporal correlation structures among different dimensions. Since many applications call for spatially and/or temporally coherent forecasts, multivariate post-processing aims to capture these possibly lost dependencies. We compare the forecast skill of different non-parametric multivariate approaches to modeling temporal dependence of ensemble weather forecasts with different forecast horizons. The focus is on two-step methods, where, after univariate post-processing, the ensemble model output statistics predictive distributions corresponding to different forecast horizons are combined to a multivariate calibrated prediction using an empirical copula. Based on global ensemble predictions of temperature, wind speed, and precipitation accumulation of the European Centre for Medium-Range Weather Forecasts from January 2002 to March 2014, we investigate the forecast skill of different versions of ensemble copula coupling (ECC) and Schaake shuffle. In general, compared with the raw and independently calibrated forecasts, multivariate post-processing substantially improves the forecast skill. Although even the simplest ECC approach with low computational cost provides a powerful benchmark method, recently proposed advanced extensions of the ECC and the Schaake shuffle are found to not provide any significant improvements over their basic counterparts.

KEYWORDS

ensemble copula coupling, ensemble model output statistics, multivariate post-processing, Schaake shuffle

1 | INTRODUCTION

In many cases, ensemble forecasts generated by numerical weather prediction (NWP) models suffer from systematic biases and underdispersion. In order to correct for these issues, over the last two decades a plethora of different post-processing approaches have been developed. They range from classical regression-type approaches, like ensemble model output statistics (EMOS; Gneiting *et al.*, 2005) or Bayesian model averaging (BMA; Raftery *et al.*, 2005), to sophisticated machine-learning-based approaches, such as, for instance, applied in Schulz and Lerch (2022), Vannitsem *et al.* (2021) provide a comprehensive summary of statistical post-processing.

Most post-processing approaches implicitly assume statistical independence between different forecast margins, like lead times, locations, or weather variables. Apparently, this assumption does not allow for providing realistic forecast scenarios. However, end-users may be interested in scenarios like total precipitation over an entire hydrological catchment, the temporal evolution of precipitation, or the interaction of precipitation and temperature, in particular when temperature is close to 0°C. Therefore, for many use cases, dependence structures need to be re-established explicitly in a second post-processing step after univariate calibration. To this end, different copula-based approaches have been proposed. Most state-of-the-art multivariate post-processing applications employ an empirical copula based on a dependence template stemming either from an NWP ensemble or historical observations. The former and the latter are referred to as ensemble copula coupling (ECC; Schefzik *et al.*, 2013) and Schaake shuffle (SSh; Clark *et al.*, 2004) respectively. We refer to Schefzik and Möller (2018) for a detailed discussion of copula-based methods to incorporate dependence structures.

Lerch *et al.* (2020a) performed an extensive simulation study comparing several variants of ECC with SSh and the parametric Gaussian copula approach (GCA; Möller *et al.*, 2013), and they concluded that the benchmark methods, ECC with equidistant quantile sampling (ECC-Q; Schefzik *et al.*, 2013) and SSh, generally performed well. Moreover, their results suggest that more sophisticated approaches like, for instance, dual ECC (dECC; Ben Bouallègue *et al.*, 2016) or the Gaussian copula approach provide substantial benefits to predictive performance only in very specific conditions. In this study, we assess whether these findings can be confirmed by using real NWP ensemble forecasts and observations. To this end, we apply univariate EMOS combined with a whole range of ECC and SSh variants to NWP ensemble forecasts of temperature, wind speed, and precipitation provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). Observations from SYNOP stations are used for verification. To the best of our knowledge, our work is the first to compare a large variety of state-of-the-art two-step methods for multivariate post-processing including recently proposed similarity-based SSh approaches and dECC.

Since we focus on the multivariate post-processing step, for the univariate calibration we follow simply the EMOS implementations by Hemri *et al.* (2014). However, for the multivariate step we compare three naive approaches, which are derived more or less directly from univariate EMOS, with four variants of ECC and five variants of the SSh. For the sake of simplicity and comprehensibility, we focus on temporal dependence.

After a short description of the data in Section 2, we provide a detailed summary of both univariate EMOS and all approaches we apply to re-establish the multivariate dependence structure in Section 3. In Section 4 we present our results, followed by a brief discussion and conclusions in Section 5.

2 | DATA

We compare the various multivariate post-processing methods with the help of global ECMWF ensemble forecasts of 2 m temperature (T2M), 10 m wind speed (V10), and 24 hr precipitation accumulation (PPT24) for the period between January 1, 2002, and March 20, 2014. The datasets at hand are identical to the ones investigated in Hemri *et al.* (2014). containing the ECMWF high-resolution forecasts, the 50-member operational ensemble generated using random perturbations, and the control run initialized at 1200 UTC with forecast horizons ranging from 1 to 10 days, together with the corresponding observations. After an initial quality control removing SYNOP stations with missing or irregular data, 4,160, 4,388, and 2,917 stations covering the entire globe remained for T2M, V10, and PPT24 respectively. For more details about the data investigated and the quality control procedure applied, see Hemri *et al.* (2014).

3 | METHODS

As mentioned in Section 1, we restrict our attention to two-step approaches to multivariate post-processing, where, after an initial univariate calibration, multivariate predictions are obtained by combining the individual post-processed forecasts with the help of an empirical copula.

In what follows, let $\boldsymbol{f}^{(\ell)} = (f_1^{(\ell)}, f_2^{(\ell)}, \ldots, f_{52}^{(\ell)})$ denote a 52-member ECMWF ensemble forecast with a lead

time of ℓ days ($\ell = 1, 2, \ldots, 10$) initialized at a given time point, where $f_1^{(\ell)} = f_{\text{HRES}}^{(\ell)}$ and $f_2^{(\ell)} = f_{\text{CTRL}}^{(\ell)}$ are the high-resolution and the control member respectively, whereas $f_3^{(\ell)}, f_4^{(\ell)}, \ldots, f_{52}^{(\ell)}$ correspond to the 50 statistically indistinguishable (and thus exchangeable) ensemble members, which are generated using random perturbations and which we will be denoted by $f_{\text{ENS},1}^{(\ell)}, f_{\text{ENS},2}^{(\ell)}, \cdots, f_{\text{ENS},50}^{(\ell)}$.

3.1 | Univariate post-processing

For calibrating ensemble forecasts for a given location and time point with a given forecasts horizon, one can choose from a multitude of state-of-the-art approaches, as mentioned in Section 1. Here, we consider the computationally simple and efficient EMOS method, where post-processed forecasts are obtained in the form of parametric predictive distributions with parameters depending on the corresponding ensemble predictions. EMOS models for various weather quantities differ in the parametric distribution family to be used and in link functions relating the parameters of the predictive distribution to the raw ensemble forecasts. For univariate calibration of T2M, V10, and PPT24 ensemble forecasts, we make use of the specific EMOS approaches of Hemri *et al.* (2014) described briefly in Sections 3.1.1,3.1.2, and 3.1.3 respectively. To simplify notation we omit the indication of the forecast horizon and use notation $f_k = f_k^{(\ell)}, k = 1, 2, \ldots, 52$, in these sections.

3.1.1 | Temperature

The normal distribution and its generalizations (skewed normal, mixture of normals, etc.) are widely used to model temperature (see e.g. Gneiting *et al.*, 2005; Raftery *et al.*, 2005; Rasp and Lerch, 2018; Taillardat, 2021). The EMOS predictive distribution suggested by Gneiting (2014) specifically for the 52-member T2M ECMWF ensemble is Gaussian with mean μ and variance σ^2, given as

$$\begin{aligned} \mu &= a_0 + a_1^2 f_{\text{HRES}} + a_2^2 f_{\text{CTRL}} + a_3^2 \bar{f}_{\text{ENS}} \quad \text{and} \\ \sigma^2 &= b_0^2 + b_1^2 S_{\text{ENS}}^2, \end{aligned}$$

where $\bar{f}_{\text{ENS}}$ and S_{ENS}^2 respectively denote the mean and the variance of the 50 exchangeable ensemble members; that is,

$$\bar{f}_{\text{ENS}} := \frac{1}{50}\sum_{k=1}^{50} f_{\text{ENS},k} \quad \text{and} \quad S_{\text{ENS}}^2 := \frac{1}{50}\sum_{k=1}^{50} (f_{\text{ENS},k} - \bar{f}_{\text{ENS}})^2.$$

Model parameters $a_0, a_1, a_2, a_3, b_0, b_1 \in \mathbb{R}$ are estimated according to the optimum score principle of Gneiting and Raftery (2007); that is, by optimizing the mean value of an appropriate verification metric (see Section 3.3) over the training data consisting of past forecast–observation pairs.

To account for seasonal variations in temperature, inspired by Scheuerer and Büermann (2014), Hemri *et al.* (2014) suggest a more complex model for the mean of the predictive distribution in the form of

$$\begin{aligned} y_t &= c_0 + c_1 \sin\left(\frac{2\pi t}{365}\right) + c_2 \cos\left(\frac{2\pi t}{365}\right) + \varepsilon_t, \\ t &\in 1, 2, \ldots, n, \end{aligned} \tag{1}$$

where the dependent variables y_t are either temperature observations for a given location or functionals of the corresponding ECMWF ensemble forecast with a given lead time ℓ; namely, the high-resolution member f_{HRES}, the control run f_{CTRL}, and the mean of the exchangeable forecasts $\bar{f}_{\text{ENS}}$ from a training period of length n. With the help of the model in Equation (1) one can calculate the ℓ-step-ahead predictions $\hat{y}$ and $\hat{f}_{\text{HRES}}, \hat{f}_{\text{CTRL}}, \hat{f}_{\text{ENS}}$ of the observations and the corresponding functionals of the ensemble and obtain a Gaussian predictive distribution with parameters

$$\begin{aligned} \mu &= \hat{y} + a_1^2(f_{\text{HRES}} - \hat{f}_{\text{HRES}}) \\ &\quad + a_2^2(f_{\text{CTRL}} - \hat{f}_{\text{CTRL}}) + a_3^2(\bar{f}_{\text{ENS}} - \hat{f}_{\text{ENS}}) \quad \text{and} \\ \sigma^2 &= b_0^2 + b_1^2 S^2, \end{aligned} \tag{2}$$

where

$$S^2 := \frac{1}{52}\sum_{k=1}^{52}(f_k - \bar{f})^2 \quad \text{with} \quad \bar{f} := \frac{1}{52}\sum_{k=1}^{52} f_k.$$

3.1.2 | Wind speed

To model wind speed one requires non-negative and skewed distributions. Here, we consider a normal law $\mathcal{N}_0^\infty(\mu, \sigma^2)$ with location μ and scale $\sigma > 0$, left-truncated at zero, which is applied in the EMOS model of Thorarinsdottir and Gneiting (2010). As a natural transformation (see e.g. Haslett and Raftery, 1989), Hemri *et al.* (2014) model the square root of V10 as

$$\begin{aligned} &\mathcal{N}_0^\infty\Big(a_0 + a_1^2\sqrt{f_{\text{HRES}}} + a_2^2\sqrt{f_{\text{CTRL}}} \\ &\quad + a_3^2\sqrt{\bar{f}_{\text{ENS}}}, b_0^2 + b_1^2 \text{MD}_{\sqrt{f}}\Big), \end{aligned} \tag{3}$$

where

$$\mathrm{MD}_{\sqrt{f}} := \frac{1}{52^2}\sum_{k,\ell=1}^{52}\left|\sqrt{f_k}-\sqrt{f_\ell}\right|.$$

The model parameters $a_0, a_1, a_2, a_3, b_0, b_1 \in \mathbb{R}$, can again be estimated using the optimum score estimation principle.

3.1.3 | Precipitation

Statistical calibration of precipitation forecasts is more challenging than the post-processing of temperature or wind speed due to the discrete-continuous nature with a positive probability of observing zero precipitation. Following Hemri *et al.* (2014), to model PPT24 we consider the EMOS model of Scheuerer (2014), where the predictive distribution is a generalized extreme-value distribution left-censored at zero. Location μ and scale σ are linked to the raw forecasts via

$$\begin{aligned}\mu &= a_0 + a_1^2 f_{\mathrm{HRES}} + a_2^2 f_{\mathrm{CTRL}} + a_3^2 \bar{f}_{\mathrm{ENS}} + a_4^2 \pi_0 \text{ and}\\ \sigma^2 &= b_0^2 + b_1^2 \mathrm{MD}_f,\end{aligned} \tag{4}$$

where $a_0, a_1, a_2, a_3, a_4, b_0, b_1 \in \mathbb{R}$, and where

$$\pi_0 := \frac{1}{52}\sum_{k=1}^{52} I_{\{f_k=0\}} \text{ and } \mathrm{MD}_f := \frac{1}{52^2}\sum_{k,\ell=1}^{52}|f_k - f_\ell|,$$

are respectively the proportion of ensemble members predicting zero precipitation and the ensemble mean difference, whereas shape ξ is kept fixed ($\xi = 0.2$).

3.2 | Multivariate methods

Independent univariate calibration of ensemble forecasts $\boldsymbol{f}^{(\ell)}$ for each forecast horizon ℓ does not take into account the temporal dependencies between predictions initialized at the same time point. These dependencies are restored in a second step with the help of the approaches described in the following, which are based on empirical copulas.

In general, a copula is a multivariate cumulative distribution function (CDF) with standard uniform marginals. According to Sklar's theorem (Sklar, 1959), any L-dimensional CDF H with marginal CDFs $F^{(\ell)}, \ell = 1, 2, \ldots, L$, can be decomposed as

$$\begin{aligned}&H(x_1, x_2, \ldots, x_L) = C(F^{(1)}(x_1), F^{(2)}(x_2), \ldots, F^{(L)}(x_L)),\\ &x_1, x_2, \ldots, x_L \in \mathbb{R},\end{aligned}$$

where C is an L-dimensional copula representing the dependencies between the marginals. In our case $F^{(\ell)}$ is the predictive distribution corresponding to lead time $\ell = 1, \ldots, 10$ obtained after univariate post-processing, and we would like to combine calibrated predictions initialized at the same time point into a temporally consistent calibrated forecast trajectory represented by a 10-dimensional predictive CDF H.

Here, we focus on non-parametric methods where marginals are represented by empirical CDFs $\hat{F}^{(\ell)}$ of samples drawn independently from the corresponding predictive distributions, and C is an empirical copula (Rüschendorf, 2009) providing a "dependence template" derived from a given discrete dataset. The learned dependence structure is then imposed on the post-processed forecasts by rearranging the marginal samples according to the specified template (see e.g. Wilks, 2019, Sect. 8.4.2).

3.2.1 | Ensemble copula coupling

In the ECC approach, the dependence template is obtained from the corresponding raw ensemble forecasts. The method, introduced by Schefzik *et al.* (2013), consists of the following two simple steps.

1 For each dimension $\ell = 1, 2, \ldots, L$, generate a sample $\hat{f}_1^{(\ell)}, \hat{f}_2^{(\ell)}, \ldots, \hat{f}_K^{(\ell)}$ of the same size as the raw ensemble ($K = 52$ here) from the calibrated marginal predictive distribution $F^{(\ell)}$, which is assumed to be arranged in ascending order.
2 Consider permutations $\boldsymbol{\pi}_\ell = (\pi_\ell(1), \pi_\ell(2), \ldots, \pi_\ell(K))$ of $\{1, 2, \ldots, K\}$ induced by the rank order structure of the raw ensemble $f_1^{(\ell)}, f_2^{(\ell)}, \ldots, f_K^{(\ell)}$; that is, $\pi_\ell(k) := \mathrm{rank}(f_k^{(\ell)})$ with ties resolved at random. The ECC calibrated sample $\tilde{f}_1^{(\ell)}, \tilde{f}_2^{(\ell)}, \ldots, \tilde{f}_K^{(\ell)}$ for dimension ℓ is obtained by rearranging the sample generated in step 1 according to permutation $\boldsymbol{\pi}_\ell$; that is:

$$\tilde{f}_k^{(\ell)} := \hat{f}_{\pi_\ell(k)}^{(\ell)}, \quad k = 1, 2, \ldots, K, \quad \ell = 1, 2, \ldots, L.$$

Similar to Lerch *et al.* (2020b), we consider three different ECC variants depending on the sampling method in step 1: ECC-R, ECC-Q, and ECC-S. ECC-R refers to random sampling from the predictive distribution $F^{(\ell)}$ and arranging the sample in ascending order, whereas in ECC-Q one considers equidistant quantiles of $F^{(\ell)}$; that is:

$$\begin{aligned}&\hat{f}_k^{(\ell)} := (F^{(\ell)})^{-1}\left(\frac{k}{K+1}\right),\\ &k = 1, 2, \ldots, K, \quad \ell = 1, 2, \ldots, L.\end{aligned}$$

Finally, with the ECC-S variant we also apply the stratified sampling approach of Hu *et al.* (2016). In this approach

$$\hat{f}_k^{(\ell)} := (F^{(\ell)})^{-1}(u_k), \quad k = 1, 2, \ldots, K, \quad \ell = 1, 2, \ldots, L,$$

where u_k is a random draw from a uniform distribution on

$$\left]\frac{k-1}{K}, \frac{k}{K}\right], \quad k = 1, 2, \ldots, K.$$

Note that in Section 4.2 we also investigate the forecast skill of three naive multivariate forecasts obtained from the univariate post-processing methods without accounting for temporal dependencies. This means that forecasts $\hat{f}_k^{(\ell)}, k = 1, 2, \ldots, K, \ell = 1, 2, \ldots, L$, drawn from the corresponding predictive distributions $F^{(\ell)}$ are simply combined into L-dimensional ensemble forecasts $\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \ldots, \hat{\mathbf{f}}_K$ with $\hat{\mathbf{f}}_k = (\hat{f}_k^{(1)}, \hat{f}_k^{(2)}, \ldots, \hat{f}_k^{(L)})^{\mathrm{T}}, k = 1, 2, \ldots K$. These "independent" forecasts derived using random sampling from the univariate predictive distributions, considering equidistant quantiles and stratified sampling are denoted by EMOS-R, EMOS-Q, and EMOS-S respectively.

We remark that, in the case of EMOS-Q, the simple use of equidistant quantiles along each margin results in a positive dependence between each pair of coordinates that might lead to misleading results. To avoid the possibly unrealistic dependence structure one can randomly rearrange the assignments of the member indices of the equidistant quantiles obtained by EMOS post-processing (Schefzik, 2017), or even consider a number of such random shuffles and average the scores obtained (Schefzik, 2016). However, to be consistent with the methods of Lerch *et al.* (2020a), none of the aforementioned techniques is used in the present study.

3.2.2 | Dual ensemble copula coupling

The dECC method of Ben Bouallègue *et al.* (2016) combines the structure of the raw ensemble forecast with the estimated forecast error autocorrelation and proceeds as follows.

1 Apply ECC-Q to generate an initial post-processed multivariate ensemble forecast $\tilde{\mathbf{f}}_1, \tilde{\mathbf{f}}_2, \ldots, \tilde{\mathbf{f}}_K$ with $\tilde{\mathbf{f}}_k = \left(\tilde{f}_k^{(1)}, \tilde{f}_k^{(2)}, \ldots, \tilde{f}_k^{(L)}\right)^{\mathrm{T}}, k = 1, 2, \ldots K$.
2 With the help of the estimated $L \times L$ autocorrelation matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{e}}$ of the forecast error of the ensemble mean generate a correction term

$$\mathbf{c}_k := \hat{\boldsymbol{\Sigma}}_{\boldsymbol{e}}^{1/2}\left(\tilde{\mathbf{f}}_k - \mathbf{f}_k\right), \qquad k = 1, 2, \ldots, K,$$

where $\mathbf{f}_k = \left(f_k^{(1)}, f_k^{(2)}, \ldots, f_k^{(L)}\right)^{\mathrm{T}}$ denotes the kth raw multivariate forecast. The estimates of the error correlations can be obtained, for example, from the training data for the univariate post-processing at the different forecast horizons.
3 Derive the adjusted multivariate ensemble $\check{\mathbf{f}}_1, \check{\mathbf{f}}_2, \ldots, \check{\mathbf{f}}_K$, where $\check{\mathbf{f}}_k := \mathbf{f}_k + \mathbf{c}_k, k = 1, 2, \ldots, K$.
4 Apply ECC-Q again; however, this time using the rank order structure of the adjusted ensemble of step 3 for rearranging the samples generated from the calibrated univariate predictive distributions.

3.2.3 | Schaake shuffle

In contrast to ECC, the SSh (SSh; Clark *et al.*, 2004) determines a dependence template based on past observations rather than the raw ensemble predictions. Samples drawn from the calibrated univariate predictive distributions are thus rearranged in the rank order structure of randomly selected historical observation trajectories of length L. Again, as historical data, one can consider the training data for the univariate post-processing. The SSh approach allows for generating post-processed forecasts of any size, provided one has a long enough historical climatological record. However, to ensure a fair comparison with the (d)ECC methods, we restrict the sample size to the size of the raw ensemble. Similar to the ECC described in Section 3.2.1, three different sampling methods from the predictive distributions are considered; the corresponding SSh variants are referred to as SSh-R (random sample), SSh-Q (equidistant quantiles), and SSh-S (stratified sample). In addition to the standard SSh, we further apply two more recently proposed variants that utilize more advanced procedures to select past observations for determining the dependence template, the minimum divergence SSh (mdSSh) and the similarity-based SSh (simSSh), which are discussed next.

3.2.4 | Minimum divergence Schaake shuffle

The mdSSh (Scheuerer *et al.*, 2017) provides a more sophisticated method for selecting historical observation trajectories used as a dependence template. The basic selection algorithm applied for T2M and V10 ensemble forecasts is as follows.

1 For each lead time ℓ, calculate the 99% central prediction interval of the corresponding marginal predictive distribution $F^{(\ell)}, \ell = 1, 2, \ldots, L$.

2 From the historical climatological record, keep those observation trajectories of length L where the corresponding central prediction intervals of step 1 contain at least m observations. Threshold m is chosen to retain at least $M \geq K$ forecast trajectories.
3 Select randomly K observation trajectories from the M remaining after step 2.

For the discrete-continuous marginal predictive distributions of PPT24 (see Section 3.1.3), Scheuerer *et al.* (2017) suggests to replace step 3 by a more complex method to reduce the number of selected forecast trajectories from M to the required K. This approach is based on selection of a K-subset of the set of forecast trajectories of step 2, which minimizes the total divergence of the EMOS predictive CDFs and the empirical CDFs of the corresponding observations (see e.g. Thorarinsdottir *et al.*, 2013) for all lead times and forecast cases. For more details and a computationally feasible algorithm, see Scheuerer *et al.* (2017).

3.2.5 | Similarity-based Schaake shuffle

In contrast to the mdSSh, where observation trajectories used as dependence templates are selected on the basis of their consistency with the corresponding EMOS predictive distributions, the simSSh (Schefzik, 2016) looks for historical forecast trajectories where the corresponding ensemble forecasts are the most similar to the actual ones.

1 For a given initialization time in the verification period, calculate the mean $\bar{f}^{(\ell)}$ and variance $S^{2(\ell)}$ of the ensemble forecast $\boldsymbol{f}^{(\ell)}$ initialized at this time point for each forecast horizon $\ell = 1, 2, \ldots, L$.
2 For each initialization time τ in the historical dataset, compute similarity

$$\Delta^{\tau} := \sqrt{\sum_{\ell=1}^{L}(\bar{f}^{(\ell)} - \bar{f}_{\tau}^{(\ell)})^2 + \frac{1}{L}\sum_{\ell=1}^{L}(S^{2(\ell)} - S_{\tau}^{2(\ell)})^2},$$

where $\bar{f}_{\tau}^{(\ell)}$ and $S_{\tau}^{2(\ell)}$ respectively denote the mean and variance of the ensemble forecast $\boldsymbol{f}_{\tau}^{(\ell)}$ with lead time ℓ initialized at time point τ.
3 Chose initialization times $\tau_1, \tau_2, \ldots, \tau_K$ resulting in the highest similarity to the actual forecasts; that is, where $\Delta^{\tau_1}, \Delta^{\tau_2}, \ldots, \Delta^{\tau_K}$ are the smallest among all similarities computed in step 2. The dependence template is given by historic observations $y_{\tau_k}^{(\ell)}$ corresponding to ensemble forecasts $\boldsymbol{f}_{\tau_k}^{(\ell)}$, $k = 1, 2, \ldots, K$, $\ell = 1, 2, \ldots, L$.

3.3 | Forecast evaluation methods

As argued in Gneiting *et al.* (2007), the main goal in probabilistic forecasting is to maximize the sharpness of the predictive distribution subject to calibration. Calibration measures the statistical consistency between the predictions and the corresponding observations, whereas sharpness refers to the concentration of the predictive distribution. Predictive performance is usually quantified with the help of proper scoring rules, which are loss functions $S(F, y)$ assigning numerical values to forecast–observation pairs (F, y). One of the most popular proper scoring rules in the atmospheric sciences assessing simultaneously both calibration and sharpness is the continuous ranked probability score (CRPS; Wilks, 2019, Sect. 9.5.1). For a predictive CDF $F(x)$ and an observation $y \in \mathbb{R}$, the CRPS is defined as

$$\begin{aligned}\mathrm{CRPS}(F, y) &:= \int_{-\infty}^{\infty}(F(x) - I_{\{x \geq y\}})^2 \, dx \\ &= \mathrm{E}|X - y| - \frac{1}{2}\mathrm{E}|X - X'|, \end{aligned} \tag{5}$$

where I_H denotes the indicator of a set H, whereas X and X' are independent random variables with CDF F and finite first moment. The CRPS is a negatively oriented score (i.e., the smaller the better), and this scoring rule serves as a loss function in parameter estimation of normal, truncated normal, and censored generalized extreme-value EMOS models for calibration of T2M, V10, and PPT24 ensemble forecasts respectively. For these distributions the CRPS can be obtained in closed form (for the corresponding formulae see e.g. Jordan *et al.*, 2019), allowing a computationally efficient estimation process.

A multivariate extension of the CRPS is the energy score (ES; Gneiting and Raftery, 2007). Given an L-dimensional CDF F and vector $\boldsymbol{y} = \left(y^{(1)}, y^{(2)}, \ldots, y^{(L)}\right)^{\mathrm{T}}$, the ES is defined as

$$\mathrm{ES}(F, \boldsymbol{y}) := \mathrm{E}\|\boldsymbol{X} - \boldsymbol{y}\| - \frac{1}{2}\mathrm{E}\|\boldsymbol{X} - \boldsymbol{X}'\|, \tag{6}$$

where $\|\cdot\|$ denotes the Euclidean distance and, similar to the univariate case in Equation (5), $\boldsymbol{X}$ and $\boldsymbol{X}'$ are independent random vectors having distribution F. For a forecast ensemble $\mathfrak{f}_1, \mathfrak{f}_2, \ldots, \mathfrak{f}_K$ one should consider the empirical CDF F_K (Gneiting *et al.*, 2008), which reduces Equation (6) to the ensemble energy score

$$\mathrm{ES}(F_K, \boldsymbol{y}) = \frac{1}{K}\sum_{j=1}^{K}\|\mathfrak{f}_j - \boldsymbol{y}\| - \frac{1}{2K^2}\sum_{j=1}^{K}\sum_{k=1}^{K}\|\mathfrak{f}_j - \mathfrak{f}_k\|. \tag{7}$$

Note that the same definition applies for reordered calibrated samples discussed in Section 3.2.

A more recently introduced multivariate proper scoring rule is the (ensemble) variogram score of order p (VS^p; Scheuerer and Hamill, 2015). For an ensemble forecast $\mathbf{f}_k = (f_k^{(1)}, f_k^{(2)}, \ldots, f_k^{(L)})^{\mathrm{T}}$, $k = 1, 2, \ldots, K$, it is defined as

$$\mathrm{VS}^p(F_K, \boldsymbol{y}) = \sum_{i=1}^{L}\sum_{j=1}^{L}\omega_{ij}\left(|y^{(i)} - y^{(j)}|^p - \frac{1}{K}\sum_{k=1}^{K}|f_k^{(i)} - f_k^{(j)}|^p\right)^2,$$

where $\omega_{ij} \geq 0$ is the weight for coordinate pair (i, j). Compared with the ES, the VS^p is more sensitive to the errors in the specification of correlations. Following Lerch *et al.* (2020b), we consider here $p = 1$ and use the notation VS for VS^1.

Further, in the case studies of Section 4, for a given forecast F the improvement in terms of a score S_F with respect to a reference forecast F_{ref} is quantified using the corresponding skill score (Gneiting and Raftery, 2007)

$$S_F^{\mathrm{skill}} := 1 - \frac{\overline{S}_F}{\overline{S}_{F_{\mathrm{ref}}}},$$

where $\overline{S}_F$ and $\overline{S}_{F_{\mathrm{ref}}}$ denote the mean score values over all forecast cases in the verification period for forecasts F and F_{ref} respectively. Thus, besides the ES and the VS, we investigate the energy skill score ESS and the variogram skill score VSS, which are positively oriented (i.e., the larger the better).

Calibration of univariate ensemble forecasts can also be diagnosed with the help of verification rank histograms displaying the ranks of observations with respect to the ensemble forecasts (see e.g. Wilks, 2019, Sect. 9.7.1). For a properly calibrated K-member ensemble the ranks follow a uniform distribution on $\{1, 2, \ldots, K+1\}$, and the deviation from uniformity can be quantified by the reliability index

$$\mathrm{RI} := \sum_{r=1}^{K+1}\left|\rho_r - \frac{1}{K+1}\right|, \tag{8}$$

where ρ_r is the relative frequency of rank r (Delle Monache *et al.*, 2006). There are several options to generalize the verification rank histogram to multivariate ensemble forecasts depending on the definition of the ranks in higher dimensions. Here, we consider the average ranking given by the average of the univariate ranks of the different coordinates. The resulting histogram has properties and interpretation that are similar to the univariate rank histogram (Thorarinsdottir *et al.*, 2016).

Further, we also investigate the mean Euclidean distance EE of the median vectors of the forecasts from the corresponding validating observations, where the multivariate L^1 ensemble median can be obtained with, for example, the algorithm of Vardi and Zhang (2000).

Finally, following Gneiting and Ranjan (2011), statistical significance of score differences between forecasts are assessed with the help of the Diebold–Mariano (DM; Diebold and Mariano, 1995) test, which is able to account for temporal dependencies in the forecast errors. Given a scoring rule S and two competing probabilistic forecasts F and G, the test statistic of the DM test is given by

$$t_N = \sqrt{N}\frac{\overline{S}_F - \overline{S}_G}{\hat{\sigma}_N}, \tag{9}$$

where $\overline{S}_F$ and $\overline{S}_G$ are the mean scores over a test set corresponding to forecasts F and G respectively, and $\hat{\sigma}_N$ is a suitable estimator of the asymptotic standard deviation of the sequence of individual score differences. Under standard regularity conditions, t_N asymptotically follows a standard Gaussian distribution under the null hypothesis of equal predictive performance. Negative values of t_N indicate a better predictive performance of F, whereas G is preferred in the case of positive values of t_N.

4 | RESULTS

The predictive performance of the different multivariate post-processing methods is investigated in three case studies based on global ECMWF T2M, V10, and PPT24 ensemble forecasts. To assess calibration of probabilistic forecasts we use the energy score ES and the variogram score of order 1 (VS), and we also investigate the multivariate rank histograms together with the corresponding reliability indices (Δ), whereas multivariate point forecasts are evaluated with the help of the mean Euclidean error EE. Note that, in addition, we studied the regularized Dawid–Sebastiani scores (RDSS; Wilks, 2020) of the raw ensemble and multivariate forecasts addressing temporal dependence. However, for our datasets, the RDSS results in a ranking of the various post-processing methods that is similar to the ES. In the interest of brevity, we restrict our attention to the ES and VS here, and we provide more detailed results for the RDSS in the Supporting Information.

Though our focus is on the multivariate predictive performance, the first step in each of the multivariate methods described in Section 3.2 is the independent calibration of ensemble forecasts with different forecast horizons. Hence, we start with a brief description of the details of univariate post-processing, followed by the results for the multivariate models.

4.1 | Univariate post-processing

For a given initialization time, the corresponding ensemble forecasts with different forecast horizons are calibrated with the help of the EMOS approaches of Section 3.1. Estimates of the EMOS model parameters minimize the mean CRPS of the predictive distribution over locally selected training data using a rolling training period. This means that, for each observation station, the training dataset consists of ensemble forecasts and corresponding validating observations of the given station for the preceding n calendar days. To avoid numerical problems in the minimization of the loss function, local models require long training periods; here, we consider the optimal training period lengths determined by Hemri *et al.* (2014).

4.1.1 | Temperature

The ℓ-day ahead ECMWF T2M ensemble forecasts ($\ell = 1, 2, \ldots 10$) are calibrated with the help of the normal EMOS model specifued by Equation (2) with a training period length of 720 days. For verification we consider the time interval between January 1, 2004, and March 20, 2014 (3,732 calendar days). Figure 1a shows the boxplots of the continuous ranked probability skill score (CRPSS) of the EMOS models over the verification period with respect to the raw ensemble forecasts. Compared with the raw ECMWF T2M forecasts, univariate post-processing substantially decreases the mean CRPS for the vast majority of stations for all lead times; however, the gain decreases with the increase of the forecast horizon (see also Feldmann *et al.*, 2019).

4.1.2 | Wind speed

The optimal training period length for the truncated normal EMOS model of Equation (3) for post-processing ECMWF V10 ensemble forecasts is 365 calendar days. This allows an almost 1-year longer verification period than in the case of T2M; however, to keep the results consistent,

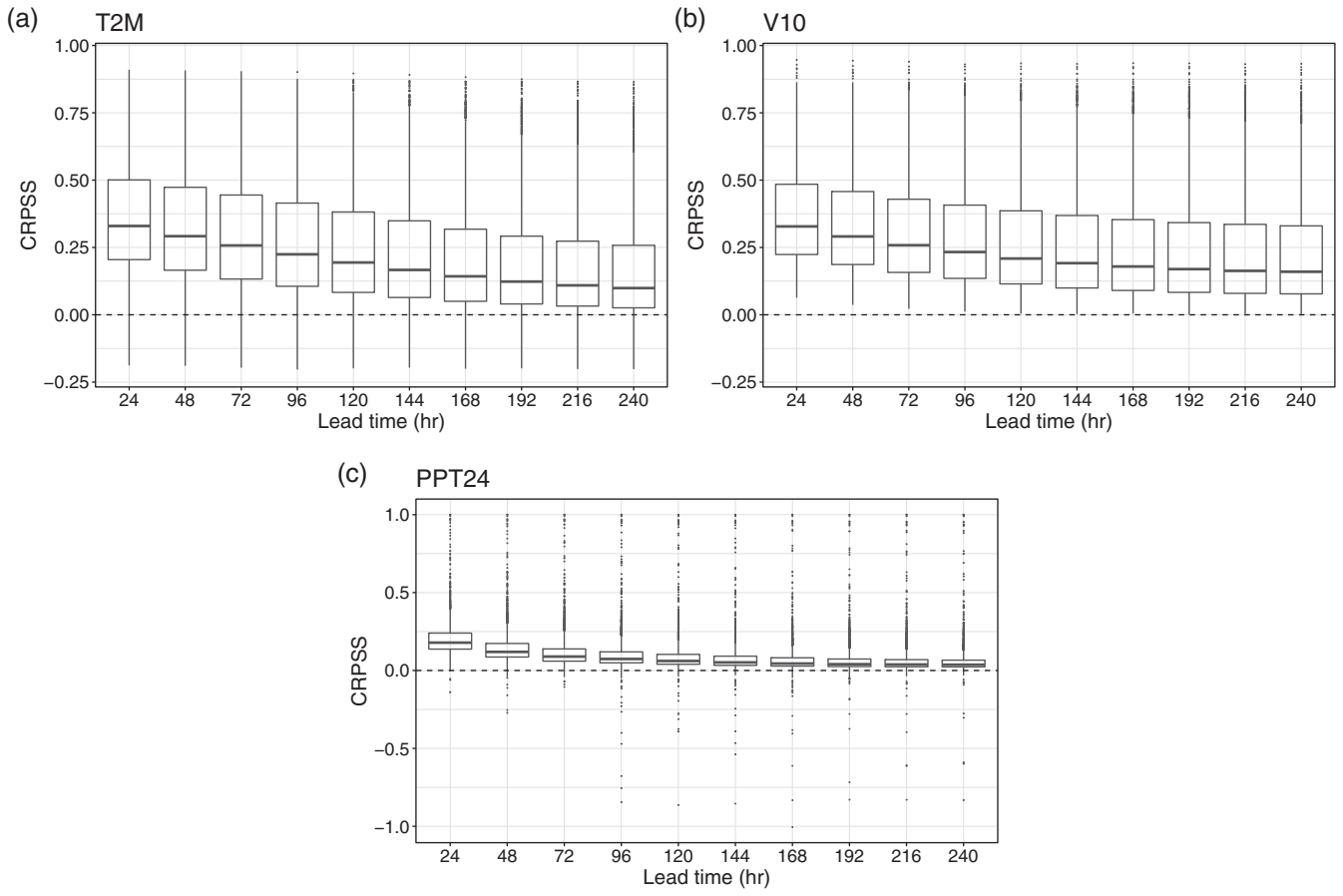

FIGURE 1 Boxplots of the continuous ranked probability skill score (CRPSS) of the ensemble model output statistics models for (a) 2 m temperature (T2M), (b) 10 m wind speed (V10), and (c) 24 hr precipitation accumulation (PPT24) over the verification period with respect to the corresponding raw ensemble forecasts

we verify both models on the same time interval starting at January 1, 2004. The forecast skill of the EMOS models for various lead times in terms of the mean CRPS over the verification period is depicted in Figure 1b. Compared with the T2M ensemble forecasts, EMOS post-processing results in a positive skill score for all stations considered, and the improvement in the mean CRPS is a bit higher, especially for longer lead times.

4.1.3 | Precipitation

Precipitation records usually contain a large number of zero observations; hence, reliable parameter estimation in EMOS modelling requires much longer training periods than for temperature or wind speed. Following Hemri *et al.* (2014), to calibrate ECMWF PPT24 forecasts with lead times of $1, 2, \ldots, 10$ days, we make use of the EMOS model of Equation (4) with a 1,816-day rolling training period, leaving the time interval from January 1, 2007, to March 20, 2014, for model verification. Again, in Figure 1c we provide the CRPSS values for various forecast horizons, where, similar to the other two weather quantities, the reference forecast is the raw ensemble. The general behaviour of the skill scores is the same as before. For almost all stations the EMOS forecasts outperform the raw PPT24 ensemble for all lead times, and the improvement decreases with the increase of the forecast horizon. However, compared with T2M and V10, the boxplots of the CRPSS values of the EMOS models for PPT24 have much shorter interquartile ranges and display far more outliers.

4.2 | Multivariate performance

We now continue with the comparison of the forecast skill of the multivariate approaches described in Section 3.2 using calibrated samples of size 52, which is the size of the raw ECMWF ensemble (see Section 2). In methods requiring historical data for providing the dependence template (estimation of the autocorrelation matrix for dECC; observation trajectories for the SSh and for its variants), we consider forecast-observation pairs from the rolling training window applied in univariate calibration. In the case of T2M, in standard SSh variants seasonality is addressed by considering historical data only within 30 calendar days before and after the actual forecast date. For T2M and V10 ensemble forecasts, in mdSSh we select dependence template from observation trajectories of the training data set where at least $m = 6$ from the $L = 10$ observations falls into the corresponding central prediction interval, whereas in the case of PPT24, due to the high computational costs (see also Scheuerer *et al.*, 2017), this method is excluded from the analysis. Further, in the case of T2M, tests are performed also with SSh approaches based on the whole available (and in this way extending) past; however, as the corresponding results of Section 4.2.1 show, in most cases these forecasts do not significantly outperform the corresponding ones using shorter rolling periods.

In the following analysis, the ECC-Q forecast is used as a reference for the computation of skill scores and in DM tests for equal predictive performance (see Section 3.3). Further, in the interest of improving the visual presentation of the results, in the boxplots presented in Sections 4.2.1–4.2.3 the extreme values (further than three times the interquartile range from the box) are not indicated.

4.2.1 | Temperature

In Figure 2, the various multivariate post-processing approaches and the raw T2M forecasts are compared in terms of the mean ES given by Equation (7) of the different observation stations over the verification period (January 1, 2004–March 20, 2014). According to Figure 2a, each calibration method improves the ES of the raw ensemble; however, the differences between the various approaches are hardly visible. A better insight can be obtained from Figure 2b displaying the ESS values with respect to the reference ECC-Q forecast. The raw ensemble forecasts resulting in very low skill scores are excluded here. All post-processing methods accounting for multivariate dependencies provide clear improvements in terms of the ES over the independent EMOS models: ECC-S shows the best predictive performance, closely followed by ECC-Q and dECC. Perhaps surprisingly, even the advanced variants of the SSh (mdSSh and simSSh) fail to outperform the reference ECC-Q method, and SSh-S is not behind mdSSh in skill. Finally, according to the results of the DM tests for equal predictive performance (Figure 2c), the differences in ES from the reference model are significant for all approaches but dECC.

A similar ranking of the multivariate post-processing methods can be observed in Figure 3, though the performance in terms of the VS of the standard SSh variants are closer to each other than in terms of the ES. ECC-Q results in the lowest mean VS; however, its advantage over dECC, ECC-S, and simSSh is often not significant (see Figure 3c). Further, in contrast to the ES, none of the independent EMOS forecasts outperforms the raw ECMWF ensemble in terms of VS.

As mentioned, SSh approaches with dependence templates selected from the whole available past meaning 721 days for the first and 4443 days for the last day of the verification period have also been tested. As before, in

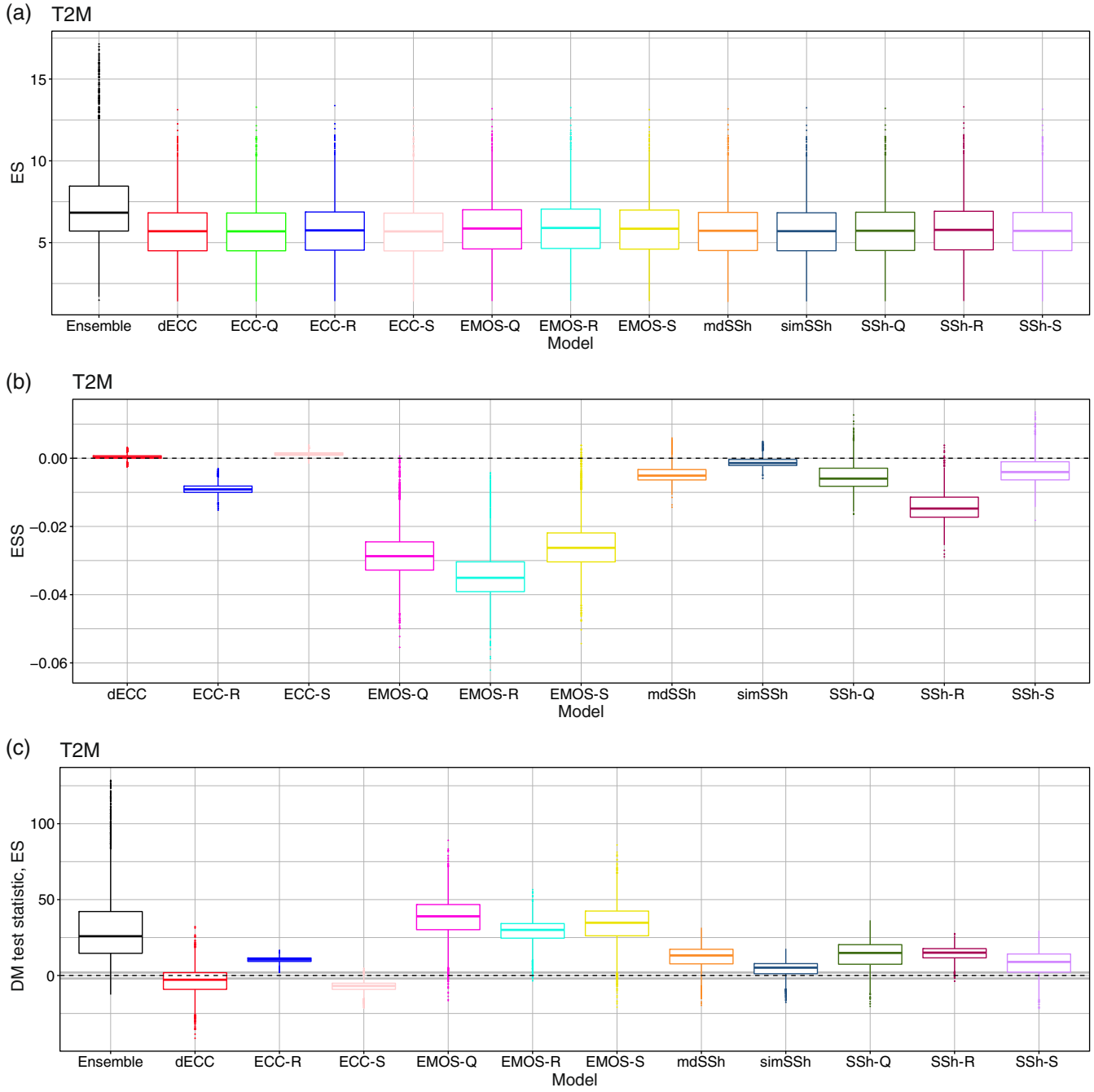

FIGURE 2 Boxplots of the (a) mean energy score (ES) over the verification period of the calibrated and raw 2 m temperature (T2M) forecasts, (b) energy skill score (ESS) with respect to the ECC-Q approach, and (c) Diebold–Mariano (DM) test statistic investigating the significance of the difference from the reference ECC-Q method. Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5% level of significance. ECC: ensemble copula coupling; EMOS: ensemble model output statistics; SSh: Schaake shuffle; d: dual; md: minimum divergence; sim: similarity-based; Q: equidistant quantiles; R: random sample; S: stratified sample [Colour figure can be viewed at wileyonlinelibrary.com]

the case of the simple SSh variants historical data only within 30 calendar days around the actual forecast date have been used. However, according to the boxplots of Figure 4, the use of the whole available past results in just minor improvements in ES and even smaller in VS, which are not significant. Hence, in the following analysis including case studies with V10 and PPT24 forecasts as well (Sections 4.2.2 and 4.2.3, respectively), we proceed with SSh approaches based on historical observations from the rolling training period used in EMOS modelling. This has the potential advantage of accounting more rapidly to substantial changes in the NWP model

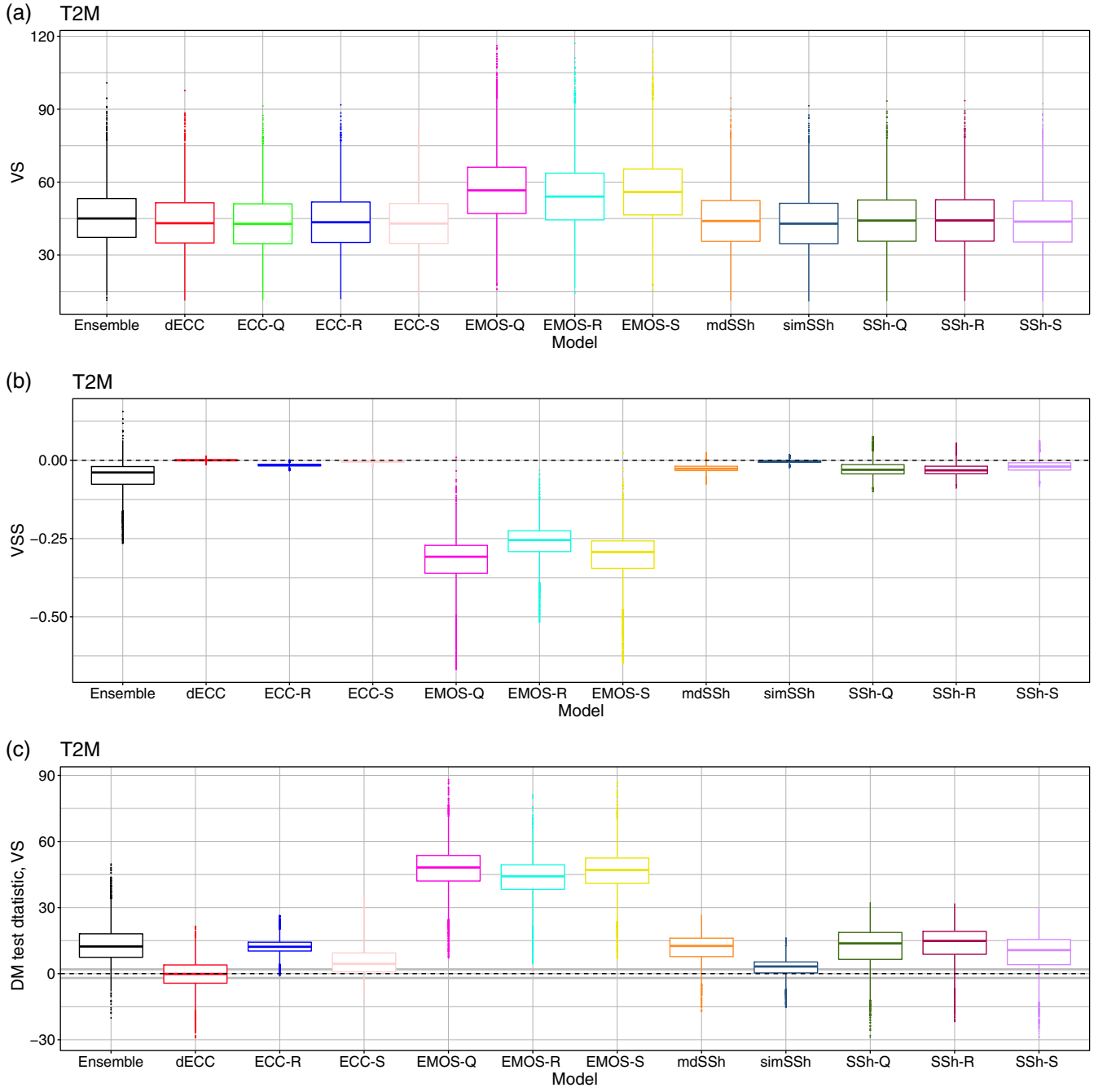

FIGURE 3 Boxplots of the (a) mean variogram score (VS) over the verification period of the calibrated and raw 2 m temperature (T2M) forecasts, (b) variogram skill score (VSS) with respect to the ECC-Q approach, and (c) Diebold–Mariano (DM) test statistic investigating the significance of the difference from the reference ECC-Q method. Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5% level of significance. ECC: ensemble copula coupling; EMOS: ensemble model output statistics; SSh: Schaake shuffle; d: dual; md: minimum divergence; sim: similarity-based; Q: equidistant quantiles; R: random sample; S: stratified sample [Colour figure can be viewed at wileyonlinelibrary.com]

setup that may influence the forecast error characteristics. For related considerations in the context of univariate post-processing, see Lang *et al.* (2020).

Boxplots of the reliability indices corresponding to average ranks over the verification period of the calibrated and raw T2M forecasts are displayed in Figure 5. ECC-S and all three independent predictions (EMOS-Q, EMOS-R, EMOS-S) underperform the raw ensemble by a wide margin, which is a consequence of the highly overdispersive character of these forecasts resulting in hump-shaped rank histograms (not shown). SSh-Q, SSh-R, and SSh-S still show some overdispersion, whereas the

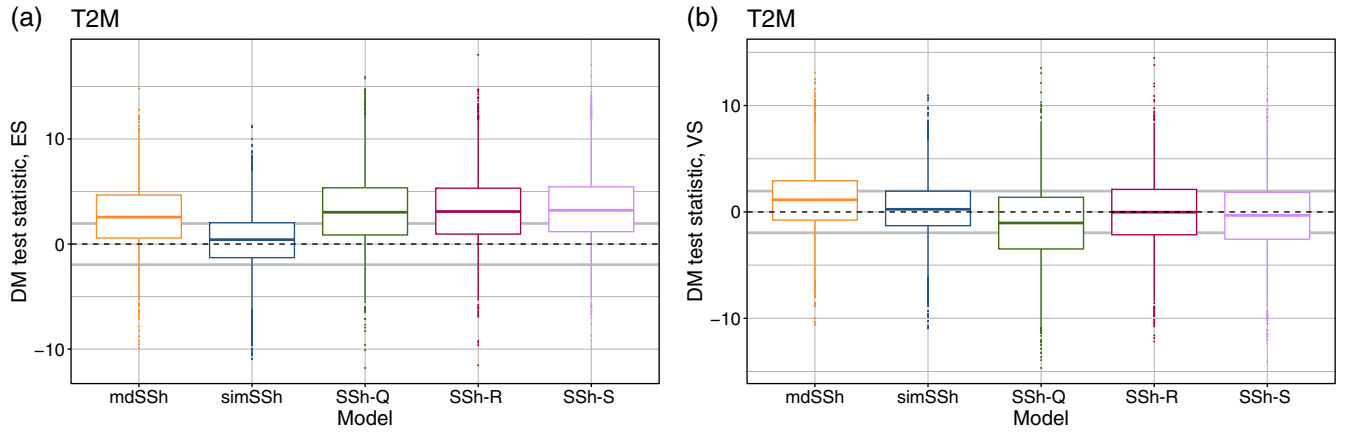

FIGURE 4 Boxplots of the Diebold–Mariano (DM) test statistic investigating the significance of the difference in the (a) mean energy score (ES) and (b) mean variogram score (VS) of Schaake shuffle (SSh) methods based on historical observations from the rolling training period used in ensemble model output statistics modelling with respect to the corresponding forecasts selecting dependence templates from the whole available past. Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5% level of significance. md: minimum divergence; sim: similarity-based; Q: equidistant quantiles; R: random sample; S: stratified sample [Colour figure can be viewed at wileyonlinelibrary.com]

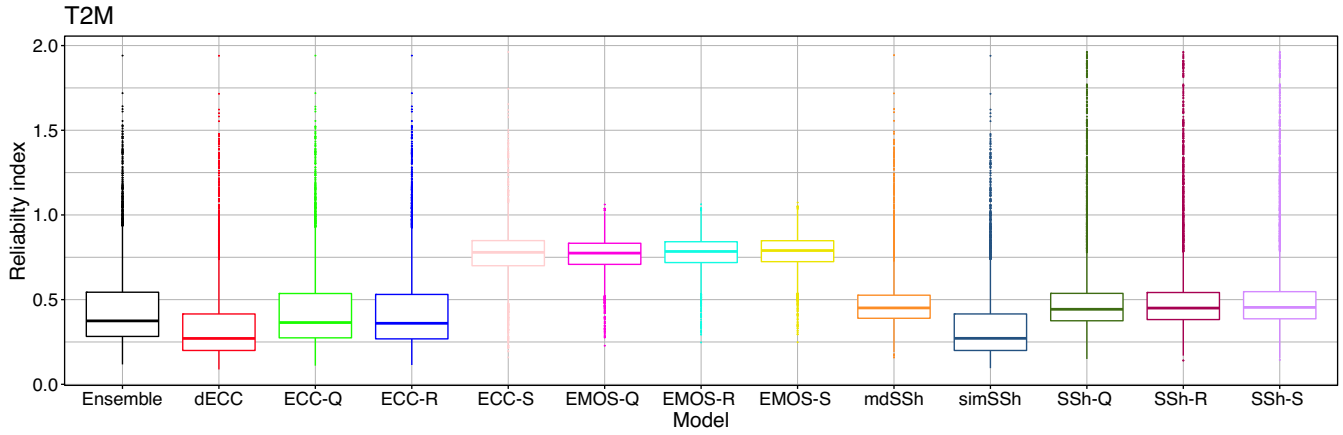

FIGURE 5 Boxplots of the reliability indices corresponding to average ranks over the verification period of the calibrated and raw 2 m temperature (T2M) forecasts. ECC: ensemble copula coupling; EMOS: ensemble model output statistics; SSh: Schaake shuffle; d: dual; md: minimum divergence; sim: similarity-based; Q: equidistant quantiles; R: random sample; S: stratified sample [Colour figure can be viewed at wileyonlinelibrary.com]

other forecasts, including the raw ensemble, are underdispersive. The average ranks of the dECC and simSSh are the closest to the uniform distribution (the corresponding mean/median reliability indices are 0.357/0.271 for both forecasts), followed by the ECC-R (0.438/0.360) and ECC-Q (0.442/0.365); however, the rank histograms of the latter two forecasts (not shown) are just slightly less underdispersive than that of the ECMWF ensemble (0.451/0.375), whereas the histogram of the mdSSh is even more U-shaped (0.493/0.451).

Finally, Figure 6 shows the boxplots of the DM test statistic investigating the significance of the difference from the reference ECC-Q method in terms of the mean EE of the L^1 median vectors. Compared with the raw ensemble, post-processing substantially improves the accuracy of the L^1 median forecast and the empirical copula-based models clearly outperform the independent approaches. The lowest mean EE correspond to the SSh-Q and SSh-S, followed by the mdSSh; however, the differences between these methods and the dECC, ECC-Q, ECC-S and simSSh are not significant.

Note that our ranking of the different post-processing methods differs from recent results of Heinrich *et al.* (2021) on multivariate calibration of sea-surface temperature forecasts, where SSh significantly outperforms ECC. However, this disagreement might be explained by the

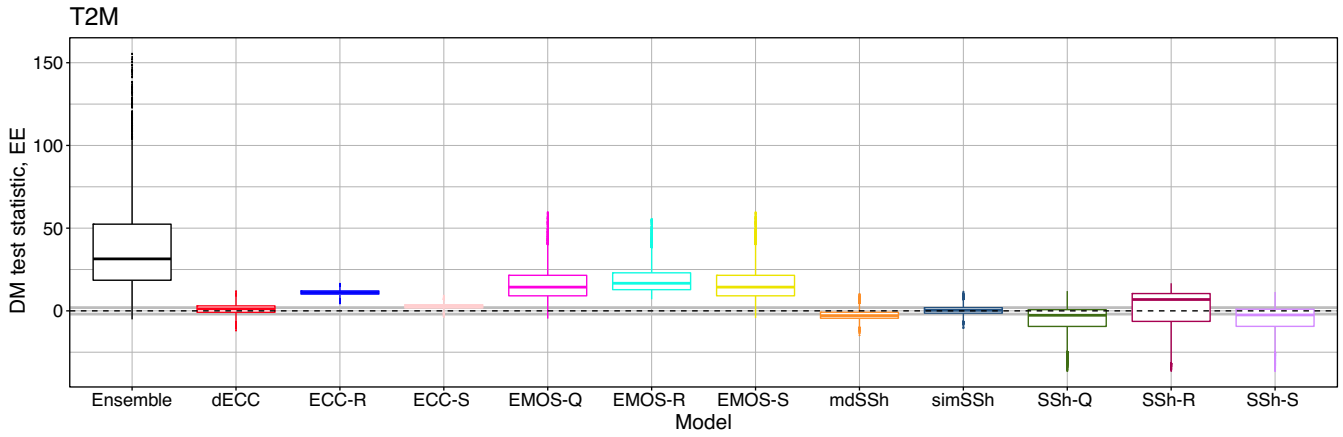

FIGURE 6 Boxplots of the Diebold–Mariano (DM) test statistic investigating the significance of the difference from the reference ECC-Q method in terms of the mean Euclidean error (EE) of the L^1 median vectors of calibrated and raw 2 m temperature (T2M) forecasts. Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5% level of significance. ECC: ensemble copula coupling; EMOS: ensemble model output statistics; SSh: Schaake shuffle; d: dual; md: minimum divergence; sim: similarity-based; Q: equidistant quantiles; R: random sample; S: stratified sample [Colour figure can be viewed at wileyonlinelibrary.com]

difference in the weather quantity studied and the specific properties of spatial dependencies of gridded sea-surface forecasts compared with the temporal dependencies considered here.

4.2.2 | Wind speed

The predictive performance of the post-processed and raw V10 ensemble forecast vectors in terms of the mean ensemble ES over the verification period can be investigated with the help of Figure 7. Compared with the raw ensemble, each of the post-processing approaches investigated substantially reduces the mean ES (see Figure 7a). According to the skill scores with respect to the ECC-Q model depicted in Figure 7b (the raw ensemble is again excluded), the best-performing methods are ECC-S and SSh-S, followed by ECC-Q, mdSSh, simSSh, and SSh-Q. However, as the DM test statistics provided in Figure 7c indicate, the forecast skill of the latter four approaches does not differ significantly. Finally, in contrast to the case of T2M ensemble forecast vectors, all empirical copula-based methods outperform the independent EMOS-Q, EMOS-R, and EMOS-S forecasts by a wide margin.

Figure 8, summarizing the results for the mean VS, is similar to Figure 3, in the sense that independent post-processing of raw V10 forecasts with different lead times (EMOS-Q, EMOS-R, EMOS-S) increases the score values. In general, the predictive performance of the empirical copula-based methods is almost identical (see Figure 8b), with only ECC-R, ECC-S, and SSh-R performing slightly worse. However, according to the results of the DM tests given in Figure 8c, the difference in skill of these three methods from the reference ECC-Q approach is significant.

In contrast to temperature, in the case of wind speed, any form of post-processing improves the multivariate calibration, in the sense that the corresponding average rank histogram is closer to the uniform distribution than the rank histogram of the raw V10 ensemble. This improvement is quantified in the reliability indices displayed in Figure 9, where the highest values again correspond to the independently calibrated EMOS-Q, EMOS-R, and EMOS-S forecasts. These approaches result in hump-shaped rank histograms (not shown) indicating overdispersion, the corresponding mean/median RI values are 0.936/0.957, 0.939/0.961, and 0.947/0.968 respectively. All other forecasts (including the raw ensemble) are underdispersive, where the least U-shaped rank histograms correspond to the three versions of the standard SSh, followed by the simSSh. The mean/median reliability indices of these forecasts are 0.639/0.441 (SSh-Q), 0.637/0.439 (SSh-R), 0.637/0.438 (SSh-S), and 0.671/0.470 (simSSh), which are substantially better than the corresponding scores of 1.029/0.952 of the raw ensemble vector.

Finally, Figure 10, showing boxplots of the DM test statistic investigating the significance of the difference from the reference ECC-Q method in terms of the mean EE of the L^1 median vectors, is almost identical to Figure 6. All post-processing methods result in more accurate median forecasts than the raw ensemble, and all empirical copula-based methods but ECC-R and SSh-R perform almost identically well.

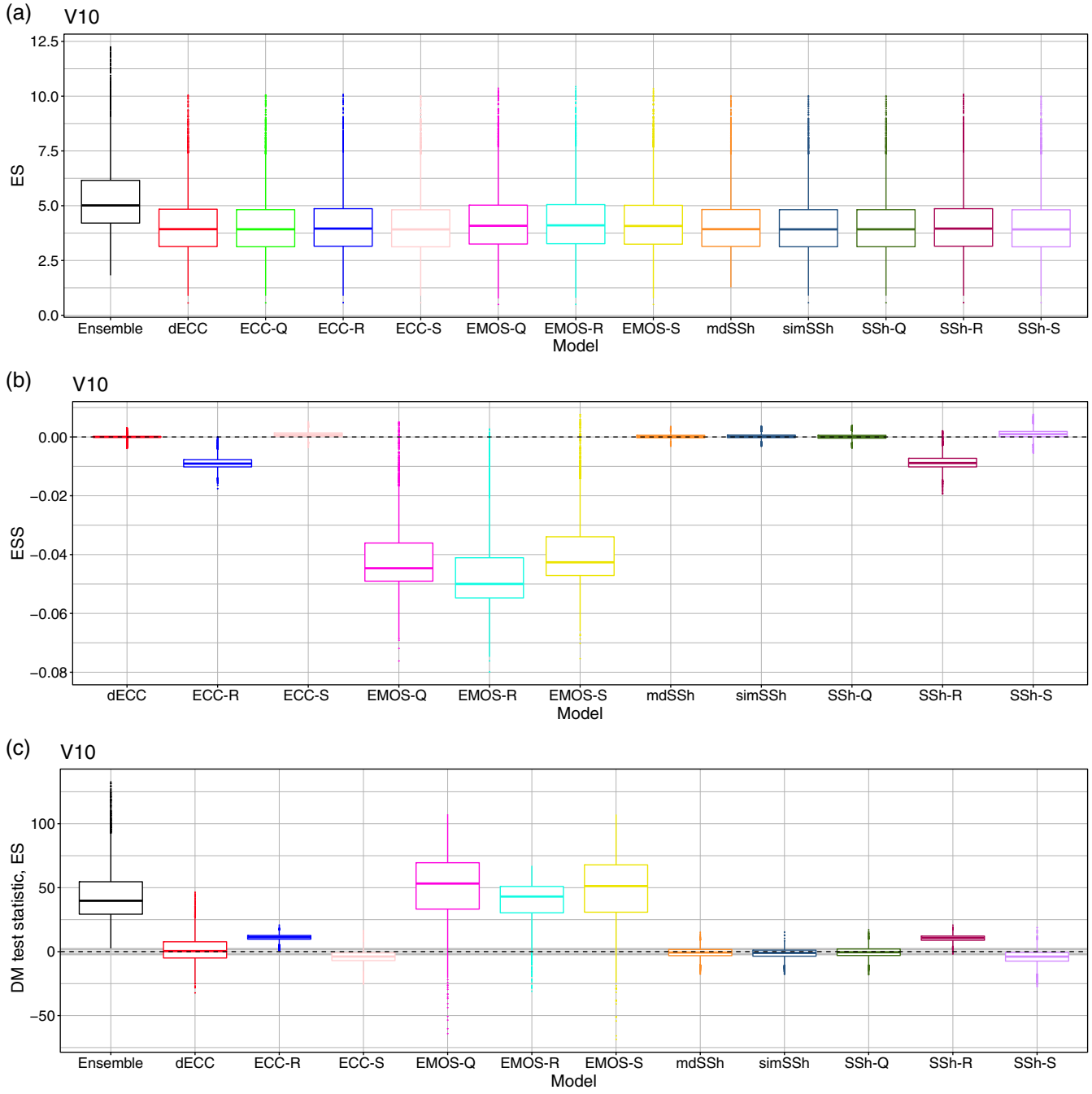

FIGURE 7 Boxplots of the (a) mean energy score (ES) over the verification period of the calibrated and raw 10 m wind speed (V10) forecasts, (b) energy skill score (ESS) with respect to the ECC-Q approach, and (c) Diebold–Mariano (DM) test statistic investigating the significance of the difference from the reference ECC-Q method. Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5% level of significance. ECC: ensemble copula coupling; EMOS: ensemble model output statistics; SSh: Schaake shuffle; d: dual; md: minimum divergence; sim: similarity-based; Q: equidistant quantiles; R: random sample; S: stratified sample [Colour figure can be viewed at wileyonlinelibrary.com]

4.2.3 | Precipitation accumulation

Figure 11a, displaying the boxplots of the mean ES over the verification period of the calibrated and raw PPT24 forecasts, does not reveal a clearly visible difference between the various predictions. The raw ensemble and the independent EMOS-Q, EMOS-R, and EMOS-S methods seem to be slightly behind the other forecasts, which is confirmed by the skill scores of Figure 11b. From the empirical copula-based approaches, ECC-S again shows the best forecast skill, followed by ECC-Q, dECC, simSSh, and SSh-S. However, as indicated by the values of the DM

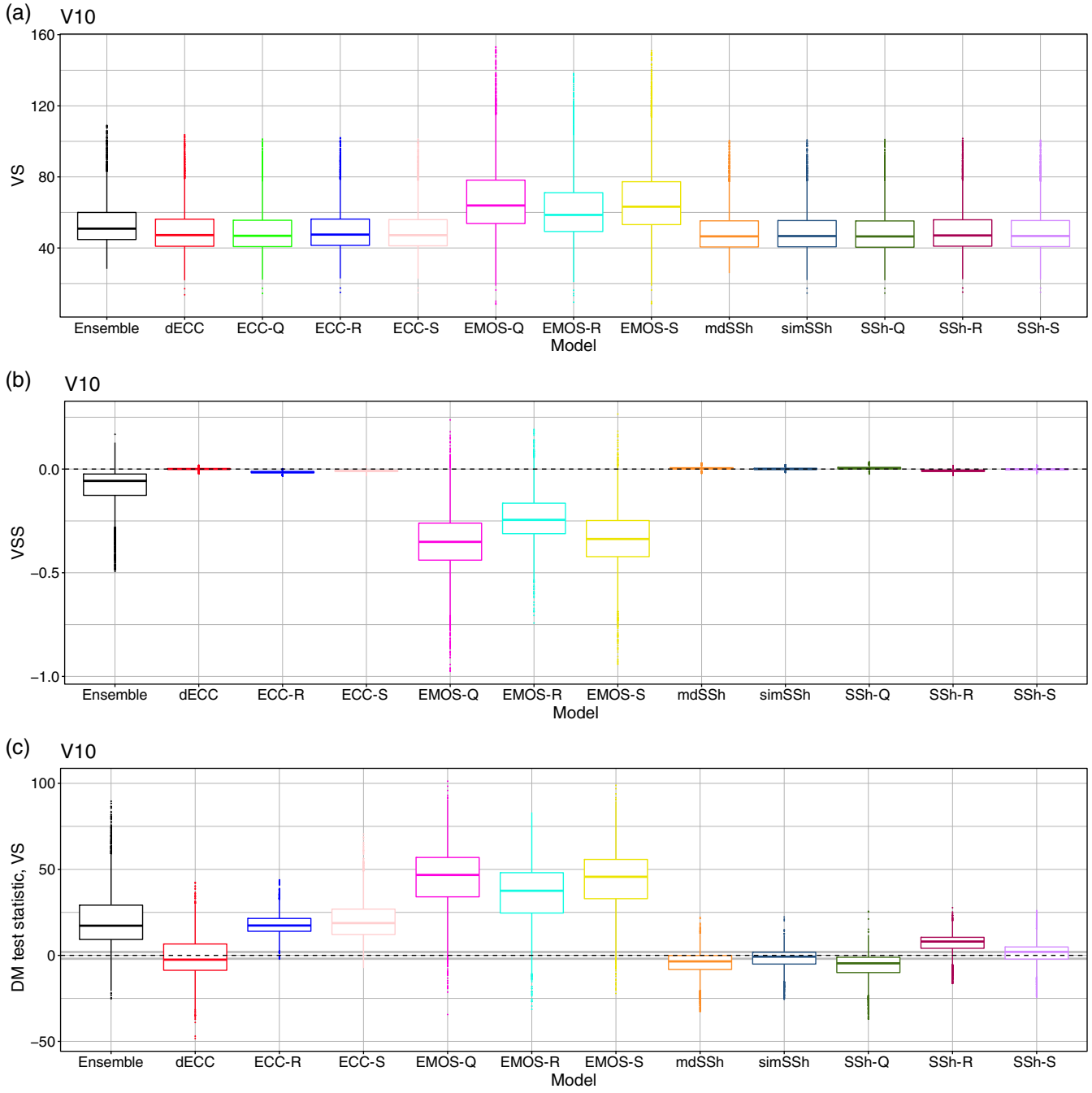

FIGURE 8 Boxplots of the (a) mean variogram skill (VS) over the verification period of the calibrated and raw 10 m wind speed (V10) forecasts, (b) variogram skill score (VSS) with respect to the ECC-Q approach, and (c) Diebold–Mariano (DM) test statistic investigating the significance of the difference from the reference ECC-Q method. Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5% level of significance. ECC: ensemble copula coupling; EMOS: ensemble model output statistics; SSh: Schaake shuffle; d: dual; md: minimum divergence; sim: similarity-based; Q: equidistant quantiles; R: random sample; S: stratified sample [Colour figure can be viewed at wileyonlinelibrary.com]

test statistics summarized in Figure 11c, the differences between the latter four forecasts in terms of the ES are not significant.

In contrast to the other two weather quantities, Figures 11 and 12 result in the same conclusions; that is, the rankings of the forecasts in terms of the ES and the VS are identical. Further, investigating the boxplots of Figure 13, one can again observe the positive effect of post-processing on multivariate calibration quantified in lower reliability indices. The average ranks of dECC and simSSh fit the uniform distribution best, followed by the independent EMOS-Q, EMOS-R, and EMOS-S

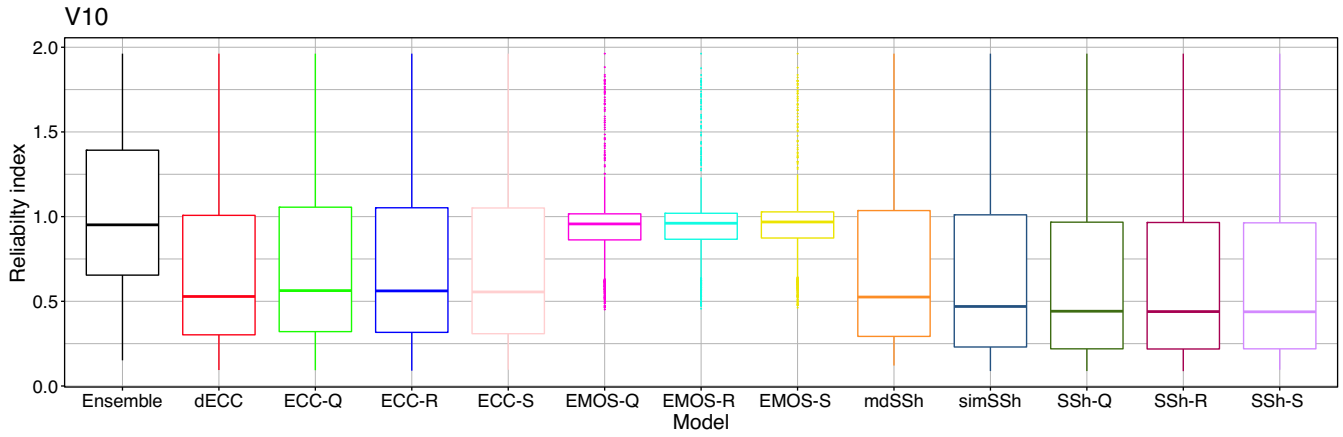

FIGURE 9 Boxplots of the reliability indices corresponding to average ranks over the verification period of the calibrated and raw 10 m wind speed (V10) forecasts. ECC: ensemble copula coupling; EMOS: ensemble model output statistics; SSh: Schaake shuffle; d: dual; md: minimum divergence; sim: similarity-based; Q: equidistant quantiles; R: random sample; S: stratified sample [Colour figure can be viewed at wileyonlinelibrary.com]

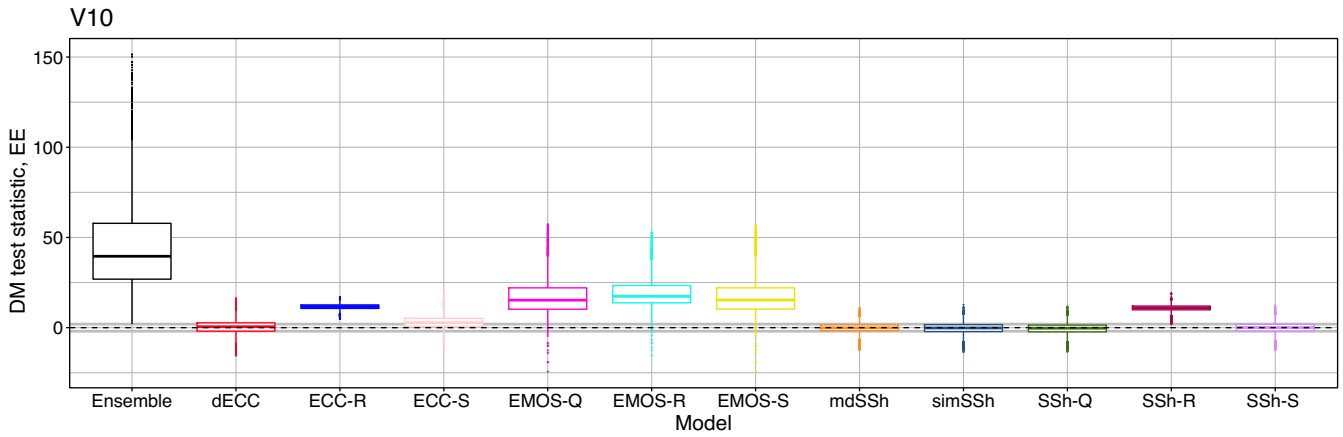

FIGURE 10 Boxplots of the Diebold–Mariano (DM) test statistic investigating the significance of the difference from the reference ECC-Q method in terms of the mean Euclidean error (EE) of the L^1 median vectors of calibrated and raw 10 m wind speed (V10) forecasts. Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5% level of significance. ECC: ensemble copula coupling; EMOS: ensemble model output statistics; SSh: Schaake shuffle; d: dual; md: minimum divergence; sim: similarity-based; Q: equidistant quantiles; R: random sample; S: stratified sample [Colour figure can be viewed at wileyonlinelibrary.com]

approaches. Note that, despite their overdispersive character resulting in hump-shaped rank histograms (not shown), these independent forecasts outperform all versions of the standard SSh and ECC, which are highly underdispersive.

Finally, Figure 14, showing the boxplots of the DM test statistic investigating the significance of the difference from the reference ECC-Q method in terms of the mean EE of the L^1 median vectors, suggests the same ranking of the various post-processing approaches as Figures 11 and 12. However, in contrast to the ES and VS, with regard to this score even the independent EMOS-Q, EMOS-R, and EMOS-S methods outperform the raw ensemble.

5 | DISCUSSION AND CONCLUSIONS

We compared a wide variety of state-of-the-art methods for multivariate ensemble post-processing with a focus on dependencies over lead times from 1 to 10 days, using three case studies of global ECMWF ensemble forecasts of temperature, wind speed, and precipitation accumulation. Across all of the three settings, all multivariate post-processing methods substantially improve all aspects of multivariate forecast quality investigated over both the raw ensemble predictions and a simple application of univariate post-processing models without accounting for multivariate dependencies. Among the basic ECC

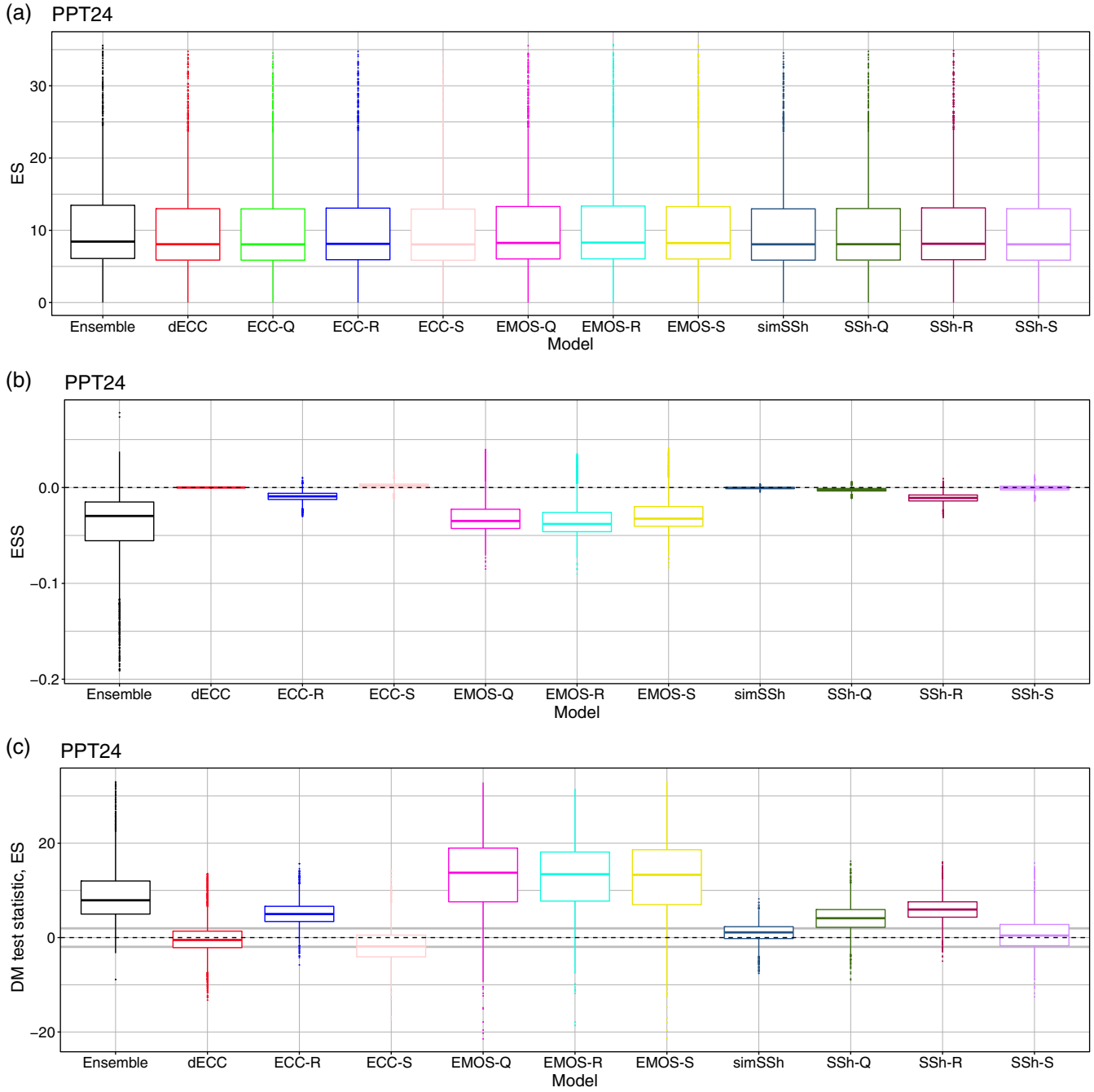

FIGURE 11 Boxplots of the (a) mean energy score (ES) over the verification period of the calibrated and raw 24 hr precipitation accumulation (PPT24) forecasts, (b) energy skill score (ESS) with respect to the ECC-Q approach, and (c) Diebold–Mariano (DM) test statistic investigating the significance of the difference from the reference ECC-Q method. Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5% level of significance. ECC: ensemble copula coupling; EMOS: ensemble model output statistics; SSh: Schaake shuffle; d: dual; md: minimum divergence; sim: similarity-based; Q: equidistant quantiles; R: random sample; S: stratified sample [Colour figure can be viewed at wileyonlinelibrary.com]

and SSh variants, the different sampling strategies only showed minor differences in the predictive performance. In particular, random sampling (ECC-R and SSh-R) generally performed worse than quantile-based (ECC-Q and SSh-Q) or stratified sampling (ECC-S and SSh-S), whereas no significant differences could be detected between the latter two approaches. Comparing the more advanced dECC, mdSSh, and simSSh approaches with their basic counterparts, we generally did not observe any benefits of these more complex methods, and we did not find a single case where they significantly outperform ECC-Q.

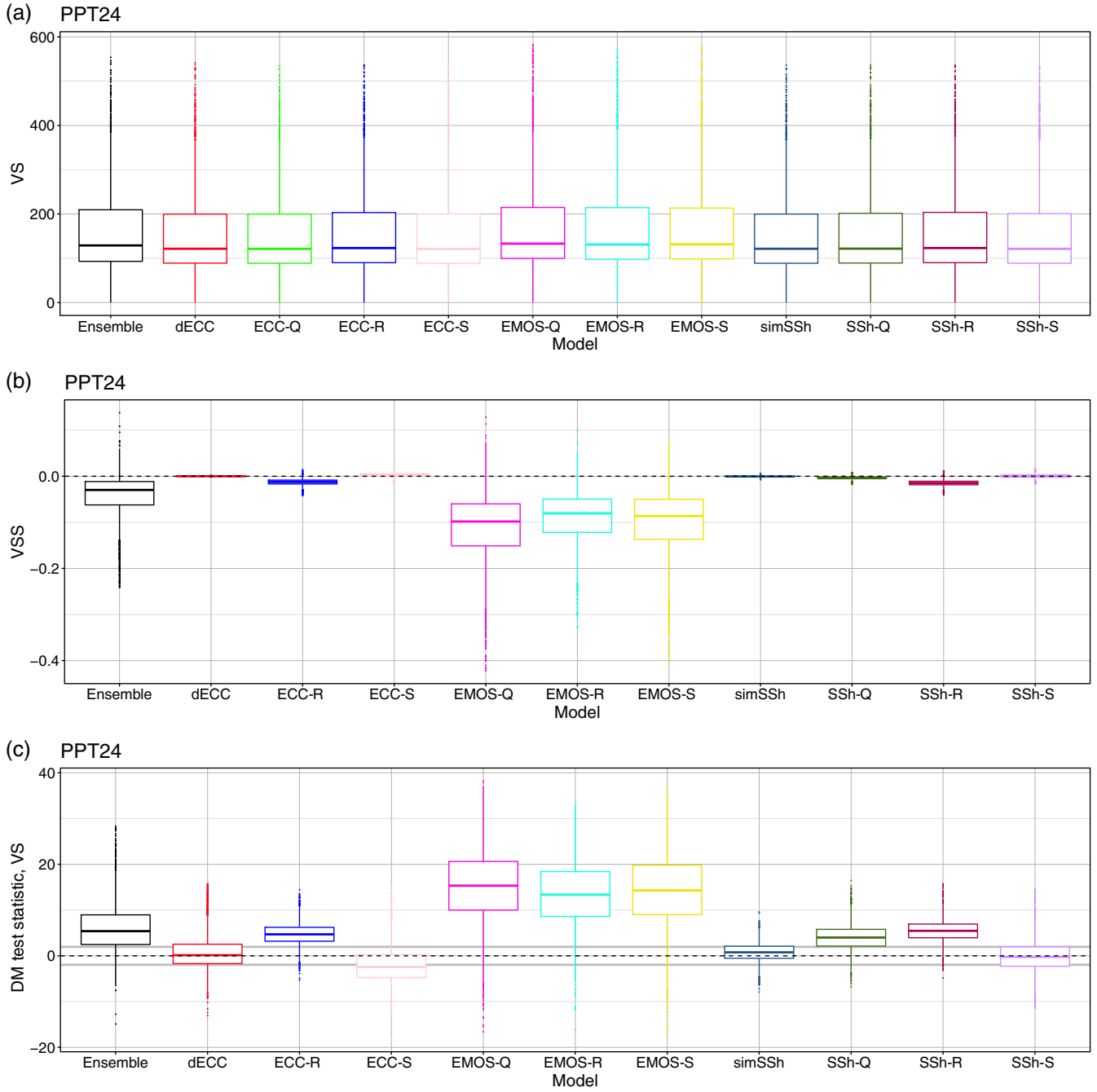

FIGURE 12 Boxplots of the (a) mean variogram skill (VS) over the verification period of the calibrated and raw 24 hr precipitation accumulation (PPT24) forecasts, (b) variogram skill score (VSS) with respect to the ECC-Q approach, and (c) Diebold–Mariano (DM) test statistic investigating the significance of the difference from the reference ECC-Q method (c). Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5% level of significance. ECC: ensemble copula coupling; EMOS: ensemble model output statistics; SSh: Schaake shuffle; d: dual; md: minimum divergence; sim: similarity-based; Q: equidistant quantiles; R: random sample; S: stratified sample [Colour figure can be viewed at wileyonlinelibrary.com]

In a nutshell, our overall findings indicate that there are generally only minor differences in the predictive performance of the various multivariate post-processing methods and that the widely used ECC-Q approach constitutes a powerful benchmark method. Its straightforward applicability and very low computational costs make it a natural first choice to apply in future comparative studies of multivariate post-processing methods. Subject to differences in the set of methods investigated, our findings are in line with the results of the simulation studies performed in Lerch *et al.* (2020a), who also did not observe a single consistently best method across all potential misspecifications

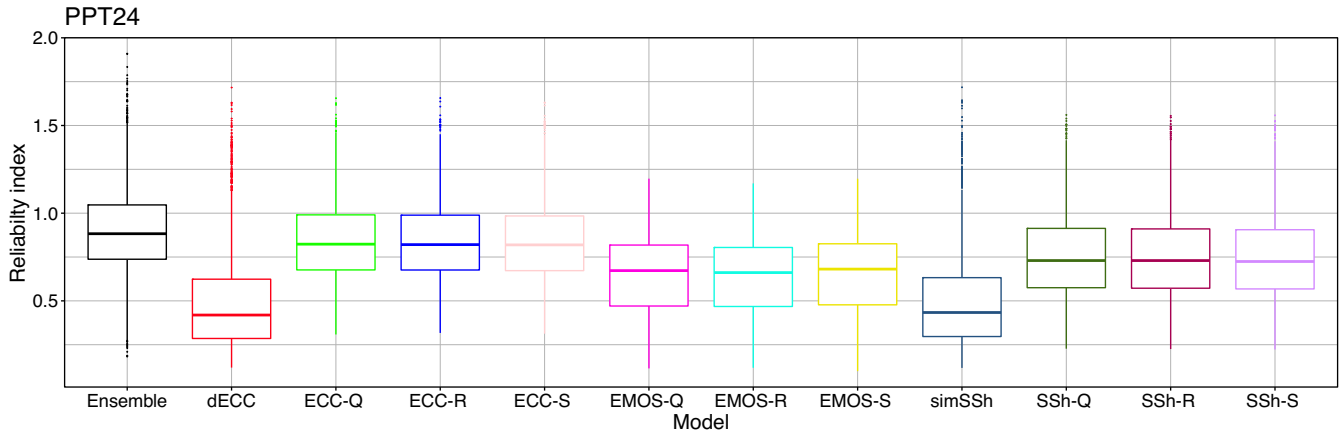

FIGURE 13 Boxplots of the reliability indices corresponding to average ranks over the verification period of the calibrated and raw 24 hr precipitation accumulation (PPT24) forecasts. ECC: ensemble copula coupling; EMOS: ensemble model output statistics; SSh: Schaake shuffle; d: dual; md: minimum divergence; sim: similarity-based; Q: equidistant quantiles; R: random sample; S: stratified sample [Colour figure can be viewed at wileyonlinelibrary.com]

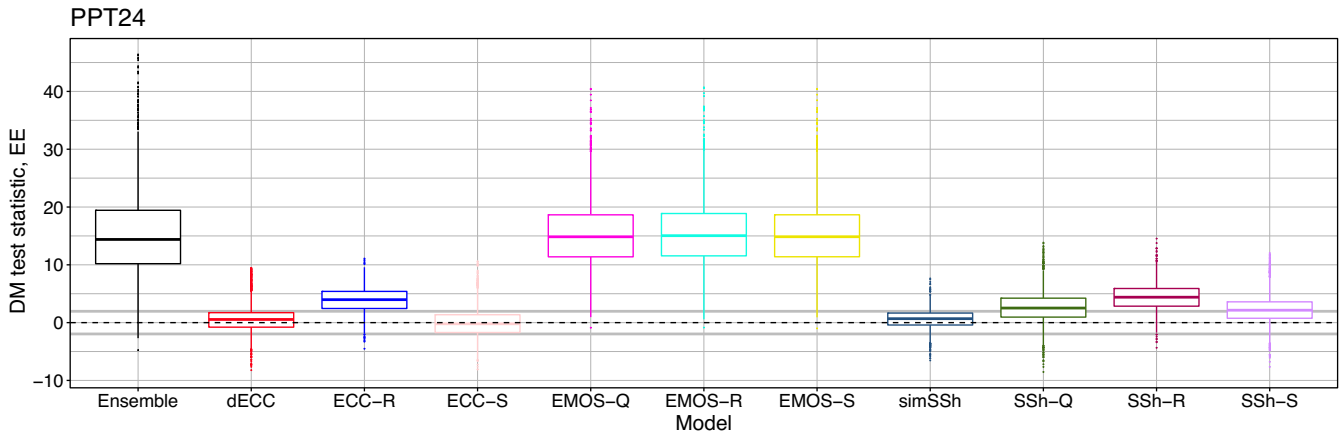

FIGURE 14 Boxplots of the Diebold–Mariano (DM) test statistic investigating the significance of the difference from the reference ECC-Q method in terms of the mean Euclidean error (EE) of the L^1 median vectors of calibrated and raw 24 hr precipitation accumulation (PPT24) forecasts. Grey lines indicate the acceptance region of the two-tailed DM test for equal predictive performance at a 5% level of significance. ECC: ensemble copula coupling; EMOS: ensemble model output statistics; SSh: Schaake shuffle; d: dual; md: minimum divergence; sim: similarity-based; Q: equidistant quantiles; R: random sample; S: stratified sample [Colour figure can be viewed at wileyonlinelibrary.com]

considered. In the interest of ensuring the direct comparability of the methods, we restricted all approaches to the fixed sample size of the raw ensemble. That said, one advantage of the SSh variants is that they, in principle, allow for generating post-processed ensemble forecasts of arbitrary size, which might be advantageous for better modelling extreme events (Lerch et al., 2017) and offers a natural starting point for future research; for example, by investigating the effect of the sample size on the performance in terms of recently proposed weighted multivariate proper scoring rules (Allen *et al.*, 2022).

The case studies considered here provide several avenues for further generalization and analysis. Though we have restricted our attention to temporal dependencies between lead times from 1 to 10 days, it would be interesting to also systematically compare the predictive performance in terms of spatial or intervariable dependencies. Further, our focus was on copula-based two-step approaches to multivariate post-processing. Alternative methods based on parametric models for the full joint distribution (Baran and Möller, 2015; Feldmann *et al.*, 2019) or quantile mapping (Whan *et al.*, 2021) have been proposed and offer a natural starting point for further comparisons.

Recent research in post-processing has demonstrated the benefits of incorporating additional predictors on

the forecasting performance of univariate methods; for example, see Rasp and Lerch (2018). Though these advanced post-processing methods can serve as building blocks of multivariate post-processing schemes, incorporating additional predictor information in the second, copula-based step is challenging, calling for the development of tailored approaches to machine-learning methods for multivariate post-processing. There have been first studies in this direction focusing on obtaining spatially coherent forecast fields via generative adversarial networks (Dai and Hemri, 2021) and multivariate post-processing using scoring-rule-based generative models that allow for incorporating additional predictors (Chen *et al.*, 2022).

Finally, the evaluation of multivariate predictive performance remains a challenging problem, and different multivariate evaluation metrics result in different rankings of the approaches considered. For example, disentangling the various contributions to multivariate forecast performance (univariate performance, multivariate dependencies, quantification of forecast uncertainty, etc.) and better understanding their effect on, for example, the differences in variability of the ES and VS observed in our case studies is difficult. Though there has been recent progress on the methodological aspects of multivariate evaluation (Ziel and Berk, 2019; Alexander *et al.*, 2022; Allen *et al.*, 2022), the need for systematic comparisons of the discrimination ability of multivariate proper scoring rules, ideally based on standardized benchmark datasets, constitutes an important pathway towards a better understanding of advantages and disadvantages of individual metrics. As noted in Lerch *et al.* (2020b), post-processing studies based on large datasets, such as the one investigated here, might provide helpful insights in this regard.

AUTHOR CONTRIBUTIONS

Mária Lakatos: dataCuration; investigation; software; validation; visualization; writingReviewEditing. **Sebastian Lerch:** formalAnalysis; investigation; validation; writingOriginalDraft; writingReviewEditing. **Stephan Hemri:** formalAnalysis; investigation; software; writingOriginalDraft; writingReviewEditing. **Sándor Baran:** formalAnalysis; investigation; software; supervision; validation; writingOriginalDraft; writingReviewEditing.

ACKNOWLEDGEMENTS

We gratefully acknowledge support by the Deutsche Forschungsgemeinschaft (DFG) through project MO-3394/1-1 "Statistische Nachbearbeitung von Ensemble-Vorhersagen für verschiedene Wettervariablen". Sándor Baran is further supported by the Hungarian National Research, Development and Innovation Office under grant no. NN125679. Sebastian Lerch gratefully acknowledges support by the Vector Stiftung through the Young Investigator Group "Artificial Intelligence for Probabilistic Weather Forecasting". We thank the two anonymous reviewers, whose constructive comments helped to improve an earlier version of this paper.

ORCID

Sebastian Lerch https://orcid.org/0000-0002-3467-4375
Stephan Hemri https://orcid.org/0000-0002-5832-509X
Sándor Baran https://orcid.org/0000-0003-1035-004X

REFERENCES

Alexander, C., Coulon, M., Han, Y. and Meng, X. (2022) Evaluating the discrimination ability of proper multi-variate scoring rules. *Annals of Operations Research*. https://doi.org/10.1007/s10479-022-04611-9.

Allen, S., Ginsbourger, D. and Ziegel, J. (2022) Evaluating forecasts for high-impact events using transformed kernel scores. *Preprint*. Available at: https://doi.org/10.48550/arXiv.2202.12732

Baran, S. and Möller, A. (2015) Joint probabilistic forecasting of wind speed and temperature using Bayesian model averaging. *Environmetrics*, 26, 120–132.

Ben Bouallègue, Z., Heppelmann, T., Theis, S.E. and Pinson, P. (2016) Generation of scenarios from calibrated ensemble forecasts with a dual-ensemble copula-coupling approach. *Monthly Weather Review*, 144, 4737–4750.

Chen, J., Janke, T., Steinke, F. and Lerch, S. (2022) Generative machine learning methods for multivariate ensemble post-processing. *Preprint*. Available at: https://doi.org/10.48550/arXiv.2211.01345

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B. and Wilby, R. (2004) The Schaake shuffle: a method for reconstructing space–time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology*, 5, 243–262.

Dai, Y. and Hemri, S. (2021) Spatially coherent postprocessing of cloud cover ensemble forecasts. *Monthly Weather Review*, 149, 3923–3937.

Delle Monache, L., Hacker, J.P., Zhou, Y., Deng, X. and Stull, R.B. (2006) Probabilistic aspects of meteorological and ozone regional ensemble forecasts. *Journal of Geophysical Research*, 111, D24307.

Diebold, F.X. and Mariano, R.S. (1995) Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13, 253–263.

Feldmann, K., Richardson, D.S. and Gneiting, T. (2019) Grid- versus station-based postprocessing of ensemble temperature forecasts. *Geophysical Research Letters*, 46, 7744–7751.

Gneiting, T. (2014) Calibration of medium-range weather forecasts. *ECMWF Technical Memorandum*, 719, 30. Available at. http://www.ecmwf.int/sites/default/files/elibrary/2014/9607-calibration-medium-range-weather-forecasts.pdf.

Gneiting, T., Balabdaoui, F. and Raftery, A.E. (2007) Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B Statistical Methodology*, 69, 243–268.

Gneiting, T. and Raftery, A.E. (2007) Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association*, 102, 359–378.

Gneiting, T., Raftery, A.E., Westveld, A.H. and Goldman, T. (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133, 1098–1118.

Gneiting, T. and Ranjan, R. (2011) Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29, 411–422.

Gneiting, T., Stanberry, L.I., Grimit, E.P., Held, L. and Johnson, N.A. (2008) Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *TEST*, 17, 211–235.

Haslett, J. and Raftery, A.E. (1989) Space–time modelling with long-memory dependence: assessing Ireland's wind power resource (with discussion). *Journal of the Royal Statistical Society Series C: Applied Statistics*, 38, 1–50.

Heinrich, C., Hellton, K.H., Lenkoski, A. and Thorarinsdottir, T.L. (2021) Multivariate postprocessing methods for high-dimensional seasonal weather forecasts. *Journal of the American Statistical Association*, 116, 1048–1059.

Hemri, S., Scheuerer, M., Pappenberger, F., Bogner, K. and Haiden, T. (2014) Trends in the predictive performance of raw ensemble weather forecasts. *Geophysical Research Letters*, 41, 9197–9205.

Hu, Y., Schmeits, M.J., van Andel, J.S., Verkade, J.S., Xu, M., Solomatine, D.P. and Liang, Z. (2016) A stratified sampling approach for improved sampling from a calibrated ensemble forecast distribution. *Journal of Hydrometeorology*, 17, 2405–2417.

Jordan, A., Krüger, F. and Lerch, S. (2019) Evaluating probabilistic forecasts with scoringRules. *Journal of Statistical Software*, 90, 1–37.

Lang, M.N., Lerch, S., Mayr, G.J., Simon, T., Stauffer, R. and Zeileis, A. (2020) Remember the past: a comparison of time-adaptive training schemes for non-homogeneous regression. *Nonlinear Processes in Geophysics*, 27, 23–34.

Lerch, S., Baran, S., Möller, A., Groß, J., Schefzik, R., Hemri, S. and Graeter, M. (2020a) Simulation-based comparison of multivariate ensemble post-processing methods. *Nonlinear Processes in Geophysics*, 27, 349–371.

Lerch, S., Thorarinsdottir, T.L., Ravazzolo, F. and Gneiting, T. (2020b) Forecaster's dilemma: extreme events and forecast evaluation. *Statistical Science*, 32, 106–127.

Möller, A., Lenkoski, A. and Thorarinsdottir, T.L. (2013) Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas. *Quarterly Journal of the Royal Meteorological Society*, 139, 982–991.

Raftery, A.E., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005) Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133, 1155–1174.

Rasp, S. and Lerch, S. (2018) Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, 146, 3885–3900.

Rüschendorf, L. (2009) On the distributional transform, Sklar's theorem, and the empirical copula process. *Journal of Statistical Planning and Inference*, 139, 3921–3927.

Schefzik, R. (2016) A similarity-based implementation of the Schaake shuffle. *Monthly Weather Review*, 144, 1909–1921.

Schefzik, R. (2017) Ensemble calibration with preserved correlations: unifying and comparing ensemble copula coupling and member-by-member postprocessing. *Quarterly Journal of the Royal Meteorological Society*, 143, 999–1008.

Schefzik, R. and Möller, A. (2018) Ensemble postprocessing methods incorporating dependence structures. In: Vannitsem, S., Wilks, D.S. and Messner, J.W. (Eds.) *Statistical postprocessing of ensemble forecasts*. Amsterdam: Elsevier, pp. 91–125.

Schefzik, R., Thorarinsdottir, T.L. and Gneiting, T. (2013) Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, 28, 616–640.

Scheuerer, M. (2014) Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*, 140, 1086–1096.

Scheuerer, M. and Büermann, L. (2014) Spatially adaptive post-processing of ensemble forecasts for temperature. *Journal of the Royal Statistical Society Series C Applied Statistics*, 63, 405–422.

Scheuerer, M. and Hamill, T.M. (2015) Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143, 1321–1334.

Scheuerer, M., Hamill, T.M., Whitin, B., He, M. and Henkel, A. (2017) A method for preferential selection of dates in the Schaake shuffle approach to constructing spatiotemporal forecast fields of temperature and precipitation. *Water Resources Research*, 53, 3029–3046.

Schulz, B. and Lerch, S. (2022) Machine learning methods for postprocessing ensemble forecasts of wind gusts: a systematic comparison. *Monthly Weather Review*, 150, 235–257.

Sklar, A. (1959) Fonctions de répartition à n dimensions et leurs marges. *Publications de Institut de Statistique de Université de Paris*, 8, 229–231.

Taillardat, M. (2021) Skewed and mixture of gaussian distributions for ensemble postprocessing. *Atmosphere*, 12, 966.

Thorarinsdottir, T.L. and Gneiting, T. (2010) Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society. Series A, Statistics in Society*, 173, 371–388.

Thorarinsdottir, T.L., Gneiting, T. and Gissibl, N. (2013) Using proper divergence functions to evaluate climate models. *SIAM/ASA Journal of Uncertainty Quantification*, 1, 522–534.

Thorarinsdottir, T.L., Scheuerer, M. and Heinz, C. (2016) Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. *Journal of Computational and Graphical Statistics*, 25, 105–122.

Vannitsem, S., Bremnes, J.B., Demaeyer, J., Evans, G.R., Flowerdew, J., Hemri, S., Lerch, S., Roberts, N., Theis, S., Atencia, A., Ben Boualègue, Z., Bhend, J., Dabernig, M., De Cruz, L., Hieta, L., Mestre, O., Moret, L., Odak Plenkovič, I., Schmeits, M., Taillardat, M., Van den Bergh, J., Van Schaeybroeck, B., Whan, K. and Ylhaisi, J. (2021) Statistical postprocessing for weather forecasts – review, challenges and avenues in a big data world. *The Bulletin of the American Meteorological Society*, 102, E681–E699.

Vardi, Y. and Zhang, C.H. (2000) The multivariate L_1-median and associated data depth. *Proceedings of the National Academy of Sciences of the USA*, 97, 1423–1426.

Whan, K., Zscheischler, J., Jordan, A.I. and Ziegel, J.F. (2021) Novel multivariate quantile mapping methods for ensemble post-processing of medium-range forecasts. *Weather and Climate Extremes*, 32, 100310.

Wilks, D.S. (2019) *Statistical methods in the atmospheric sciences*, 4th edition. Amsterdam: Elsevier.

Wilks, D.S. (2020) Regularized Dawid–Sebastiani score for multivariate ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 146, 2421–2431.

Ziel, F. and Berk, K. (2019) Multivariate forecasting evaluation: on sensitive and strictly proper scoring rules. *Preprint*. Available at: https://doi.org/10.48550/arXiv.1910.07325

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Lakatos, M., Lerch, S., Hemri, S. & Baran, S. (2023) Comparison of multivariate post-processing methods using global ECMWF ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 1–22. Available from: https://doi.org/10.1002/qj.4436