Received: 7 September 2020 | Accepted: 11 March 2021

DOI: 10.1002/env.2678

RESEARCH ARTICLE

WILEY

Truncated generalized extreme value distribution-based ensemble model output statistics model for calibration of wind speed ensemble forecasts

Sándor Baran[1] | **Patrícia Szokol**[1] | **Marianna Szabó**[2]

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

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

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

Funding information
European Social Fund, Grant/Award Numbers: EFOP-3.6.1-16-2016-00022, EFOP-3.6.2-16-2017-00015; Ministry for Innovation and Technology, Hungary, Grant/Award Number: ÚNKP-19-3; National Research, Development and Innovation Office, Grant/Award Number: NN125679

Abstract

In recent years, ensemble weather forecasting has become a routine at all major weather prediction centers. These forecasts are obtained from multiple runs of numerical weather prediction models with different initial conditions or model parametrizations. However, ensemble forecasts can often be underdispersive and also biased, so some kind of postprocessing is needed to account for these deficiencies. One of the most popular state of the art statistical postprocessing techniques is the ensemble model output statistics (EMOS), which provides a full predictive distribution of the studied weather quantity. We propose a novel EMOS model for calibrating wind speed ensemble forecasts, where the predictive distribution is a generalized extreme value (GEV) distribution left truncated at zero (TGEV). The truncation corrects the disadvantage of the GEV distribution-based EMOS models of occasionally predicting negative wind speed values, without affecting its favorable properties. The new model is tested on four datasets of wind speed ensemble forecasts provided by three different ensemble prediction systems, covering various geographical domains and time periods. The forecast skill of the TGEV EMOS model is compared with the predictive performance of the truncated normal, log-normal and GEV methods and the raw and climatological forecasts as well. The results verify the advantageous properties of the novel TGEV EMOS approach.

KEYWORDS

continuous ranked probability score, ensemble calibration, ensemble model output statistics, truncated generalized extreme value distribution

1 | INTRODUCTION

Wind speed has become one of the most important weather quantities in our rapidly changing economy, hence precise and reliable wind forecasting is of utmost importance in renewable energy production or in air pollution modeling (see e.g., Tagle et al., 2020, the corresponding discussion and references therein). At the base of forecasting such—and many other—weather variables lie the calculations of numerical weather prediction (NWP) models, which rely on the physical and chemical models of the atmosphere and the oceans. Accounting for the uncertainties of the process and the sometimes unreliable initial conditions it is customary to run multiple instances of the NWP models with its initial conditions perturbed. The resulting system is called an ensemble of forecasts (Leith, 1974), and it provides the possibility of probabilistic forecasting (Gneiting & Raftery, 2005), where together with the forecasts the corresponding information about forecast uncertainty is also estimated. However, as has been observed with several operational ensemble prediction systems (EPSs), ensemble forecasts often suffer from systematic errors such as bias or lack of calibration, which problems need to be accounted for (see e.g., Buizza et al., 2005). A popular approach is to use some form of statistical postprocessing (Buizza, 2018).

In the last decades various statistical calibration methods have been developed for a wide range of weather quantities including parametric models providing full predictive distributions (Gneiting et al., 2005; Raftery et al., 2005), non-parametric approaches (see e.g., Bremnes, 2019; Friederichs & Hense, 2007) or most recently, machine learning techniques (Rasp & Lerch, 2018; Taillardat & Mestre, 2020). This paper focuses on parametric postprocessing where one of the most widely used methods is the ensemble model output statistics (EMOS) suggested by Gneiting et al. (2005). It fits a single probability distribution to the ensemble forecast with its parameters depending on the ensemble members. Different weather quantities require different probability laws as predictive distributions, moreover, the link functions connecting the parameters of these distributions to the ensemble members might also differ. For example, a normal distribution provides a reasonable model for temperature and pressure (Gneiting et al., 2005), whereas for the nonnegative and skew-distributed wind speed, according to Thorarinsdottir and Gneiting (2010), a truncated normal (TN) distribution makes a good choice. In order to provide a better fit to high wind speed values, Lerch and Thorarinsdottir (2013) and Baran and Lerch (2015) suggest models based on generalized extreme value (GEV) and log-normal (LN) distributions, respectively, and a regime-switching approach combining the advantages of these heavy tailed laws with those of the light tailed TN model. More flexibility can be obtained by mixture EMOS models combining light and heavy tailed distributions, where the parameters and weights of a mixture of two forecast laws are estimated jointly (Baran & Lerch, 2016). However, a general disadvantage of these latter approaches is the increased computation cost. A more general approach to improving forecast skill is based on a two-step combination of predictive distributions from individual postprocessing models. In the first step, individual EMOS models based on single parametric distributions are estimated, whereas in the second step the forecast distributions are combined utilizing state of the art forecast combination techniques (see e.g., Baran & Lerch, 2018; Bassetti et al., 2018; Gneiting & Ranjan, 2013).

In the present work we concentrate on EMOS models based on a single parametric distribution. The case studies of Lerch and Thorarinsdottir (2013) and Baran and Lerch (2015) revealed the superiority of the GEV EMOS model compared with the competing TN and LN EMOS approaches, especially for high wind speeds. However, the GEV model has the disadvantage of assigning positive probability to negative wind speed values. We propose a novel EMOS approach to calibrating wind speed ensemble forecasts, where the predictive distribution is a left truncated GEV distribution with cut-off at 0 (TGEV). On the basis of four case studies using wind speed forecasts of three different EPSs, the forecast skill of the TGEV EMOS model is compared with the predictive performance of the TN, LN and GEV EMOS models, the climatological forecasts and the raw ensemble as well.

The paper is organized as follows. Section 2 contains the detailed description of the four wind speed datasets. In Section 3 the applied EMOS models, including the novel TGEV EMOS approach, are reviewed, and the methods of parameter estimation and model verification are given. The results of the four case studies are provided in Section 4, followed by a concluding Section 5. Finally, details of calculations and some additional results are given in Appendix.

2 | DATA

In order to provide a fair comparison with the existing distribution-based EMOS models, first we consider the same three datasets of ensemble forecasts and corresponding observations as in Baran and Lerch (2015) (and later studied in Baran & Lerch, 2016, 2018), which differ in the observed wind quantity, in the forecast lead time and in the stochastic properties of the ensemble. Each data set contains ensemble predictions for a single forecast horizon ranging from 24 to 48 h, hence we call them short-range forecasts. For these data we limit the description to a short summary and refer to Baran and Lerch (2015) and the references therein for more details. Further, we compare the predictive performance of the different EMOS models on a much larger database, providing ensemble forecasts with different lead times ranging up to 360 h.

2.1 | Short-range ensemble forecasts

2.1.1 | UWME forecasts

The eight members of the University of Washington mesoscale ensemble (UWME) are generated by separate runs of the fifth generation Pennsylvania State University-National Center for Atmospheric Research mesoscale model (PSU-NCAR MM5) with different initial conditions (Grell et al., 1995). The EPS domain covers the Pacific Northwest region of North America with a 12-km grid and the dataset at hand contains 48-h ahead forecasts and the corresponding validating observations of the 10-m maximal wind speed (maximum of the hourly instantaneous wind speeds over the previous 12 h, given in m/s, see e.g., Sloughter et al., 2010) for 152 stations in the Automated Surface Observing Network (National Weather Service, 1998) in the U.S. states of Washington, Oregon, Idaho, California, and Nevada for calendar years 2007–2008. The forecasts are initialized at 0000 UTC and the generation of the ensemble ensures that its members are clearly distinguishable. Our analysis is focused on calendar year 2008 with additional data from December 2007 used for model training. Removing days and locations with missing data and stations where data are only available on a very few days results in 101 stations with a total of 27,481 individual forecast cases.

2.1.2 | ALADIN-HUNEPS ensemble

The Aire Limitée Adaptation dynamique Développement International-Hungary Ensemble Prediction System (ALADIN-HUNEPS) of the Hungarian Meteorological Service (HMS) covers a large part of continental Europe with a horizontal resolution of 8 km. The forecasts are obtained by dynamical downscaling of the global ARPEGE[1]-based PEARP[2] system of Metéo-France (Descamps et al., 2015; Horányi et al., 2006). The EPS provides one control member obtained from the unperturbed analysis and 10 members calculated using perturbed initial conditions. These members are statistically indistinguishable and thus can be considered as exchangeable, which fact should be taken into account in the formulation of post-processing models. We use ensembles of 42-h ahead forecasts (initialized at 1800 UTC) of the 10-m instantaneous wind speed (in m/s) issued for 10 major cities in Hungary for the 1-year period April 1, 2012 to March 31, 2013, together with the corresponding validation observations. 6 days with missing forecasts and/or observations are excluded from the analysis.

2.1.3 | ECMWF ensemble forecasts for Germany

The operational EPS of the European Centre for Medium-Range Weather Forecasts (ECMWF) comprises 50 perturbed (thus exchangeable) members and operates on a global 18-km grid (Leutbecher & Palmer, 2008; Molteni et al., 1996). First we consider 24-h ahead ECMWF forecasts of 10-m daily maximum wind speed initialized at 0000 UTC for the period between February 1, 2010 and April 30, 2011 along with corresponding verifying observations of 228 synoptic observation (SYNOP) stations over Germany. This dataset is identical to the one studied in Lerch and Thorarinsdottir (2013) and in Baran and Lerch (2015, 2016). Postprocessed forecasts are verified on the 1-year period between May 1, 2010 and April 30, 2011 containing 83,220 individual forecast cases, whereas forecast–observation pairs from April 2010 are used for training purposes.

2.2 | Global ECMWF forecasts with different forecast horizons

In order to compare the predictive performance of the various EMOS models for different prediction horizons, we also investigate a global dataset of ECMWF ensemble forecasts of 10-m daily maximal wind speed with lead times from 1 day up until 15 days initialized at 1200 UTC between January 1, 2014 and June 24, 2018, and validating SYNOP observations for calendar years 2014–2018. Thus, one has observations and corresponding ensemble forecasts with 15 different lead times for the period January 16, 2014 to June 25, 2018 with the exception of 2 days in between with missing forecast data.

[1] Action de Recherche Petite Echelle Grande Echelle
[2] Prévisino d'Ensemble ARPEGE

For the sake of consistency our analysis is restricted to SYNOP stations with complete data, meaning 1059 stations in Europe and Asia.

3 | ENSEMBLE MODEL OUTPUT STATISTICS

As already mentioned in the Introduction, EMOS is a commonly used method of statistical postprocessing, which fits a single probability distribution to the ensemble forecast with parameters depending on the ensemble members. In what follows, let $f_1, f_2, \dots, f_K$ denote a wind speed ensemble forecast for a given location, time and lead time under the assumption that the ensemble members are not exchangeable, so the individual members can be clearly distinguished and tracked either based on the individual initial conditions, or as one can depict a systematic behaviour of each ensemble member. Examples of EPSs with nonexchangeable members are the UWME introduced in Section 2.1.1 or the 30-member Consortium for Small-scale Modelling EPS of the German Meteorological Service (Ben Bouallègue et al., 2013).

However, recently most operational EPSs incorporate ensembles where at least some members are generated using perturbed initial conditions. Such groups of exchangeable forecasts appear, for example, in the ALADIN-HUNEPS ensemble and in the operational ECMWF ensemble described in Sections 2.1.2 and 2.1.3, respectively, but one can also mention multimodel EPSs such as the Grand Limited Area Model EPS ensemble (Iversen et al., 2011). In the following sections, if we have M ensemble members divided into K exchangeable groups, where the kth group contains $M_k \geq 1$ ensemble members $(\sum_{k=1}^{K} M_k = M)$, then notation $\overline{f}_k$ will be used for the mean of the corresponding kth ensemble group. Further, the overall ensemble mean and variance will be denoted by $\overline{f}$ and S^2, respectively.

3.1 | EMOS models for wind speed

To model wind speed a nonnegative and skewed distribution is required, such as Weibull (Justus et al., 1978) or gamma (Garcia et al., 1998) laws. Gamma distribution also serves as underlying law in a Bayesian model averaging (Sloughter et al., 2010) approach to parametric postprocessing of wind speed ensemble forecasts, whereas in EMOS modeling TN, LN, and GEV distributions have been utilized so far. Note, that TN and LN EMOS models have already been implemented in the `ensembleMOS` package of `R` (Yuen et al., 2018).

3.1.1 | TN EMOS model

Starting with the fundamental work of Thorarinsdottir and Gneiting (2010), TN distribution became a popular base for EMOS predictive distributions of wind speed (see e.g. Bremnes, 2019; Lerch & Baran, 2017). Denote by $\mathcal{N}_0(\mu, \sigma^2)$ the TN distribution with location μ, scale $\sigma > 0$, and lower truncation at 0, having probability density function (PDF)

$$g(x|\mu, \sigma) := \frac{1}{\sigma}\varphi\left((x-\mu)/\sigma\right)/\Phi(\mu/\sigma), \quad \text{if } x \geq 0,$$

and $g(x|\mu, \sigma) := 0$, otherwise, where φ is the PDF, while Φ denotes the cumulative distribution function (CDF) of the standard normal distribution. For the TN EMOS predictive distribution the location and scale are linked to the ensemble members via equations

$$\mu = a_0 + a_1 f_1 + \dots + a_K f_K \quad \text{and} \quad \sigma^2 = b_0 + b_1 S^2. \tag{3.1}$$

where $a_0 \in \mathbb{R}$ and $a_1, \dots, a_K, b_0, b_1 \geq 0$.

If the ensemble can be split into K groups of exchangeable members, then forecasts within a given group will share the same location parameter (Gneiting, 2014; Wilks, 2018) resulting in link functions

$$\mu = a_0 + a_1 \overline{f}_1 + \dots + a_K \overline{f}_K \quad \text{and} \quad \sigma^2 = b_0 + b_1 S^2. \tag{3.2}$$

According to the optimum score estimation principle of Gneiting and Raftery (2007), model parameters $a_0, a_1, \dots, a_K$ and b_0, b_1 are estimated by optimizing the mean value of a proper verification score over the training data, see Section 3.2.

3.1.2 | LN EMOS model

To address the modeling of large wind speeds Baran and Lerch (2015) propose an EMOS approach based on an LN distribution. This distribution is more applicable for high wind speed values due to its heavier upper tail. The PDF of the LN distribution $\mathcal{LN}(\mu, \sigma)$ with parameters μ and $\sigma > 0$ is

$$h(x|\mu, \sigma) := \frac{1}{x\sigma}\varphi\left((\log x - \mu)/\sigma\right), \quad \text{if } x \geq 0,$$

and $h(x|\mu, \sigma) := 0,$ otherwise, while the mean m and variance v are

$$m = \mathrm{e}^{\mu+\sigma^2/2} \quad \text{and} \quad v = \mathrm{e}^{2\mu+\sigma^2}\left(\mathrm{e}^{\sigma^2} - 1\right),$$

respectively. Obviously, an LN distribution can also be parametrized by these latter two quantities via equations

$$\mu = \log\left(\frac{m^2}{\sqrt{v + m^2}}\right) \quad \text{and} \quad \sigma = \sqrt{\log\left(1 + \frac{v}{m^2}\right)},$$

and in the LN EMOS model of Baran and Lerch (2015) m and v are affine functions of the ensemble and the ensemble variance, respectively, that is

$$m = \alpha_0 + \alpha_1 f_1 + \ldots + \alpha_K f_K \quad \text{and} \quad v = \beta_0 + \beta_1 S^2. \tag{3.3}$$

To estimate mean parameters $\alpha_0 \in \mathbb{R}, \alpha_1, \ldots, \alpha_K \geq 0$ and variance parameters $\beta_0, \beta_1 \geq 0$, one can again use the optimum score estimation principle and minimize an appropriate verification score over the training data.

In the case of existence of groups of exchangeable ensemble members, similar to (3.2), the equation for the mean in (3.3) is replaced by

$$m = \alpha_0 + \alpha_1 \overline{f}_1 + \ldots + \alpha_K \overline{f}_K. \tag{3.4}$$

3.1.3 | GEV distribution-based EMOS models

As an alternative to the TN EMOS approach exhibiting good predictive performance for high wind speed values, one can consider the EMOS model of Lerch and Thorarinsdottir (2013) based on a GEV distribution $\mathcal{GEV}(\mu, \sigma, \xi)$ with location μ, scale $\sigma > 0$ and shape ξ defined by CDF

$$G(x|\mu, \sigma, \xi) := \begin{cases} \exp\left(-\left[1 + \xi(\frac{x-\mu}{\sigma})\right]^{-1/\xi}\right), & \text{if } \xi \neq 0; \\ \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right), & \text{if } \xi = 0, \end{cases} \tag{3.5}$$

for $1 + \xi(\frac{x-\mu}{\sigma}) > 0$ and $G(x|\mu, \sigma, \xi) := 0,$ otherwise.

The model proposed by Lerch and Thorarinsdottir (2013) uses location and scale parameters

$$\mu = \gamma_0 + \gamma_1 f_1 + \ldots + \gamma_K f_K \quad \text{and} \quad \sigma = \sigma_0 + \sigma_1 \overline{f}, \tag{3.6}$$

with $\sigma_0, \sigma_1 \geq 0,$ while the shape parameter ξ does not depend on the ensemble members.

However, as argued in Lerch and Thorarinsdottir (2013) and in Baran and Lerch (2015), the GEV EMOS model has the disadvantage of forecasting negative wind speed with a positive probability. As a solution we propose a novel EMOS model where the predictive GEV distribution is truncated from below at 0. For $x \geq 0$ the CDF of this truncated GEV (TGEV) distribution $\mathcal{TGEV}(\mu, \sigma, \xi)$ with location μ, scale $\sigma > 0$ and shape ξ equals

$$G_0(x|\mu, \sigma, \xi) = \begin{cases} \frac{G(x|\mu,\sigma,\xi) - G(0|\mu,\sigma,\xi)}{1 - G(0|\mu,\sigma,\xi)}, & \text{if } G(0|\mu, \sigma, \xi) < 1; \\ 1, & \text{if } G(0|\mu, \sigma, \xi) = 1, \end{cases} \tag{3.7}$$

whereas negative values are obviously excluded from the support set of the TGEV distribution. For $\xi < 1$ (and $G(0|\mu,\sigma,\xi) < 1$) the $\mathcal{TGEV}(\mu,\sigma,\xi)$ distribution has a finite mean of

$$\begin{cases} \mu + (\Gamma(1-\xi) - 1)\frac{\sigma}{\xi}, & \text{if } \xi > 0 \text{ and } \xi\mu - \sigma > 0; \\ \mu - \frac{\sigma}{\xi} + \frac{\sigma(\Gamma_\ell(1-\xi,[1-\xi\mu/\sigma]^{-1/\xi}))/\xi}{1-\exp(-[1-\xi\mu/\sigma]^{-1/\xi})}, & \text{if } \xi \neq 0 \text{ and } \xi\mu - \sigma \leq 0; \\ \frac{\mu+\sigma(C-\mathrm{Ei}(-\exp[\mu/\sigma]))}{1-\exp(-\exp[\mu/\sigma])}, & \text{if } \xi = 0, \end{cases} \tag{3.8}$$

where Γ and Γ_ℓ denote the gamma and the lower incomplete gamma function, respectively, defined as

$$\Gamma(a) = \int_0^\infty t^{a-1}\mathrm{e}^{-t}\mathrm{d}t \quad \text{and} \quad \Gamma_\ell(a,x) = \int_0^x t^{a-1}\mathrm{e}^{-t}\mathrm{d}t,$$

and $\mathrm{Ei}(x)$ is the exponential integral

$$\mathrm{Ei}(x) = \int_{-\infty}^x \frac{e^t}{t}\mathrm{d}t = C + \ln|x| + \sum_{k=1}^\infty \frac{x^k}{k!k}$$

with C being the Euler–Mascheroni constant. It is important to emphasize, that the case $\xi < 0$ and $\xi\mu - \sigma > 0$ does not appear in the formula (3.8), since in that case the PDF of $\mathcal{GEV}(\mu,\sigma,\xi)$ is positive only on $]-\infty, \mu - \sigma/\xi] \subset \mathbb{R}_-$. Further, as for $\xi > 0$ and $\xi\mu - \sigma > 0$ the support of $\mathcal{GEV}(\mu,\sigma,\xi)$ is $[\mu - \sigma/\xi, \infty[\subset \mathbb{R}_+$, truncation does not change the distribution and the means of $\mathcal{GEV}(\mu,\sigma,\xi)$ and $\mathcal{TGEV}(\mu,\sigma,\xi)$ distributions coincide. For the proof of the remaining two cases of (3.8) see Appendix A.

The parameters of the TGEV EMOS model are also linked to the ensemble members according to (3.6), which is replaced by

$$\mu = \gamma_0 + \gamma_1\overline{f}_1 + \ldots + \gamma_K\overline{f}_K \quad \text{and} \quad \sigma = \sigma_0 + \sigma_1\overline{f}, \tag{3.9}$$

in the exchangeable case. Note that alternative expressions

$$\sigma = \sigma_0 + \sigma_1 S, \quad \sigma = \sqrt{\sigma_0 + \sigma_1 S^2} \quad \text{and} \quad \sigma = \sigma_0 + \sigma_1\mathrm{MD},$$

of the scale have also been tested, where

$$\mathrm{MD} := \frac{1}{K^2}\sum_{k,\ell=1}^K |f_k - f_\ell|,$$

is the ensemble mean absolute difference (see e.g., Baran et al., 2020; Scheuerer, 2014). However, in our case studies TGEV EMOS models with link functions (3.6) and (3.9) show the best predictive performance.

3.2 | Training data selection and verification scores

As mentioned before, estimates of the unknown parameters of the EMOS models described in Sections 3.1.1–3.1.3 can be obtained with the help of the optimum score estimation principle of Gneiting and Raftery (2007), that is by optimizing a proper scoring rule over an appropriately chosen training dataset. Here we consider the standard approach in EMOS modeling and use rolling training periods. This means that model parameters for a given date are obtained using ensemble forecasts and corresponding validating observations for the preceding n calendar days. Given a training period length, there are two traditional approaches to spatial selection of training data (Thorarinsdottir & Gneiting, 2010). The global (or regional) approach uses ensemble forecasts and validating observations from all available stations during the rolling training period resulting in a single set of parameters for the whole ensemble domain. By contrast, the local estimation produces distinct parameter estimates for different stations by using only the training data of the given station. Local models typically result in better predictive performance compared with regional models (see e.g., Thorarinsdottir & Gneiting, 2010); however, require significantly longer training periods to avoid numerical stability issues (Lerch & Baran, 2017). In the case studies of Section 4 examples of both estimation techniques are shown.

In atmospheric sciences the most popular scoring rules are the logarithmic score (LogS; Good, 1952) and the continuous ranked probability score (CRPS; see e.g., Wilks, 2011). The former is the negative logarithm of the predictive PDF evaluated at the verifying observation, whereas for a (predictive) CDF F and real value (verifying observation) x the latter is defined as

$$\mathrm{CRPS}(F,x) := \int_{-\infty}^{\infty} \left[F(y) - \mathbb{I}_{\{y\geq x\}}\right]^2 \mathrm{d}y = \mathsf{E}|X-x| - \frac{1}{2}\mathsf{E}|X-X'|, \tag{3.10}$$

where X and X' are independent random variables distributed according to F and having a finite first moment, while $\mathbb{I}_H$ denotes the indicator function of a set H. Note that both LogS and CRPS are negative-oriented scores, that is, the smaller the better. Further, the optimization with respect to the logarithmic score results in the maximum likelihood (ML) estimation of the parameters, while the second expression in (3.10) implies that the CRPS can be expressed in the same unit as the observation. For all wind speed models of Sections 3.1.1–3.1.3 the CRPS can be expressed in closed form allowing efficient optimization procedures; for TN, LN and GEV laws we refer to Thorarinsdottir and Gneiting (2010), Baran and Lerch (2015), and Friederichs and Thorarinsdottir (2012), respectively. The CRPS of a TGEV distribution $\mathcal{TGEV}(\mu,\sigma,\xi)$ with CDF $G_0(x)$ derived from a GEV CDF $G(x)$ equals

$$\begin{aligned}\mathrm{CRPS}(G_0,x) = {}&(2G_0(x)-1)\left(x-\mu+\frac{\sigma}{\xi}\right) + \frac{\sigma}{\xi(1-G(0))^2}\left[-2^{\xi}\Gamma_{\ell}\left(1-\xi,-2\ln G(0)\right)\right.\\ &\left.+2G(0)\Gamma_{\ell}\left(1-\xi,-\ln G(0)\right)+2\left(1-G(0)\right)\Gamma_{\ell}\left(1-\xi,-\ln G(x)\right)\right],\end{aligned} \tag{3.11}$$

for $\xi \neq 0$, whereas for $\xi = 0$ we have

$$\begin{aligned}\mathrm{CRPS}(G_0,x) = {}&(x-\mu)\left(2G_0(x)-1\right) + \frac{\sigma}{(1-G(0))^2}\\ &\times\left(C-\ln 2+\mathrm{Ei}\left(2\ln G(0)\right)+(G(0))^2\ln\left[-\ln G(0)\right]-2G(0)\mathrm{Ei}\left(\ln G(0)\right)\right)\\ &+\frac{2\sigma}{1-G(0)}\left[G(x)\ln\left[-\ln G(x)\right]-\mathrm{Ei}\left(\ln G(x)\right)\right].\end{aligned} \tag{3.12}$$

For the proof of (3.11) and (3.12) see Appendix B.

In order to compare the predictive performance of the EMOS models for high wind speed values we also consider the threshold-weighted continuous ranked probability score (twCRPS; Gneiting & Ranjan, 2011)

$$\mathrm{twCRPS}(F,x) := \int_{-\infty}^{\infty} \left[F(y) - \mathbb{I}_{\{y\geq x\}}\right]^2 \omega(y)\mathrm{d}y, \tag{3.13}$$

where $\omega(y) \geq 0$ is a weight function. Setting $\omega(y) \equiv 1$ results in the traditional CRPS (3.10), whereas with the help of $\omega(y) = \mathbb{I}_{\{y\geq r\}}$ one can address wind speeds above a given threshold r. Note that in the case studies of Section 4 the thresholds correspond approximately to the 90th, 95th and 98th percentiles of the wind speed observations of all considered stations.

The improvement in terms of CRPS and twCRPS for a forecast F with respect to a reference forecast F_{ref} can be quantified using the continuous ranked probability skill score (CRPSS; see e.g., Gneiting & Raftery, 2007; Murphy, 1973) and the threshold-weighted continuous ranked probability skill score (twCRPSS; Lerch & Thorarinsdottir, 2013)

$$\mathrm{CRPSS} := 1 - \frac{\overline{\mathrm{CRPS}}}{\overline{\mathrm{CRPS}}_{\mathrm{ref}}} \quad \text{and} \quad \mathrm{twCRPSS} := 1 - \frac{\overline{\mathrm{twCRPS}}}{\overline{\mathrm{twCRPS}}_{\mathrm{ref}}},$$

respectively, where $\overline{\mathrm{CRPS}}$, $\overline{\mathrm{twCRPS}}$ and $\overline{\mathrm{CRPS}}_{\mathrm{ref}}$, $\overline{\mathrm{twCRPS}}_{\mathrm{ref}}$ denote the mean score values corresponding to F and F_{ref} over the verification data. Skill scores are obviously positively oriented, that is larger skill scores mean better predictive performance.

Point forecasts such as EMOS and ensemble medians and means can be evaluated using the mean absolute errors (MAEs) and the root mean squared errors (RMSEs), where the former is optimal for the median, whereas the latter is optimal for the mean forecasts (Gneiting, 2011).

The uncertainty in the verification scores is assessed with the help of confidence intervals for mean score values and skill scores. These intervals are calculated from 2,000 block bootstrap samples, which are obtained using the stationary bootstrap scheme with mean block length computed according to Politis and Romano (1994).

Simple and widely used tools of graphically assessing the calibration of probabilistic forecasts are the verification rank histogram (or Talagrand diagram) of ensemble predictions and its continuous counterpart, the probability integral transform (PIT) histogram. The verification rank is the rank of the verifying observation with respect to the corresponding ensemble forecast (see e.g., Wilks, 2011, section 8.7.2), whereas the PIT is the value of the predictive CDF evaluated at the verifying observation (Dawid, 1984; Raftery et al., 2005). In the case of a properly calibrated K-member ensemble the verification ranks follow a uniform distribution on $\{1, 2, \dots, K+1\}$, while PIT values of calibrated predictive distributions are uniformly distributed on the $[0,1]$ interval.

Finally, calibration and sharpness of a predictive distribution can also be investigated by examining the coverage and average width of the $(1-\alpha)100\%$, $\alpha \in]0,1[$, central prediction interval, respectively. Here the coverage is the proportion of the validating observations located between the lower and upper $\alpha/2$ quantiles of the predictive CDF, and level α should be chosen to match the nominal coverage of the raw ensemble, that is $(K-1)/(K+1)100\%$, where again, K is the ensemble size. As the coverage of a calibrated predictive distribution should be around $(1-\alpha)100\%$, such a choice of α allows a direct comparison with the ensemble coverage.

4 | RESULTS

The forecast skill of the novel TGEV EMOS model proposed in Section 3.1.3 is tested both on short-range (24–48 h) wind speed forecasts of the eight-member UWME, of the 11-member ALADIN-HUNEPS ensemble, and of the 50-member ECMWF ensemble, and on more recent global surface wind forecasts of the operational EPS of the ECMWF with lead times 1,2, … ,15 days, for more details see Section 2. As reference models we consider the TN, LN and GEV EMOS approaches described in Sections 3.1.1–3.1.3, respectively, and the raw ensemble and climatological forecasts (observations of the training period are considered as an ensemble) as well. For the sake of brevity here we report only the main results, further details can be found in Appendices C and D.

4.1 | Implementation details

In the case studies presented here the estimates of TN and LN EMOS model parameters minimize the mean CRPS of forecast-observation pairs over the training data. Objective functions are optimized using the popular Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (see e.g., Press et al., 2007, section 10.9). However, for the more complex GEV and TGEV models the estimation methods used in the case studies of Sections 4.2 and 4.3 differ. For the short-range forecasts of Section 4.2 we follow the suggestions of Lerch and Thorarinsdottir (2013) and calculate the ML estimates of the GEV parameters, whereas for the TGEV model we consider the box constrained version of BFGS (L-BFGS-B; Byrd et al., 1995) and keep the shape parameter ξ in the interval $]-0.278, 1/3[$ to ensure a finite mean and a positive skewness. Note that the ML estimates of the shape of the GEV model also remain in this interval. For the global ECMWF forecasts of Section 4.2 both GEV and TGEV parameters are estimated by minimizing the mean CRPS of the training data with the help of a BFGS algorithm, where the constrains on scale and shape parameters are forced using appropriate transformations. All optimization tasks are performed using the `optim` function of `R` allowing at most 200 iteration steps. In the case of TN and LN models starting parameters of location/mean are computed with a linear regression of the observations on the corresponding forecasts, whereas the starting points for the scale parameters are fixed. In the case of GEV and TGEV models all iterations are started from fixed initial points.

4.2 | Short-range ensemble forecasts

The case studies of this section are based on those three wind speed data sets that have already been investigated in Baran and Lerch (2015, 2016). We use the same training and verification data for the TGEV modeling (global training

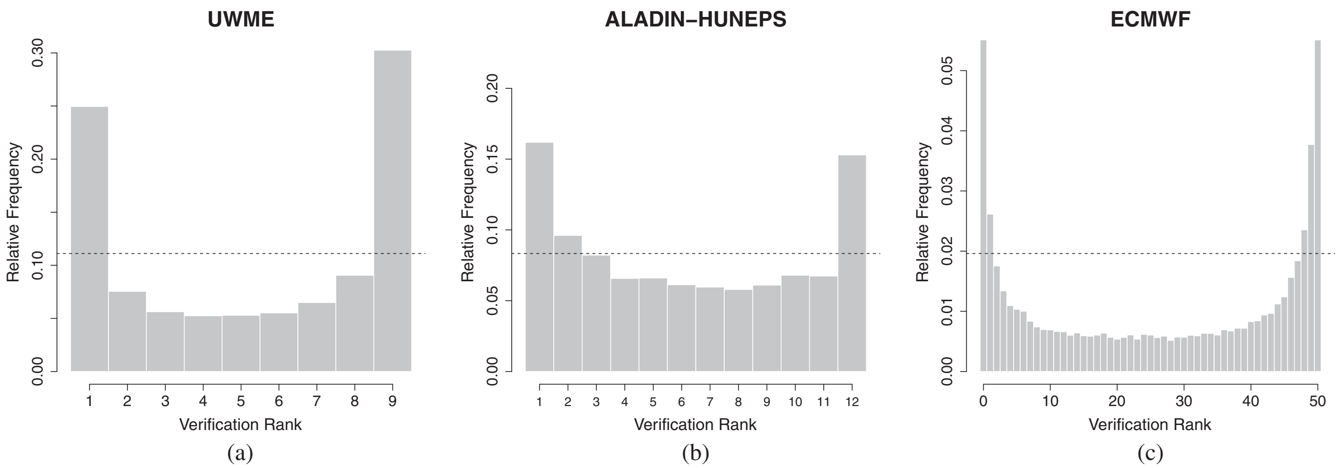


FIGURE 1 Verification rank histograms. (a) University of Washington mesoscale ensemble for the calendar year 2008; (b) Aire Limitée Adaptation dynamique Développement International-Hungary Ensemble Prediction System ensemble for the period April 1, 2012 to March 31, 2013; (c) European Centre for Medium-Range Weather Forecasts ensemble for the period May 1, 2010 to April 30, 2011

TABLE 1 Mean continuous ranked probability score (CRPS) and mean absolute error (MAE) of median forecasts together with 95% confidence intervals, root mean squared error (RMSE) of mean forecasts and coverage and average width of 77.78% central prediction intervals for the University of Washington mesoscale ensemble. Mean and maximal probability of predicting negative wind speed by the generalized extreme value (GEV) model: 0.05% and 4%

Forecast	CRPS (m/s)	MAE (m/s)	RMSE (m/s)	Cover. (%)	Av. w. (m/s)
Truncated normal	1.114 (1.052,1.188)	1.550 (1.466,1.655)	2.048	78.65	4.67
Log-normal	1.114 (1.052,1.188)	1.554 (1.465,1.658)	2.052	77.29	4.69
GEV	1.100 (1.041,1.174)	1.554 (1.463,1.656)	2.047	77.20	4.69
Truncated GEV	1.099 (1.038,1.173)	1.551 (1.464,1.656)	2.046	76.69	4.62
Ensemble	1.353 (1.274,1.460)	1.655 (1.554,1.775)	2.169	45.24	2.53
Climatology	1.412 (1.291,1.539)	1.987 (1.820,2.170)	2.629	81.10	5.90

with matching training period lengths) as in the earlier works, allowing a direct comparison with the performance of the previously investigated TN, LN and GEV EMOS models.

4.2.1 | EMOS models for the UWME

As one can observe on Figure 1(a), the verification rank histogram of the eight-member UWME wind speed forecasts for calendar year 2008 is highly U-shaped, indicating a strongly underdispersive character. The ensemble range contains the validating observation in only 45.24% of cases, which is far below the nominal coverage of 77.78%, calling for some form of calibration.

As the eight members of the UWME are nonexchangeable, for postprocessing we make use of TN and LN EMOS models (3.1) and (3.3), respectively, and GEV and TGEV EMOS with parametrization (3.6), where $K=8$. Ensemble forecasts for calendar year 2008 are calibrated using a 30-day training period, which training period length is a result of a detailed preliminary analysis, see Baran and Lerch (2015).

In Table 1 a summary of verification scores and coverage and average width of nominal 77.78% central prediction intervals are given for the competing EMOS models and the raw and climatological UWME forecasts (27,481 forecast cases), whereas Table 2 reports the mean twCRPS values corresponding to various thresholds. Climatological forecasts

TABLE 2 Mean threshold-weighted continuous ranked probability score (twCRPS) for various thresholds r together with 95% confidence intervals for the University of Washington mesoscale ensemble

	twCRPS (m/s)		
Forecast	$r = 9$	$r = 10.5$	$r = 14$
Truncated normal	0.150 (0.116,0.189)	0.074 (0.054,0.099)	0.010 (0.005,0.016)
Log-normal	0.149 (0.115,0.186)	0.073 (0.053,0.098)	0.010 (0.005,0.017)
Generalized extreme value (GEV)	0.145 (0.112,0.183)	0.072 (0.052,0.095)	0.010 (0.005,0.018)
Truncated GEV	0.145 (0.112,0.180)	0.072 (0.052,0.096)	0.010 (0.005,0.017)
Ensemble	0.175 (0.134,0.226)	0.085 (0.061,0.115)	0.011 (0.005,0.019)
Climatology	0.173 (0.132,0.220)	0.081 (0.058,0.111)	0.010 (0.005,0.017)

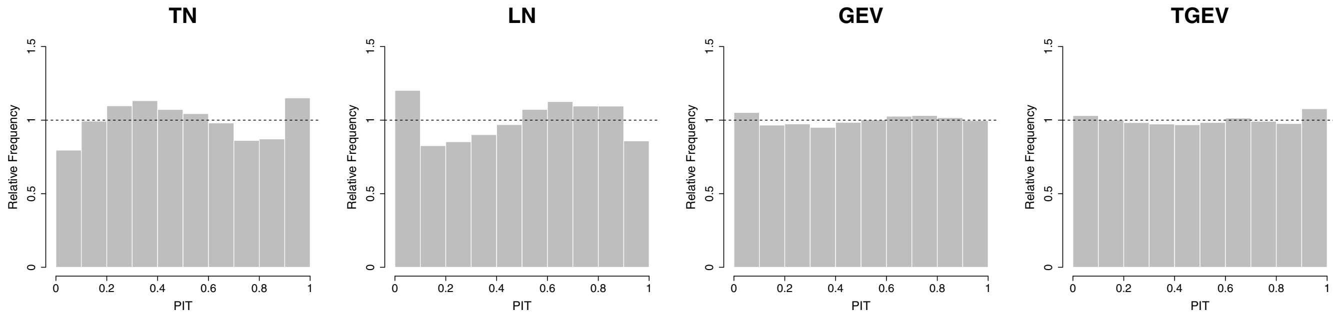


FIGURE 2 Probability integral transform histograms of the ensemble model output statistics-calibrated University of Washington mesoscale ensemble forecasts

underperform the raw ensemble in terms of mean CRPS, MAE, and RMSE, but have better skill on the tails which is quantified in lower mean twCRPS values. As mentioned before, the underdispersive character of the raw forecasts leads to poor coverage and very sharp central prediction intervals, whereas the climatological prediction intervals are much wider resulting in a far better coverage. EMOS postprocessing improves the calibration and forecast skill of the raw ensemble by a wide margin as all EMOS scores but the mean twCRPS corresponding to most extreme wind speeds are much lower than the corresponding scores of raw and climatological forecasts. The advantage in terms of the mean CRPS is significant. The coverage of each calibrated forecast is very close to the nominal value; however, one should also note that these central prediction intervals are less sharp than the intervals calculated from the raw ensemble. From the competing EMOS approaches, the novel TGEV model results in the lowest mean CRPS, RMSE, and twCRPS values (which are either identical with or very close to the corresponding GEV EMOS scores), whereas in terms of MAE it is slightly outperformed by the TN EMOS method. Further, the TGEV model leads to the sharpest central prediction intervals, which is naturally connected with a slight decrease in coverage. For a deeper analysis of the tail behaviour of the different EMOS approaches we refer to Figure C1(a) of Appendix C showing the twCRPSS with respect to the TN EMOS as function of the threshold.

Finally, compared with the verification rank histogram of the raw UWME forecasts (Figure 1(a)), the PIT histograms of the different EMOS models displayed in Figure 2 are much closer to the desired uniform distribution, indicating an improved calibration. TN and LN EMOS result in slightly biased and hump-shaped histograms, whereas the histograms of GEV and TGEV approaches are almost perfectly flat. These shapes are nicely in line with the corresponding CRPS values of Table 1.

Based on the above results one can conclude that in the case of the UWME forecasts, from the competing EMOS approaches the novel TGEV model shows the best forecast skill, closely followed by the GEV EMOS. This conclusion is also supported by the results of Appendix D, where the calibration of the investigated EMOS models at different forecast levels is addressed. However, in connection with the GEV model one should not forget about the positive probability of predicting negative wind speed values. For the UWME forecasts at hand the mean and maximum of these probabilities are 0.05% and 4%, respectively (Baran & Lerch, 2015).

TABLE 3 Mean continuous ranked probability score (CRPS) and mean squared error (MAE) of median forecasts together with 95% confidence intervals, root mean square error (RMSE) of mean forecasts and coverage and average width of 83.33% central prediction intervals for the Aire Limitée Adaptation dynamique Développement International-Hungary Ensemble Prediction System ensemble. Mean and maximal probability of predicting negative wind speed by the generalized extreme value (GEV) model: 0.33% and 9.46%

Forecast	CRPS (m/s)	MAE (m/s)	RMSE (m/s)	Cover. %	Av.w. (m/s)
Truncated normal	0.738 (0.689,0.793)	1.037 (0.966,1.112)	1.357	83.59	3.53
Log-normal	0.741 (0.690,0.799)	1.038 (0.960,1.125)	1.362	80.44	3.57
GEV	0.737 (0.685,0.793)	1.041 (0.970,1.117)	1.355	81.21	3.54
Truncated GEV	0.736 (0.685,0.793)	1.037 (0.969,1.114)	1.356	82.13	3.53
Ensemble	0.803 (0.749,0.865)	1.069 (1.001,1.136)	1.373	68.22	2.88
Climatology	1.046 (0.944,1.149)	1.481 (1.333,1.627)	1.922	82.54	4.92

TABLE 4 Mean threshold-weighted continuous ranked probability score (twCRPS) for various thresholds r together with 95% confidence intervals for the Aire Limitée Adaptation dynamique Développement International-Hungary Ensemble Prediction System ensemble

	twCRPS (m/s)		
Forecast	$r=6$	$r=7$	$r=9$
Truncated normal	0.102 (0.062,0.147)	0.054 (0.027,0.085)	0.012 (0.003,0.022)
Log-normal	0.102 (0.062,0.145)	0.054 (0.028,0.084)	0.011 (0.004,0.022)
Generalized extreme value (GEV)	0.098 (0.062,0.143)	0.052 (0.026,0.081)	0.011 (0.003,0.021)
Truncated GEV	0.099 (0.058,0.145)	0.052 (0.026,0.082)	0.011 (0.003,0.022)
Ensemble	0.112 (0.069,0.163)	0.059 (0.030,0.093)	0.013 (0.004,0.026)
Climatology	0.127 (0.076,0.190)	0.064 (0.031,0.102)	0.012 (0.003,0.023)

4.2.2 | EMOS models for the ALADIN–HUNEPS ensemble

Compared with the UWME discussed in the previous section, the ALADIN-HUNEPS ensemble is better calibrated. Although the verification rank histogram given in Figure 1(b) still shows overconfidence, resulting in large bins at the sides, it is much closer to the uniform distribution than the one in Figure 1(a), and the ensemble coverage of 61.21% is also closer to the nominal 83.33%.

The structure of the ALADIN-HUNEPS ensemble induces a natural division of the ensemble members into two exchangeable groups: the first contains just the control member, while the second consists of the members obtained from random perturbations of the initial conditions $(M=11,\ K=2,\ M_1=1,\ M_2=10)$. Hence, calibration is performed using EMOS models with distribution locations/means linked to the ensemble members via (3.2), (3.4), and (3.9).

The detailed data analysis of Baran et al. (2014) suggests a 43-day training period for EMOS postprocessing of ALADIN-HUNEPS ensemble forecasts, leaving 315 calendar days (3, 150 forecast cases) between May 15, 2012 and March 31, 2013 for forecast verification.

Again, Table 3 showing the verification scores of different forecasts and the coverage and average width of nominal 83.33% central prediction intervals justifies the use of statistical post-processing. All EMOS models result in reasonably sharp forecasts with coverage values close to the nominal one outperforming both the raw and climatological forecasts in terms of all reported scores. The positive effect of statistical calibration can also be observed on mean twCRPS values provided in Table 4; however, one should also be aware of the large uncertainty in the forecasts. Among the different post-processing approaches, the TGEV EMOS yields the lowest mean CRPS and MAE and the sharpest central prediction interval combined with a coverage that is the second closest to the nominal one. However, in terms of twCRPS addressing the predictive performance at high wind speed values, GEV EMOS seems to show better forecast skill. This can also be

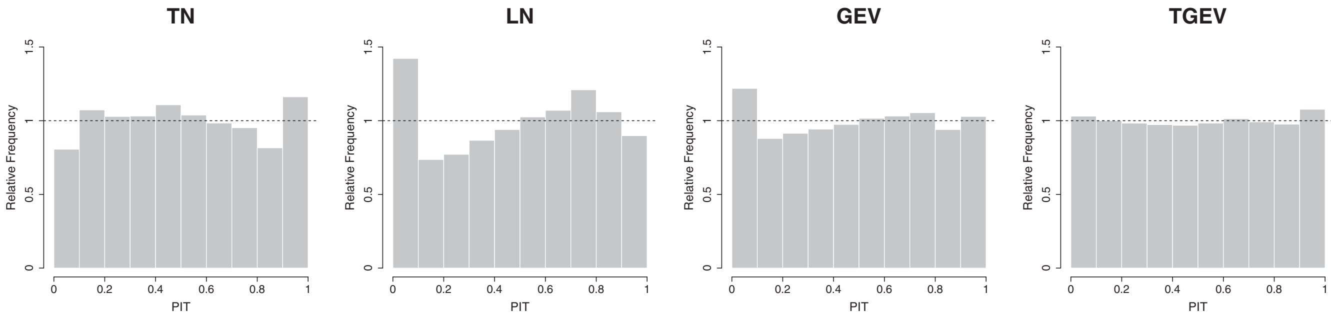


FIGURE 3 Probability integral transform histograms of the ensemble model output statistics-calibrated Aire Limitée Adaptation dynamique Développement International-Hungary Ensemble Prediction System ensemble forecasts

observed in Figure C1(b) of Appendix C, where the twCRPSS values with respect to the TN EMOS are plotted as function of the threshold. The GEV EMOS clearly outperforms the competitors; however, the situation is nuanced by the fact that in the case of ALADIN-HUNEPS ensemble forecasts the maximal probability of predicting negative wind speed is 9.46%, and the mean value of these probabilities is also 0.33%.

The improved calibration of postprocessed ALADIN-HUNEPS forecasts can also be observed on PIT histograms of Figure 3, which are much closer to uniformity than the corresponding verification rank histogram, see Figure 1(b). Here the TGEV model results in the flattest histogram, whereas the PIT histograms of TN, LN and GEV models are slightly hump-shaped and biased. Hence, keeping in mind also the results of Appendix D, one can conclude that in the case of the ALADIN-HUNEPS ensemble forecasts, from the presented four EMOS approaches the TGEV has the best overall performance.

4.2.3 | EMOS models for the ECMWF forecasts for Germany

From the three EPSs investigated in Section 4.2, the ECMWF ensemble exhibits the lack of calibration to the highest extent. In most cases the ensemble forecasts either under-, or overestimate the validating observation, resulting in a coverage of 43.40%, whereas the nominal coverage is 96.08%. The underdispersive character of the forecasts can also be clearly observed on the corresponding verification rank histogram (see Figure 1(c)).

The 50 members of operational ECMWF EPS are regarded as exchangeable, so in the link functions (3.2), (3.4), and (3.9) we have $K=1$ and $\overline{f}_1$ equals the ensemble mean. Following the suggestions of Baran and Lerch (2015), the parameters of the EMOS models for calibrating ECMWF ensemble forecast for the period May 1, 2010 to April 30, 2011 (83,220 forecast cases) are estimated globally using a rolling training period of length 20 days.

Similar to Sections 4.2.1 and 4.2.2, in Table 5 the mean CRPS, MAE, and RMSE of postprocessed, raw, and climatological forecasts are reported together with the corresponding coverage and average width of 96.08% (nominal) central prediction intervals, while Table 6 provides the mean twCRPS scores for three different thresholds. The picture we get after examining these values is also similar to the previous cases: postprocessing results in improved predictive performance and better calibration. The lowest CRPS, MAE, and twCRPS values belong to the TGEV EMOS model, which has a fair coverage, but slightly less sharp than the TN and LN EMOS.

Although the mean twCRPS values and the corresponding 95% confidence intervals of GEV and TGEV models given in Table 6 are almost identical, Figure C1(c) of Appendix C displaying again the twCRPSS with respect to TN EMOS reveals the differences between the tail behavior of the two methods and indicates the superiority of the novel TGEV EMOS approach. Note also that here the mean and maximal probabilities of predicting negative wind speed by the GEV model are 0.01% and 5%, respectively.

Finally, the comparison of the PIT histograms of Figure 4 with the verification rank histogram of the raw ECMWF ensemble (see Figure 1(c)) again shows that postprocessing substantially improves the calibration of forecasts. However, one should also note that none of the competing EMOS methods results in uniformly distributed PIT values. For example the GEV EMOS model is slightly overdispersive having heavy tails, which is fully in line with the wide nominal central prediction intervals (see Table 5), whereas the tails of the TN EMOS model are slightly too light. TGEV and LN EMOS PIT values show the smallest deviation from uniformity, hence, for the studied ECMWF forecasts again the TGEV EMOS

TABLE 5 Mean continuous ranked probability score (CRPS) and mean squared error (MAE) of median forecasts together with 95% confidence intervals, root mean square error (RMSE) of mean forecasts and coverage and average width of 96.08% central prediction intervals for the European Centre for Medium-Range Weather Forecasts ensemble forecasts for Germany. Mean and maximal probability of predicting negative wind speed by the generalized extreme value (GEV) model: 0.01% and 5%

Forecast	CRPS (m/s)	MAE (m/s)	RMSE (m/s)	Cover. %	Av.w. (m/s)
Truncated normal	1.045 (0.974,1.125)	1.388 (1.298,1.488)	2.148	92.19	6.39
Log-normal	1.037 (0.970,1.112)	1.386 (1.298,1.482)	2.138	93.16	6.91
GEV	1.034 (0.960,1.114)	1.388 (1.300,1.488)	2.134	94.84	8.22
Truncated GEV	1.031 (0.962,1.112)	1.385 (1.298,1.480)	2.135	92.89	7.37
Ensemble	1.263 (1.194,1.345)	1.441 (1.373,1.523)	2.232	45.00	1.80
Climatology	1.550 (1.406,1.700)	2.144 (1.948,2.340)	2.986	95.84	11.91

TABLE 6 Mean threshold-weighted continuous ranked probability score (twCRPS) for various thresholds r together with 95% confidence intervals for the European Centre for Medium-Range Weather Forecasts ensemble forecasts for Germany

	twCRPS (m/s)		
Forecast	$r = 10$	$r = 12$	$r = 15$
Truncated normal	0.200 (0.150,0.255)	0.110 (0.075,0.147)	0.042 (0.024,0.062)
Log-normal	0.198 (0.146,0.254)	0.109 (0.075,0.149)	0.042 (0.024,0.062)
Generalized extreme value (GEV)	0.195 (0.145,0.250)	0.106 (0.072,0.145)	0.041 (0.024,0.059)
Truncated GEV	0.194 (0.143,0.248)	0.106 (0.072,0.143)	0.041 (0.024,0.060)
Ensemble	0.211 (0.155,0.272)	0.113 (0.077,0.152)	0.043 (0.025,0.061)
Climatology	0.251 (0.182,0.326)	0.128 (0.087,0.172)	0.045 (0.026,0.066)

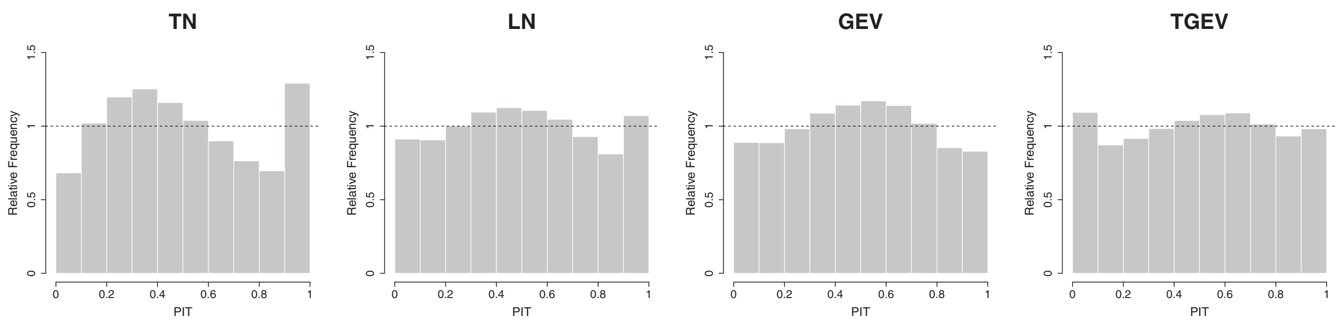


FIGURE 4 Probability integral transform histograms of the ensemble model output statistics-calibrated European Centre for Medium-Range Weather Forecasts for Germany

model has the best overall performance. Note that this conclusion is rather in line with the corresponding results of Appendix D.

4.3 | EMOS models for the global ECMWF ensemble forecasts

The case studies of Section 4.3 verify the positive effect of EMOS post-processing on calibration of short-term wind speed ensemble forecasts in general, and the superiority of the TGEV EMOS approach as well. However, as argued in the

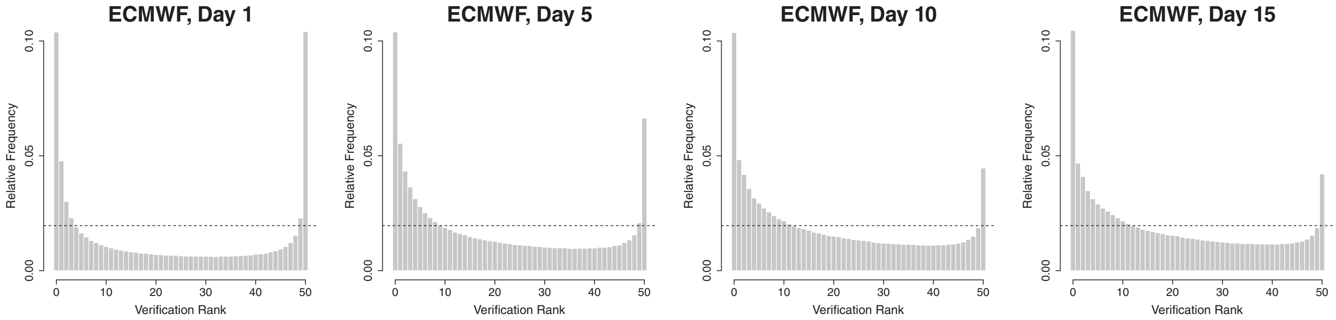
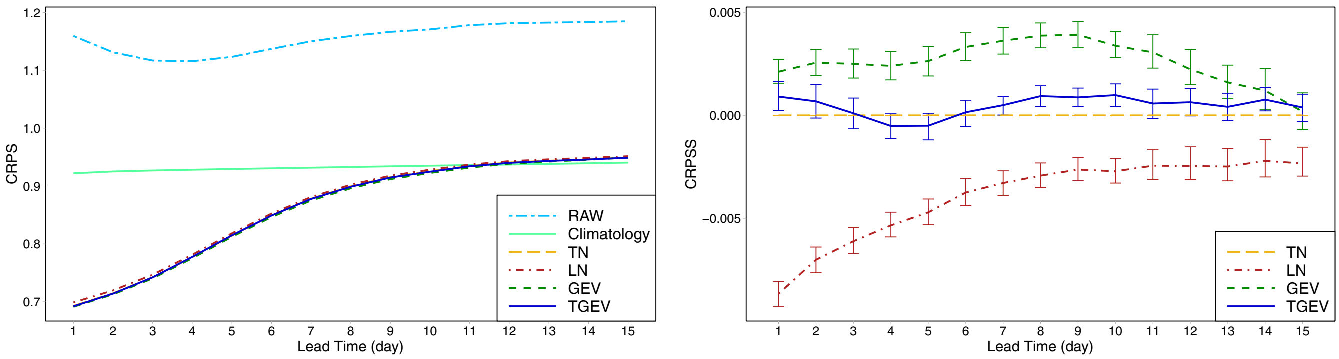


FIGURE 5 Verification rank histograms of the global European Centre for Medium-Range Weather Forecasts ensemble forecasts for the period January 16, 2014 to June 25, 2018



(a) (b)

FIGURE 6 (a) Continuous ranked probability score (CRPS) of the raw, climatological, and calibrated European Centre for Medium-Range Weather Forecasts global forecasts; (b) Continuous ranked probability skill score with respect to the truncated normal ensemble model output statistics model together with 95% confidence intervals

discussion of Feldmann et al. (2019), the longer the lead time, the more training data is needed for postprocessing to outperform the raw ensemble, and a similar conclusion can be derived from the results of Baran et al. (2020), too. This motivates the case study presented in this section, where calibration of global ECMWF wind speed ensemble forecasts with lead times $1, 2, \dots, 15$ days covering a very long time period of almost four and a half years is considered.

As one can observe on the verification rank histograms of Figure 5, the global ECMWF forecasts are strongly U-shaped for all lead times; however, the increase of the forecast horizon reduces underdispersion. This might be explained by the increase of forecast uncertainty resulting in wider ensemble range and better coverage, which improves from 52.05% of day 1 to 85.74% of day 15 (see also Figure 8).

For calibration we use the same EMOS model settings as in Section 4.2.3 considering a single group of exchangeable ensemble members; however in this case the large ensemble domain does not allow global modeling. Thus, local estimation with a rolling training period of 100 days is applied, which ensures a reasonably stable parameter estimation for all investigated EMOS approaches and leaves the period May 10, 2014 to June 25, 2018 (1,508 calendar days after excluding the 2 days with missing data) for validation purposes (1,596,972 individual forecast cases for each lead time).

In contrast to the case of ECMWF temperature forecasts investigated in Feldmann et al. (2019) or Baran et al. (2020), in terms of the mean CRPS all considered EMOS models outperform the raw wind speed ensemble forecasts for all lead times by a wide margin (see Figure 6(a)). Note that the nonmonotonic shape of the mean CRPS of the raw ensemble is a result of representativeness error in the verification, which can be partially corrected by adding up observation uncertainty to the ensemble spread (Ben Bouallègue, 2020). For shorter lead times EMOS models are also superior to climatology, but the advantage is decreasing with the lead time and disappears after day 11. To make visible the differences between the various EMOS approaches in terms of the mean CRPS, Figure 6(b) shows the CRPSS values with respect to the TN EMOS model. LN EMOS exhibits the worst forecast skill but the disadvantage decreases with the increase of the forecast horizon. GEV EMOS outperforms its competitors, followed by the TGEV EMOS, which has a significantly positive skill

TABLE 7 Mean and the 90th, 95th and 99th quantiles of probabilities (in %) of predicting negative wind speed by the generalized extreme value model

Day	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Mean	2.48	2.48	2.48	2.49	2.51	2.53	2.54	2.56	2.57	2.59	2.60	2.61	2.63	2.63	2.65
Q90	7.36	7.30	7.28	7.32	7.32	7.30	7.30	7.30	7.37	7.42	7.48	7.51	7.58	7.59	7.61
Q95	14.20	13.95	13.76	13.59	13.38	13.14	13.02	12.92	13.00	13.02	12.99	13.12	13.22	13.25	13.30
Q99	32.95	32.32	31.65	30.82	29.79	29.23	28.53	28.18	27.90	27.83	27.63	27.84	27.84	27.77	27.94

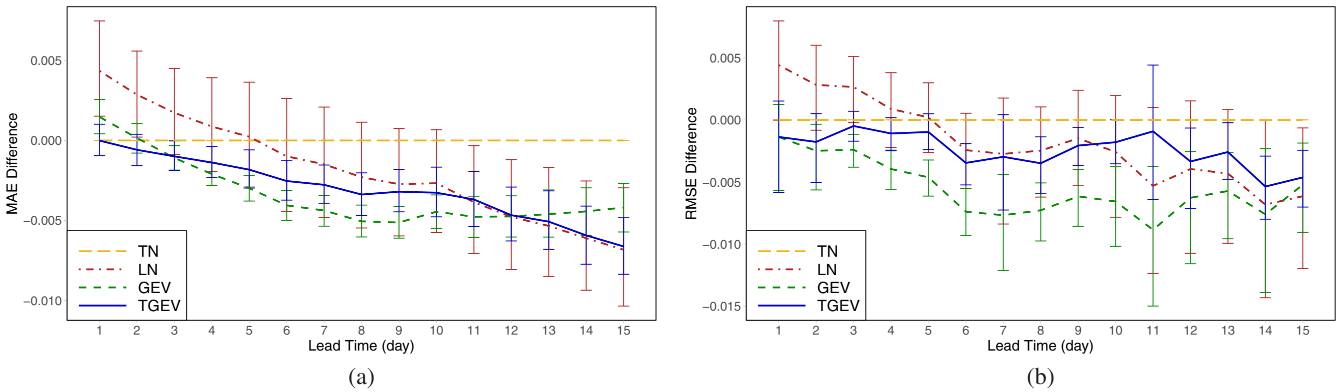


FIGURE 7 Difference in mean absolute error (a) and root mean square error (b) values from the reference truncated normal ensemble model output statistics model together with 95% confidence intervals

score for almost all lead times. Similar conclusions can be drawn from Figure D1 of Appendix D showing the skill scores separately for different forecast levels. However, for this global data set the problem of predicting negative wind speed values by the GEV EMOS approach is far more pronounced than in the case studies of Section 4.2. According to Table 7, the mean of these probabilities is around 2.5%, whereas the 99th quantiles range from 27.63% to 32.95%, which makes a possible operational use problematic.

In Figure 7(a),(b) the differences in MAE and RMSE from the reference TN EMOS model are given (the smaller the better). For short and very long lead times the TGEV EMOS results in the lowest MAE values, whereas between 4 and 10 days the GEV EMOS significantly outperforms its competitors. After day 11 the performance of the LN EMOS is similar to that of the TGEV EMOS; however, the uncertainty of the former is much higher. A different ranking can be observed in Figure 7(b), where the GEV EMOS results in the lowest score values, followed by the TGEV EMOS model, which for medium lead times behaves very similarly to the LN EMOS.

As expected, climatological forecasts result in the best coverage (Figure 8(a)), closely followed by the GEV EMOS. The coverage values of TGEV, TN, and LN EMOS approaches are slightly below 90% for all lead times and the corresponding curves are rather flat and very close to each other. In terms of sharpness, Figure 8(b) shows a clear ranking of the competing post-processing methods for all lead times. TN EMOS results in the narrowest central prediction intervals followed by TGEV, GEV and LN EMOS models.

To compare the tail behavior of the competing EMOS models we consider the twCRPSS values with respect to the TN EMOS approach for thresholds corresponding again to 90th, 95th, and 98th quantiles of the wind speed observations (see Figure 9). The ranking of the different EMOS models is consistent for all three investigated thresholds; after day 3 TN EMOS results in the best forecast skill, whereas the LN EMOS approach, similar to Figure 6(b), is far behind its competitors.

Finally, the PIT histograms of EMOS postprocessed forecasts for lead times 1, 5, 10, and 15 days plotted in Figure 10 again show the positive effect of postprocessing. They are much closer to uniformity than the verification rank histograms of the raw ECMWF ensemble forecasts of Figure 5; moreover, the shapes of the presented PIT histograms are nicely in line with the corresponding CRPS scores (Figure 6) and coverage and average widths of nominal central prediction intervals (Figure 8). PIT histograms of the LN EMOS approach show the largest deviation from uniformity, whereas the histograms

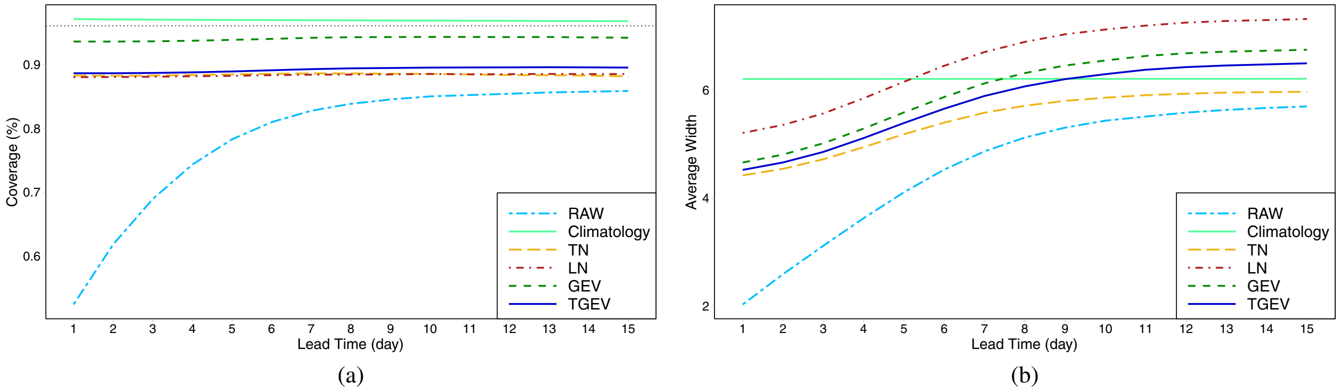


FIGURE 8 Coverage (a) and average width (b) of nominal 96.08% central prediction intervals. In panel (a) the ideal coverage is indicated by the horizontal dotted line

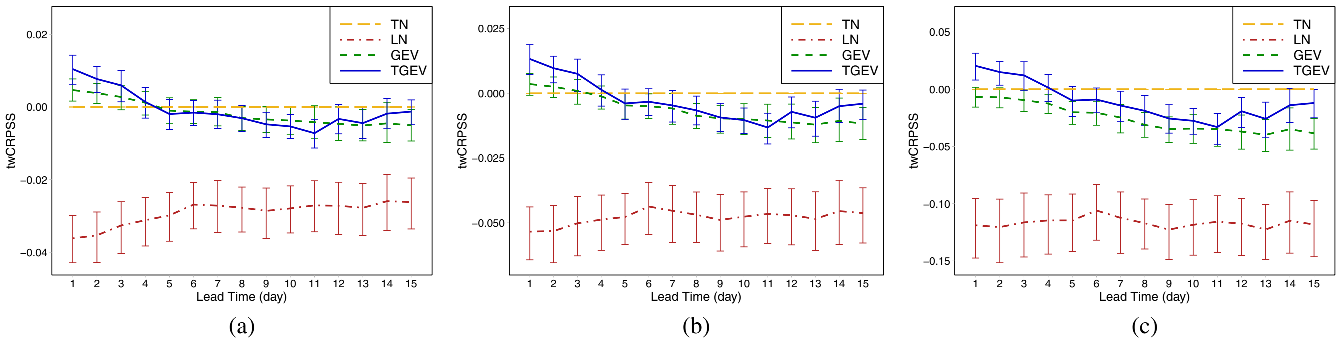


FIGURE 9 Threshold-weighted continuous ranked probability skill score values with respect to the truncated normal ensemble model output statistics model for thresholds 6 m/s (a), 7 m/s (b) and 9 m/s (c) together with 95% confidence intervals

of the GEV model are almost perfectly flat with a slight underdispersion, especially for longer lead times. TGEV EMOS also results in rather flat PIT histograms with slightly light lower tails for all lead times.

For the ECMWF data set at hand the GEV EMOS model shows the best overall predictive performance for all lead times, followed by the TGEV EMOS. However, looking back again to the mean probabilities of predicting negative wind speed by the GEV model given in Table 7, one should prefer the slightly less skillful novel TGEV EMOS approach.

5 | CONCLUSIONS

For the purpose of calibrating wind speed ensemble forecasts we propose a novel EMOS approach based on a truncated GEV distribution. The aim is to correct the deficiency of the efficient GEV EMOS method of Lerch and Thorarinsdottir (2013) of occasionally predicting negative wind speed. The TGEV EMOS model is tested both on short-range (24–48 h) wind speed forecasts of three completely different EPSs (eight-member UWME, 11-member ALADIN–HUNEPS and 50-member ECMWF) covering different and relatively small geographical regions and on a much larger dataset of global ECMWF forecasts for four and a half calendar years with lead times from 1 to 15 days. For model verification we use the CRPS of the probabilistic forecasts, the MAE of the median and the RMSE of the mean forecasts, and we also analyze the coverage and the average width of nominal central prediction intervals, which serve as measures of calibration and sharpness, respectively. Further, the predictive performance at high wind speed values is assessed with the help of the twCRPS for thresholds corresponding approximately to the 90th, 95th, and 98th percentiles of the observed wind speed.

The forecast skill of the TGEV EMOS model is compared to that of the TN, LN, and GEV EMOS approaches, and the raw and climatological forecasts. According to the results of the presented four case studies, postprocessing always improves the calibration of probabilistic and accuracy of point forecasts and all EMOS models outperform

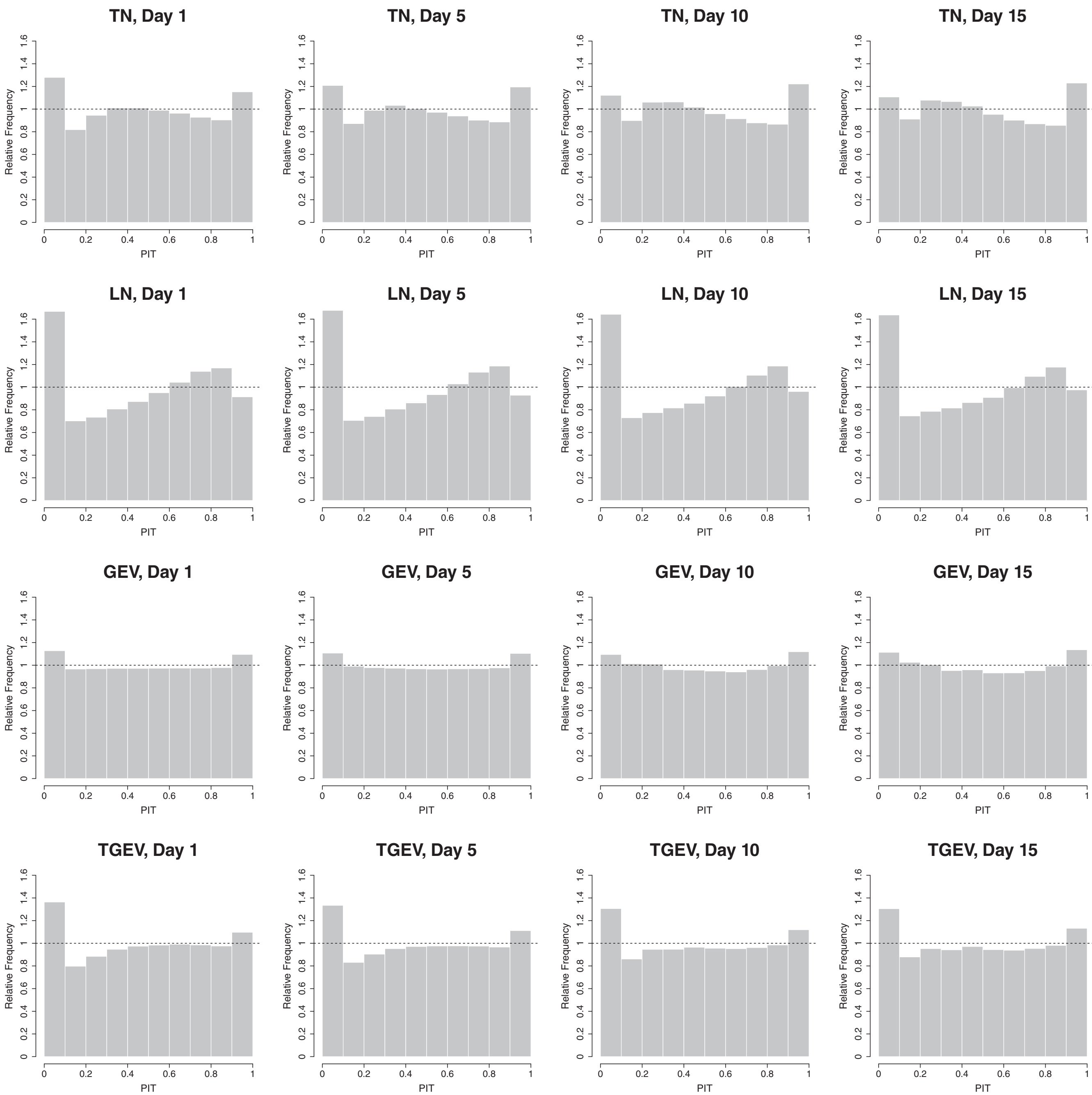


FIGURE 10 Probability integral transform histograms of the ensemble model output statistics postprocessed European Centre for Medium-Range Weather Forecasts global forecasts for days 1, 5, 10, and 15

both the raw ensemble and climatology. One can also observe that the TGEV EMOS approach has the best overall performance—regarding the four presented methods—closely followed by the GEV EMOS model. However, for the latter, at least in the case study of Section 4.3, the mean probability of predicting negative wind speed values is around 2.5% for all considered lead times.

In the present study our focus is restricted to univariate forecasts for a single location and lead time. However, most practical applications (e.g., in the context of wind energy forecasting, see Pinson & Messner, 2018) require an accurate modeling of spatial and temporal dependencies. Hence, multivariate extension of the proposed TGEV EMOS model in order to provide spatially and temporally consistent calibrated wind speed forecasts might be an interesting direction of future research. For a detailed overview of the possible approaches, see for example, Lerch et al. (2020).

ACKNOWLEDGEMENTS

Sándor Baran and Marianna Szabó were supported by the ÚNKP-19-3 New National Excellence Program of The Ministry for Innovation and Technology. Sándor Baran and Patrícia Szokol were supported by the National Research, Development and Innovation Office under Grant No. NN125679. Sándor Baran also acknowledges the support of the EFOP-3.6.2-16-2017-00015 project, while Patrícia Szokol was supported by the EFOP-3.6.1-16-2016-00022 project. Both projects were co-financed by the Hungarian Government and the European Social Fund. The authors are indebted to Zied Ben Bouallègue for providing the ECMWF data, the University of Washington MURI group for providing the UWME data, and Mihály Szűcs from the HMS for providing the ALADIN-HUNEPS data. Constructive comments of the two anonymous referees are gratefully acknowledged.

DATA AVAILABILITY STATEMENT

The data we used in this study are proprietary and the authors are not allowed to share it. However, ECMWF datasets studied in Sections 4.2.3 and 4.3 may be obtained from the European Centre for Medium-Range Weather Forecasts directly for research purposes.

ORCID

Sándor Baran https://orcid.org/0000-0003-1035-004X

REFERENCES

Baran, S., Baran, Á., Pappenberger, F., & Ben Bouallègue, Z. (2020). Statistical post-processing of heat index ensemble forecasts: is there a royal road? *Quarterly Journal of the Royal Meteorological Society*, *146*, 3416–3434.

Baran, S., Horányi, A., & Nemoda, D. (2014). Comparison of the BMA and EMOS statistical methods in calibrating temperature and wind speed forecast ensembles. *Időjárás*, *118*, 217–241.

Baran, S., & Lerch, S. (2015). Log-normal distribution based ensemble model output statistics models for probabilistic wind speed forecasting. *Quarterly Journal of the Royal Meteorological Society*, *141*, 2289–2299.

Baran, S., & Lerch, S. (2016). Mixture EMOS model for calibrating ensemble forecasts of wind speed. *Environmetrics*, *27*, 116–130.

Baran, S., & Lerch, S. (2018). Combining predictive distributions for the statistical post-processing of ensemble forecasts. *International Journal of Forecasting*, *34*, 477–496.

Bassetti, F., Casarin, R., & Ravazzolo, F. (2018). Bayesian nonparametric calibration and combination of predictive distributions. *Journal of the American Statistical Association*, *113*, 675–685.

Ben Bouallègue, Z. (2020). *Accounting for representativeness in the verification of ensemble forecasts*. (ECMWF Technical Memorandum, No. 865). https://doi.org/10.21957/5z6esc7wr

Ben Bouallègue, Z., Theis, S. E., & Gebhardt, C. (2013). Enhancing COSMO-DE ensemble forecasts by inexpensive techniques. *Meteorologische Zeitschrift*, *22*, 49–59.

Bremnes, J. B. (2019). Constrained quantile regression splines for ensemble postprocessing. *Monthly Weather Review*, *147*, 1769–1780.

Buizza, R. (2018). *Ensemble forecasting and the need for calibration*. In S. Vannitsem, D. S. Wilks, & J. W. Messner (Eds.), *Statistical postprocessing of ensemble forecasts* (pp. 15–48). Elsevier.

Buizza, R., Houtekamer, P. L., Toth, Z., Pellerin, G., Wei, M., & Zhu, Y. (2005). A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Monthly Weather Review*, *133*, 1076–1097.

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *16*, 1190–1208.

Dawid, A. P. (1984). Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *147*, 278–292.

Descamps, L., Labadie, C., Joly, A., Bazile, E., Arbogast, P., & Cébron, P. (2015). PEARP, the Météo-France short-range ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, *141*, 1671–1685.

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

Friederichs, P., & Hense, A. (2007). Statistical downscaling of extreme precipitation events using censored quantile regression. *Monthly Weather Review*, *135*, 2365–2378.

Friederichs, P., & Thorarinsdottir, T. L. (2012). Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics*, *23*, 579–594.

Garcia, A., Torres, J. L., Prieto, E., & De Francisco, A. (1998). Fitting wind speed distributions: A case study. *Solar Energy*, *62*, 139–144.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, *106*, 746–762.

Gneiting, T. (2014). *Calibration of medium-range weather forecasts* (ECMWF Technical Memorandum No. 719). https://doi.org/10.21957/8xna7glta

Gneiting, T., & Raftery, A. E. (2005). Weather forecasting with ensemble methods. *Science*, *310*, 248–249.

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

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

Gneiting, T., & Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, *7*, 1747–1782.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *14*, 107–114.

Grell, G. A., Dudhia, J., & Stauffer, D. R. (1995) *A description of the fifth-generation Penn state/NCAR mesoscale model (MM5)* (NCAR Technical Note, NCAR/TN-398+STR. 122). Boulder. https://doi.org/10.5065/D60Z716B.

Horányi, A., Kertész, S., Kullmann, L., & Radnóti, G. (2006). The ARPEGE/ALADIN mesoscale numerical modeling system and its application at the Hungarian Meteorological Service. *Időjárás*, *110*, 203–227.

Iversen, T., Deckmin, A., Santos, C., Sattler, K., Bremnes, J. B., Feddersen, H., & Frogner, I.-L. (2011). Evaluation of 'GLAMEPS' – A proposed multimodel EPS for short range forecasting. *Tellus A*, *63*, 513–530.

Justus, C. G., Hargraves, W. R., Mikhail, A., & Graber, D. (1978). Methods for estimating wind speed frequency distributions. *Journal of Applied Meteorology*, *17*, 350–353.

Leith, C. E. (1974). Theoretical skill of Monte-Carlo forecasts. *Monthly Weather Review*, *102*, 409–418.

Lerch, S., & Baran, S. (2017). Similarity-based semi-local estimation of EMOS models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *66*, 29–51.

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

Lerch, S., & Thorarinsdottir, T. L. (2013). Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus A*, *65*, 21206.

Leutbecher, M., & Palmer, T. N. (2008). Ensemble forecasting. *Journal of Computational Physics*, *227*, 3515–3539.

Molteni, F., Buizza, R., & Palmer, T. N. (1996). The ECMWF ensemble prediction system: methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, *122*, 73–119.

Murphy, A. H. (1973). Hedging and skill scores for probability forecasts. *Journal of Applied Meteorology*, *12*, 215–223.

National Weather Service (1998). *Automated Surface Observing System (ASOS) users guide*. National Weather Service: Silver Spring, MD. Retrieved from https://www.weather.gov/media/asos/aum-toc.pdf. Accessed September 5, 2020

Pinson, P., & Messner, J. W. (2018). *Application of postprocessing for renewable energy*. In S. Vannitsem, D. S. Wilks, & J. W. Messner (Eds.), *Statistical postprocessing of ensemble forecasts* (pp. 241–266). Elsevier.

Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, *89*, 1303–1313.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. T. (2007). *Numerical recipes in C: The art of scientific computing* (3rd ed.). Cambridge: Cambridge University Press.

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

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

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

Sloughter, J. M., Gneiting, T., & Raftery, A. E. (2010). Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *Journal of the American Statistical Association*, *105*, 25–37.

Tagle, F., Genton, M. G., Yip, A., Mostamandi, S., Stenchikov, G., & Castruccio, S. (2020). A high-resolution bilevel skew-*t* stochastic generator for assessing Saudi Arabia's wind energy resources. *Environmetrics*, *31*, e2628.

Taillardat, M., & Mestre, O. (2020). From research to applications – Examples of operational ensemble post-processing in France using machine learning. *Nonlinear Processes in Geophysics*, *27*, 329–347.

Thorarinsdottir, T. L., & 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.

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences* (3rd ed.). Elsevier.

Wilks, D. S. (2018). *Univariate ensemble forecasting*. In S. Vannitsem, D. S. Wilks, & J. W. Messner (Eds.), *Statistical postprocessing of ensemble forecasts* (pp. 49–89). Elsevier.

Yuen, R. A., Baran, S., Fraley, C., Gneiting, T., Lerch, S., Scheuerer, M., & Thorarinsdottir, T. L. (2018) R package ensembleMOS (Version 0.8.2). Ensemble model output statistics. Retrieved from https://cran.r-project.org/package=ensembleMOS. Accessed September 5, 2020

How to cite this article: Baran S, Szokol P, Szabó M. Truncated generalized extreme value distribution-based ensemble model output statistics model for calibration of wind speed ensemble forecasts. *Environmetrics*. 2021;e2678. https://doi.org/10.1002/env.2678

APPENDIX A. MEAN OF A TRUNCATED GEV DISTRIBUTION

To simplify the formulation of the results, similar to the notations of Section 3.2, in what follows we set aside the indication of the parameters of the GEV and TGEV CDFs G and G_0 defined by (3.5) and (3.7), respectively.

The present section is devoted to verification of the formula (3.8) for the TGEV mean in the nontrivial cases when G and G_0 differ. Let $\xi < 1$ and $0 < G(0) < 1$. The PDF $g_0(x)$ of a $\mathcal{TGEV}(\mu, \sigma, \xi)$ distribution defined by (3.7) equals

$$g_0(x) = \begin{cases} \dfrac{\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi-1} \exp\left(-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right)}{\sigma(1-G(0))}, & \text{if } \xi \neq 0; \\ \dfrac{\exp\left(\frac{x-\mu}{\sigma}\right)\exp\left(-\exp\left[-\frac{x-\mu}{\sigma}\right]\right)}{\sigma(1-G(0))}, & \text{if } \xi = 0, \end{cases} \tag{A1}$$

for $x \geq 0$ and $x\xi \geq \mu\xi - \sigma$, and $g_0(x) = 0$ otherwise, where

$$G(0) = \begin{cases} \exp(-[1-\xi\mu/\sigma]^{-1/\xi}), & \text{if } \xi \neq 0, \\ \exp(-\exp[\mu/\sigma]), & \text{if } \xi = 0. \end{cases}$$

Let X be a TGEV random variable and assume $\xi \neq 0$ and $\xi\mu - \sigma \leq 0$. If $\xi > 0$, then the support of $g_0(x)$ is $[0, \infty[$, so

$$\mathsf{E}X = \frac{1}{\sigma(1-G(0))}\int_0^\infty x\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi-1} \exp\left(-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right)\mathrm{d}x. \tag{A2}$$

For $\xi < 0$ the support of $g_0(x)$ changes to $[0, \mu - \sigma/\xi]$, so the integral in (A2) should be taken over this particular interval. However, in both cases the change of variables leads to

$$\begin{aligned} \mathsf{E}X &= \frac{1}{1-G(0)}\int_0^{\left(1-\frac{\xi\mu}{\sigma}\right)^{-1/\xi}} \left[\frac{(t^{-\xi}-1)\sigma}{\xi} + \mu\right]\exp(-t)\mathrm{d}t \\ &= \mu - \frac{\sigma}{\xi} + \frac{\sigma(\Gamma_\ell(1-\xi, [1-\xi\mu/\sigma]^{-1/\xi}))/\xi}{1-\exp(-[1-\xi\mu/\sigma]^{-1/\xi})}. \end{aligned}$$

Finally, let $\xi = 0$. In this case

$$\mathsf{E}X = \frac{1}{\sigma(1-G(0))}\int_0^\infty x\exp\left(\frac{x-\mu}{\sigma}\right)\exp\left(-\exp\left[-\frac{x-\mu}{\sigma}\right]\right)\mathrm{d}x,$$

where the change of variables with respect to $t = \exp\left(-\frac{x-\mu}{\sigma}\right)$ results in

$$\mathsf{E}X = \frac{1}{\sigma(1-G(0))}\int_0^{\exp(\mu/\sigma)} (\mu - \sigma\ln t)\exp(-t)\mathrm{d}t = \frac{\mu + \sigma(C - \mathrm{Ei}(-\exp[\mu/\sigma]))}{1-\exp(-\exp[\mu/\sigma])}.$$

□

APPENDIX B. CRPS OF A TRUNCATED GEV DISTRIBUTION

Following the ideas of Friederichs and Thorarinsdottir (2012), the CRPS of a TGEV distribution is derived using representation

$$\mathrm{CRPS}(G_0, x) = x(2G_0(x)-1) - 2\int_0^1 tG_0^{-1}(t)\mathrm{d}t + 2\int_{G_0(x)}^1 G_0^{-1}(t)\mathrm{d}t, \tag{B1}$$

where G_0^{-1} denotes the quantile function corresponding to G_0. Short calculation shows that for $0<y<1$

$$G_0^{-1}(y) = \begin{cases} \mu + \frac{\sigma}{\xi}\left(-1 + [-\ln \tau(y)]^{-\xi}\right), & \text{if } \xi \neq 0, \\ \mu - \sigma\left(\ln\left[-\ln \tau(y)\right]\right), & \text{if } \xi = 0, \end{cases} \quad \text{where} \quad \tau(y) := (1 - G(0))\,y + G(0).$$

Assume first $\xi \neq 0$. Then the first integral of (B1) equals

$$\begin{aligned} 2\int_0^1 tG_0^{-1}(t)\mathrm{d}t &= \mu - \frac{\sigma}{\xi} + \frac{2\sigma}{\xi}\int_0^1 t[-\ln \tau(t)]^{-\xi}\mathrm{d}t = \mu - \frac{\sigma}{\xi} + \frac{2\sigma}{\xi}\int_{G(0)}^1 \frac{\tau - G(0)}{(1-G(0))^2}[-\ln \tau]^{-\xi}\mathrm{d}\tau \\ &= \mu - \frac{\sigma}{\xi} + \frac{2\sigma}{\xi}\frac{1}{(1-G(0))^2}\left[\int_{G(0)}^1 \tau[-\ln \tau]^{-\xi}\mathrm{d}\tau - G(0)\int_{G(0)}^1 [-\ln \tau]^{-\xi}\mathrm{d}\tau\right]. \end{aligned}$$

Now, let Γ_u denote the upper incomplete gamma functions, defined as

$$\Gamma_u(a,x) = \int_x^\infty t^{a-1}\mathrm{e}^{-t}\mathrm{d}t.$$

Using $\Gamma(a) = \Gamma_\ell(a,x) + \Gamma_u(a,x)$, short calculations involving appropriate changes of variables show

$$\begin{aligned} \int_{G(0)}^1 \tau[-\ln \tau]^{-\xi}\mathrm{d}\tau &= 2^{\xi-1}\left[\Gamma(1-\xi) - \Gamma_u\left(1-\xi, -2\ln G(0)\right)\right] = 2^{\xi-1}\Gamma_\ell\left(1-\xi, -2\ln G(0)\right), \\ \int_{G(0)}^1 [-\ln \tau]^{-\xi}\mathrm{d}\tau &= \Gamma(1-\xi) - \Gamma_u\left(1-\xi, -\ln G(0)\right) = \Gamma_\ell\left(1-\xi, -\ln G(0)\right). \end{aligned}$$

Hence,

$$2\int_0^1 tG_0^{-1}(t)\mathrm{d}t = \mu - \frac{\sigma}{\xi} + \frac{\sigma}{\xi(1-G(0))^2}\left[2^\xi\Gamma_\ell\left(1-\xi, -2\ln G(0)\right) - G(0)\Gamma_\ell\left(1-\xi, -\ln G(0)\right)\right]. \tag{B2}$$

The second integral of (B1) can be evaluated in a similar way, resulting in

$$\int_{G_0(x)}^1 G_0^{-1}(t)\mathrm{d}t = (1 - G_0(x))\left(\mu - \frac{\sigma}{\xi}\right) + \frac{\sigma}{\xi(1-G(0))}\Gamma_\ell\left(1-\xi, -\ln G(x)\right). \tag{B3}$$

Finally, the combination of Equations (B1), (B2), and (B3) gives

$$\begin{aligned} \mathrm{CRPS}(G_0,x) = (2G_0(x)-1)\left(x - \mu + \frac{\sigma}{\xi}\right) + \frac{\sigma}{\xi(1-G(0))^2}&\left[-2^\xi\Gamma_\ell\left(1-\xi, -2\ln G(0)\right)\right. \\ &\left.+2G(0)\Gamma_\ell\left(1-\xi, -\ln G(0)\right) + 2\left(1-G(0)\right)\Gamma_\ell\left(1-\xi, -\ln G(x)\right)\right]. \end{aligned}$$

Now, let $\xi = 0$. In this case for the integrals in (B1) we have

$$\begin{aligned} 2\int_0^1 tG_0^{-1}(t)\mathrm{d}t &= \mu - 2\sigma\int_0^1 t\ln\left[-\ln \tau(t)\right]\mathrm{d}t = \mu - 2\sigma\int_{G(0)}^1 \frac{\tau - G(0)}{(1-G(0))^2}\ln[-\ln \tau]\mathrm{d}\tau \\ &= \mu - \frac{2\sigma}{(1-G(0))^2}\left[\int_{G(0)}^1 \tau\ln[-\ln \tau]\mathrm{d}\tau - G(0)\int_{G(0)}^1 \ln[-\ln \tau]\mathrm{d}\tau\right], \\ \int_{G_0(x)}^1 G_0^{-1}(t)\mathrm{d}t &= \mu\left(1 - G_0(x)\right) - \sigma\int_{G_0(x)}^1 \ln\left[-\ln \tau(t)\right]\mathrm{d}t \\ &= \mu\left(1 - G_0(x)\right) - \frac{\sigma}{1-G(0)}\int_{G(x)}^1 \ln\left[-\ln \tau\right]\mathrm{d}\tau. \end{aligned}$$

Hence, keeping in mind that

$$\int \tau \ln[-\ln \tau]\,\mathrm{d}\tau = \frac{\tau^2}{2}\ln[-\ln \tau] - \frac{1}{2}\mathrm{Ei}(2\ln \tau)\Big] \quad \text{and} \quad \int \ln[-\ln \tau]\,\mathrm{d}\tau = \tau \ln[-\ln \tau] - \mathrm{Ei}(\ln \tau)],$$

we obtain

$$\begin{aligned}\mathrm{CRPS}(G_0,\ x) = x(2G_0(x)-1) + \mu - 2\mu G_0(x) + \frac{2\sigma}{(1-G(0))^2}\Bigg\{\left[\frac{s^2}{2}\ln[-\ln s] - \frac{1}{2}\mathrm{Ei}(2\ln s)\right]_{s=G(0)}^{s=1} \\ -G(0)\left[(s\ln[-\ln s] - \mathrm{Ei}(\ln s)\right]_{s=G(0)}^{s=1} - (1-G(0))\left[s\ln[-\ln s] - \mathrm{Ei}(\ln s)\right]_{s=G(x)}^{s=1}\Bigg\}.\end{aligned}$$

Finally, since

$$\begin{aligned}&s^2\ln[-\ln s] - \mathrm{Ei}(2\ln s) - 2G(0)\,(s\ln[-\ln s] - \mathrm{Ei}(\ln s)) - 2\,(1-G(0))\,(s\ln[-\ln s] - \mathrm{Ei}(\ln s)) \\ &\quad = s^2\ln[-\ln s] - 2s\ln[-\ln s] - \mathrm{Ei}(2\ln s) + 2\mathrm{Ei}(\ln s) \\ &\quad = C - \ln 2 + (s-1)^2\ln[-\ln s] + \sum_{k=1}^{\infty}\frac{-(2\ln s)^k + 2(\ln s)^k}{k!k} \to C - \ln 2 \quad \text{as} \quad s \uparrow 1,\end{aligned}$$

the CRPS of a TGEV distribution with $\xi = 0$ equals

$$\begin{aligned}\mathrm{CRPS}(G_0, x) = (x-\mu)\,(2G_0(x)-1) + \frac{\sigma}{(1-G(0))^2} \\ \times\left(C - \ln 2 + \mathrm{Ei}\,(2\ln G(0)) + (G(0))^2\ln[-\ln G(0)] - 2G(0)\mathrm{Ei}\,(\ln G(0))\right) \\ + \frac{2\sigma}{1-G(0)}\left[G(x)\ln[-\ln G(x)] - \mathrm{Ei}\,(\ln G(x))\right].\end{aligned}$$

□

APPENDIX C. DEPENDENCE OF twCRPS OF SHORT-RANGE FORECASTS ON THE THRESHOLD

Beyond comparing the twCRPS values reported in Tables 2,4, and 6, one can get a deeper insight into the tail behavior of the different EMOS approaches by examining Figure C1 showing the twCRPSS with respect to the TN EMOS as function of the threshold. For the UWME forecasts (Figure C1(a)), GEV and TGEV models show very similar behavior and up to 13 m/s both approaches outperform the TN and LN EMOS methods. For lower threshold values TGEV EMOS results in the highest skill score, but after 8 m/s GEV shows the best predictive performance. A similar ranking of the methods can be observed in Figure C1(b); however, here the interval where the GEV and TGEV methods perform almost identically is much shorter. Finally, in the case of the ECMWF ensemble the TGEV EMOS results in the highest skill score for all thresholds, see Figure C1(c).

The difference between the first two cases and the third one in terms of the GEV and TGEV EMOS models might be explained with the difference in the support of these distributions. In the case of UWME and ALADIN-HUNEPS forecasts, the shape parameter ξ is negative for all forecast cases. Hence, the supports of GEV and TGEV predictive distributions are $]-\infty, \mu - \sigma/\xi]$ and $[0, \mu - \sigma/\xi]$, respectively, moreover, the upper bounds of the GEV are in general higher than those of the TGEV. In this way GEV can capture higher wind speeds, which results in better forecast skill in the upper tail. However, for the ECMWF ensemble the shape parameter of GEV and TGEV distributions is positive in 99.18% and 92.88% of all forecast cases, respectively, meaning that in these cases the supports of the predictive distributions are not bounded from above.

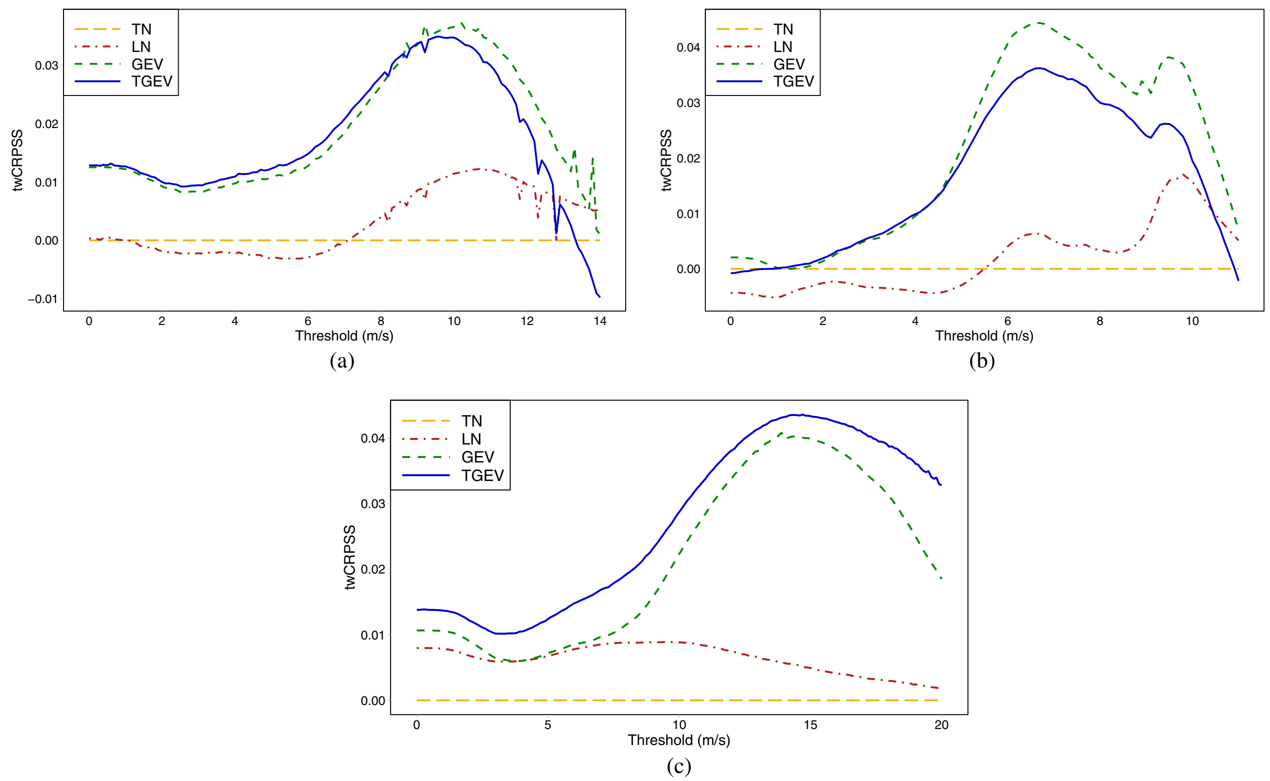


FIGURE C1 Threshold-weighted continuous ranked probability skill score values with respect to the truncated normal ensemble model output statistics model. (a) University of Washington mesoscale ensemble; (b) Aire Limitée Adaptation dynamique Développement International-Hungary Ensemble Prediction System ensemble; (c) European Centre for Medium-Range Weather Forecasts ensemble

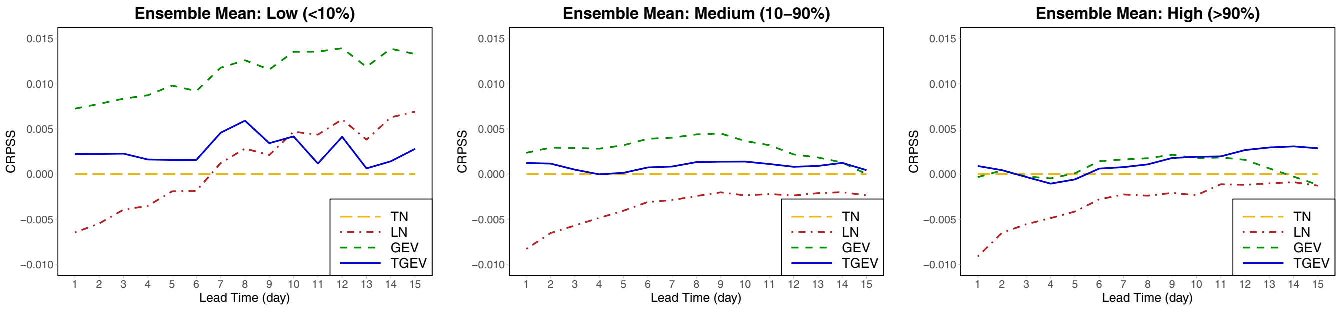
APPENDIX D. CALIBRATION AT DIFFERENT FORECAST LEVELS

To investigate calibration at different forecast levels, using the idea of Bremnes (2019), we group the forecast cases of verification data according to whether the ensemble mean is low (less than the 10th percentile of the means for the given lead time), medium (between 10 and 90 percentiles) or high (greater than the 90th percentile). As in terms of the mean CRPS all calibrated forecasts outperform the raw ensemble by a wide margin in all of our case studies, here we focus on the comparison of the competing EMOS approaches. Table D1 contains the CRPSS values with respect to the reference TN EMOS model at different levels for the short-range forecasts of Section 4.2. Note that for low and high forecasts the GEV and TGEV EMOS approaches outperform both the TN and LN EMOS models and the TGEV EMOS shows the best overall performance.

TABLE D1 Continuous ranked probability skill score values with respect to the truncate normal ensemble model output statistics model for forecasts with low (less than the 10th percentile), medium (between 10 and 90 percentiles) or high (greater than the 90th percentile) ensemble mean

	UWME			ALADIN-HUNEPS			ECMWF		
Forecast	**Low**	**Medium**	**High**	**Low**	**Medium**	**High**	**Low**	**Medium**	**High**
LN	0.000	0.000	−0.000	−0.031	−0.004	0.003	0.013	0.008	0.006
GEV	0.035	0.008	0.025	0.005	−0.001	0.018	0.027	0.008	0.016
TGEV	0.034	0.008	0.028	0.008	−0.001	0.020	0.019	0.010	0.028

Abbreviations: ALADIN-HUNEPS, Aire Limitée Adaptation dynamique Développement International-Hungary Ensemble Prediction System; ECMWF, European Centre for Medium-Range Weather Forecasts; GEV, generalized extreme value; TGEV, truncated GEV; UWME, University of Washington mesoscale ensemble.



(a) (b) (c)

FIGURE D1 Continuous ranked probability skill score values with respect to the truncate normal ensemble model output statistics model for forecasts with low (a), medium (b) or high (c) ensemble mean

A different behavior can be observed in Figure D1 showing the same skill score as function of the lead time for the global ECMWF forecasts investigated in Section 4.3. Here the GEV EMOS is the overall winner; however, for high wind speed forecasts the differences between the various EMOS approaches are rather small, especially for long forecast horizons. Note also that for the medium and high groups the ranking of the models is consistent with Figure 6(b).