Calibrated EMOS: applications to temperature and wind speed forecasting

Carlo Gaetan[1], Federica Giummolè[1] and Valentina Mameli[2*]

[1]Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, via Torino 155, Venice, 30172, Italy.

[2*]Department of Economics and Statistics, University of Udine, via Tomadini 30/A, Udine, 33100, Italy.

*Corresponding author(s). E-mail(s): valentina.mameli@uniud.it;
Contributing authors: gaetan@unive.it; giummole@unive.it;

Abstract

Ensembles of meteorological quantities obtained from numerical models can be used for forecasting weather variables. Unfortunately, such ensembles are often biased and under-dispersed and therefore need to be post-processed. Ensemble Model Output Statistics (EMOS) is a widely used post-processing technique to reduce bias and dispersion errors of numerical ensembles. In the EMOS approach, a full probabilistic prediction is given in the form of a predictive distribution with parameters depending on the ensemble forecast members. Parameters are then estimated and substituted, thus obtaining a so-called estimative predictive distribution. Nonetheless, estimative distributions may perform poorly in terms of the coverage probability of the corresponding quantiles. In this work, we suggest the use of calibrated predictive distributions based on a bootstrap adjustment of estimative predictive distributions, in the context of EMOS models. The corresponding calibrated quantiles give exact coverage probabilities. We evaluate the performance of the suggested calibrated EMOS in two simulation studies, comparing the different predictive distributions using the log-score, the continuous ranked probability score, and the coverage of the corresponding predictive quantiles. The results of these simulation studies show that the proposed calibrated predictive distributions improve estimative solutions, both reducing the mean scores and producing quantiles with exact coverage levels. The good performance of the new calibrated EMOS is further stressed in two real data applications, one about maximum

daily temperatures at sites located in the Veneto region (Italy) and the other one about wind speed forecasts at weather stations over Germany.

Keywords: calibration, Continuous Ranked Probability Score (CRPS), Ensemble Model Output Statistics (EMOS), log-score, predictive distribution

1 Introduction

In every field of knowledge, successful decisions need the support of accurate representations of the future. In particular, weather forecasts play a fundamental role nowadays, since meteorological conditions are of primary importance in almost all aspects of our lives. In the last decades, forecasts - in the form of Numerical Weather Predictions (NWP) ([7] and [32]) - have gradually improved their accuracy, mainly due to advances in technology and the coming of powerful computers. Despite this, simulated ensembles of forecasts based on physic models exhibit systematic bias and are often under-dispersive ([10], [23]).

To refine, improve, and calibrate NWP, statistical post-processing methods have been introduced in literature, including frequentist and Bayesian methods ([18, 31]). Among the most popular post-processing techniques, we focus on a parametric frequentist approach, the Ensemble Model Output Statistics (EMOS) ([18]). The EMOS is based on a heteroschedastic regression model, the parameters of which are determined by the ensemble forecasts. It is capable of reducing systematic biases and dispersion errors.

Different EMOS have been suggested in literature to model different weather quantities. For example, classic EMOS based on normal distribution may provide a reasonable model for temperature and pressure ([18]). To model high wind speed values, [34] propose an extended EMOS based on truncated normal distribution, [1] suggest log-normal distribution, [2] and [3] propose a combination of different EMOS models, and [5] and [25] use generalised extreme value distribution. [4] model rainfalls using censored and shifted gamma EMOS.

In all these applications, the unknown parameters of the EMOS model are estimated using past observations and then replaced to obtain a whole predictive distribution for the variable of interest. Despite its simplicity, this estimative approach does not take into account the uncertainty introduced by estimating the unknown parameters. As a result, estimative predictive distributions may be excessively concentrated, in particular when the number of past observations is small compared to the number of ensemble members.

This paper proposes an adjustment of the extended estimative EMOS based on a bootstrap calibration procedure introduced by [14]. The superiority of the bootstrap calibrated EMOS over the usual estimative EMOS is evaluated in different settings using suitable measures of calibration and sharpness, the most desirable properties that should characterise every predictive

model ([20]). In particular, we perform two simulation studies to evaluate and compare estimative and calibrated EMOS models, one with truncated normal and one with log-normal distributions. We assess the goodness of the considered models using the log-score, the Continuous Ranked Probability Score (CRPS), and the true coverage of the corresponding predictive quantiles. We then address the analyses of two real datasets. The first one regards maximum daily temperatures at measurement sites located in the Veneto region, northern Italy. The aim of this study is to explore more in depth the effect of bootstrap calibration in the context of classic EMOS, as already suggested in [16]. In the second application, we consider wind speed data for stations located in Germany. This data set includes the exchangeable 50-member ensemble of the European Center for Medium-range Weather Forecasts (ECMWF) and has been recently investigated in [11]. This example allows to more fully assess and compare the performance of various extended EMOS, with non-normal distributions. Our analyses show that calibrated EMOS is more accurate than estimative EMOS both in the presented simulation studies and in the applications. Moreover, they suggest the new technique's great potential in providing calibrated and sharp predictive models.

The paper is organised as follows. In Section 2 we outline the methodology used in this research. In Section 3 we study the performance of calibrated EMOS conducting two simulation studies on extended EMOS with various distributions. In Section 4 we introduce and analyse temperature data in Veneto (Italy) and we assess the superiority of calibrated classic EMOS in comparison with estimative classic EMOS. In Section 5 we present wind speed data for Germany, and we evaluate the superiority of calibrated extended EMOS versus estimative extended EMOS based on the truncated normal, the truncated logistic, the log-normal, and the generalized extreme value (GEV) distributions. Finally, in Section 6 we present some concluding remarks.

2 The method

In this section, we present our proposal which consists of a bootstrap procedure for calibration in the context of EMOS. We first recall some basics about EMOS and then we revise the bootstrap calibration method.

2.1 Ensemble Model Output Statistics

EMOS produces probabilistic forecasts of weather variables by pooling together the raw ensembles in a parametric predictive distribution with parameters depending on the ensemble forecast members ([18]). In its basic version, EMOS is nothing but a normal linear regression model with heteroschedastic errors. The EMOS mean is a linear combination of the ensemble member forecasts, with unknown coefficients that represent the contributions of each member of the ensemble to the relevant weather variable. The EMOS variance is a linear function of the ensemble variance that accounts for the spread relationship. Formally, it is assumed that the weather variable Z depends on the

ensemble forecasts $X_1, \ldots, X_m$ in such a way that

$$Z = \beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m + \varepsilon,$$

where ε is a normally distributed error term with 0 mean and variance $\sigma^2 = \gamma_0 + \gamma_1 S^2$ to account for dispersion errors in the ensemble members. Here, $S^2 = \sum_{j=1}^{m}(X_j - \bar{X})^2/(m-1)$ denotes the ensemble variance and $\bar{X} = \sum_{j=1}^{m} X_j/m$ the ensemble mean. The parameters $\beta_0, \ldots, \beta_m$, γ_0 and γ_1 are non-negative unknown coefficients. The distribution of Z is given by

$$Z \sim \Phi\left(\frac{z-\mu}{\sigma}\right), \tag{1}$$

with mean $\mu = \beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m$ and variance $\sigma^2 = \gamma_0 + \gamma_1 S^2$, where $\Phi(\cdot)$ denotes the standard normal distribution function. In the sequel, we refer to model (1) as the classic EMOS.

Classic EMOS can also be extended beyond the normal case, allowing for skewed or heavier tail distributions like log-normal, truncated normal, gamma, and generalised extreme value distributions. The unknown parameters of the chosen distribution for Z are then written as suitable functions of the ensemble members $X_1, \ldots, X_m$. We call all these models extended EMOS, in contrast to classic EMOS (1). Two examples of the application of EMOS with log-normal and truncated normal distributions are considered in the simulation section of this paper and the application to wind speed data, together with the truncated logistic and generalised extreme value distributions.

The unknown parameters of EMOS are usually estimated by minimising proper scoring rules such as the log-score and the CRPS. Minimisation of the log-score corresponds to the well-known Maximum Likelihood Estimator (MLE). The CRPS is given by the general formula:

$$\mathrm{CRPS}(F, x) = \int_{\mathbb{R}} [F(u) - \mathbb{I}(u \geq x)]^2 du,$$

where F is a predictive distribution function to be evaluated at the observed value x and $\mathbb{I}(A)$ denotes the indicator function of the set A. Minimisation of the CRPS gives rise to the minimum CRPS estimator, with good robust properties and prediction ability, see [19]. The scores are minimised by using the ensemble forecasts and the corresponding observed values referring to suitably chosen training periods ([18]). Training sets are selected basically by using two approaches: the local and the regional methods. In the local approach, only observations from a single station of interest are considered for parameter estimation, while in the regional approach observations from all available stations are considered. Although local estimation generally yields better predictive performance, it may suffer from numerical instability due to the limited availability of training data ([34]). In contrast, regional estimation has typically no numerical instability issues, but in such conditions, a single set of parameters

is found for all the stations, without taking into account geographical and climatological variability ([1]). An intermediate solution is proposed by [26] with similarity-based semilocal models to estimate the EMOS coefficients. Here, we limit our discussion to the local estimation approach since our proposal overcomes the problem of numerical instability due to small sets of training data.

In the classic EMOS, after minimising the log-score or the CRPS, the estimated parameters are replaced in (1) obtaining what is known as an estimative distribution for the future weather quantity Z: $\Phi\left((z-\hat{\mu})/\hat{\sigma}\right)$, with $\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_m X_m$, and $\hat{\sigma}^2 = \hat{\gamma}_0 + \hat{\gamma}_1 S^2$, where $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m, \hat{\gamma}_0$, and $\hat{\gamma}_1$ are the estimates of $\beta_0, \beta_1, \ldots, \beta_m, \gamma_0$, and γ_1, respectively. A similar procedure is easily applied for obtaining estimative predictive distributions in the case of extended EMOS.

Unfortunately, estimative distributions can perform poorly, particularly when the number of past observations is small in comparison to the number of ensemble members because estimates can be highly unstable in this case. In particular, the calibration requirement is not met by estimative distributions. In fact, the estimative procedure does not account for the variability introduced by substituting fixed parameter values with estimates. Thus, estimative distributions are often under-dispersed and too sharp.

2.2 Calibrated predictive distributions

There are many different properties that a good predictive distribution should possess. As suggested in [20], here we focus on a calibration that is a sort of consistency between a predictive distribution and future observations. It is based on the fact that a good predictive distribution $\hat{F}(z)$ should resemble the true distribution $F(z)$ so that, for the integral transform theorem, $\hat{F}(Z) \sim U(0,1)$, at least approximately, where $U(0,1)$ denotes a uniform distribution in $(0,1)$. The PIT (Probability Integral Transform) histogram is a graphical representation useful for checking calibration ([31]). For the construction of a PIT histogram, each observed data z is transformed through the predictive distribution $\hat{F}(\cdot)$, and then the histogram of transformed values $\hat{F}(z)$ is displayed. The histogram should be flat and similar to the histogram of random values from a uniform distribution in $(0,1)$.

It can be shown that a predictive distribution whose quantiles give the correct coverage probability is always calibrated. Thus, in this section, we briefly review the calibrating approach proposed by [14], which provides predictive distributions whose quantiles give well-calibrated coverage probabilities. The approach has recently been adapted to the EMOS context in [16], where only the classic EMOS (1) has been considered.

Suppose that $\{Z_i\}_{i\geq 1}$ is a sequence of independent continuous random variables. We assume that $Z^{(n)} = (Z_1, \ldots, Z_n)$, $n > 1$, is observable, while $Z = Z_{n+1}$ is a future or not yet available variable of the process, with probability distribution $F(z;\ \theta)$ depending on an unknown parameter θ. This general setting includes the basic EMOS specified in (1) and all the extended EMOS

as particular cases. We indicate with $z_\alpha(\theta)$ the α-quantile of Z, so that $z_\alpha(\theta) = F^{-1}(\alpha;\ \theta)$. Given the observed sample $z^{(n)} = (z_1, \ldots, z_n)$, an α-prediction limit for Z is a function $c_\alpha(z^{(n)})$ such that, exactly or approximately,

$$P_{Z^{(n)},Z}\{Z \leq c_\alpha(Z^{(n)});\ \theta\} = \alpha,$$

for every $\theta \in \Theta$ and for every fixed $\alpha \in (0, 1)$. The above probability is called coverage probability and it is calculated with respect to the joint distribution of $(Z^{(n)}, Z)$.

Consider a suitable asymptotically efficient estimator $\hat{\theta} = \hat{\theta}(Z^{(n)})$ for θ and the estimative prediction limit $z_\alpha(\hat{\theta})$, which is obtained as the α-quantile of the estimative distribution function $F(z;\ \hat{\theta})$. The associated coverage probability is

$$P_{Z^{(n)},Z}\{Z \leq z_\alpha(\hat{\theta}(Z^{(n)}));\ \theta\} = E_{Z^{(n)}}[F\{z_\alpha(\hat{\theta}(Z^{(n)}));\ \theta\};\ \theta] = C(\alpha, \theta)$$

and, although its explicit expression is rarely available, it is well-known that it does not match the target value α even if, asymptotically, $C(\alpha, \theta) = \alpha + O(n^{-1})$, as $n \to +\infty$, see e.g. [6]. As proved in [14], the function

$$F_c(z;\ \hat{\theta}, \theta) = C\{F(z;\ \hat{\theta}), \theta\}, \tag{2}$$

which is obtained by substituting α with $F(z;\ \hat{\theta})$ in $C(\alpha, \theta)$, is a proper predictive distribution function, provided that $C(\cdot, \theta)$ is a sufficiently smooth function. Furthermore, it gives, as quantiles, prediction limits $z^c_\alpha(\hat{\theta}, \theta)$ with coverage probability equal to the target nominal value α, for all $\alpha \in (0, 1)$.

The calibrated predictive distribution (2) is not useful in practice, since it depends on the unknown parameter θ. However, a suitable parametric bootstrap estimator for $F_c(z;\ \hat{\theta}, \theta)$ may be readily defined. Let $\hat{\theta}^b$, $b = 1, \ldots, B$, be estimates obtained from B bootstrap samples generated from the estimative distribution of the data. Since $C(\alpha, \theta) = E_{Z^{(n)}}[F\{z_\alpha(\hat{\theta}(Z^{(n)}));\ \theta\};\ \theta]$, we define the bootstrap calibrated predictive distribution as

$$F_c^{boot}(z;\ \hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B} F\{z_\alpha(\hat{\theta}^b);\ \hat{\theta}\}\ |_{\alpha = F(z;\ \hat{\theta})}\ . \tag{3}$$

The corresponding α-quantile defines, for each $\alpha \in (0, 1)$, a prediction limit having coverage probability equal to the target α, with an error term that depends on the efficiency of the bootstrap simulation procedure. This makes $F_c^{boot}(z;\ \hat{\theta})$ a well calibrated predictive distribution for Z.

In the following, we show that the proposed bootstrap adjustment on the EMOS estimative distributions significantly outperforms the estimative EMOS both in terms of calibration and sharpness, the most desirable properties that characterise predictive models ([20]).

3 Simulation studies

In this section, we present two simulation studies to compare estimative predictive distributions with their calibrated counterparts, in the context of EMOS with log-normal and truncated normal distributions. The classic EMOS with normal errors has already been considered in [16]. Both the considered models are estimated with the R package `ensembleMOS` ([35]). For the optimisation of the log-score and of the CRPS over the training data, we use the constrained optimisation algorithm L-BFGS-B ([8]). In both simulations, we have chosen a small training sample size with a quite high number of ensemble members. This is a setting where estimates of the unknown parameters suffer instability due to a small number of observations. Typically, in this situation the estimative distribution is under-dispersed with U-shaped PIT histograms. Indeed, this is where the bootstrap calibration is more compelling.

3.1 Log-normal EMOS

In [1] an EMOS approach based on the log-normal distribution is proposed for modelling wind speed values. The density of the log-normal distribution with parameters μ and $\sigma > 0$ is

$$f(z;\ \mu, \sigma) = \frac{1}{\sigma}\phi\left(\frac{\log z - \mu}{\sigma}\right), \qquad z > 0,$$

where $\phi(\cdot)$ denotes the density of a standard normal distribution. The mean m and the variance v of the interest variable Z are related to μ and σ through the equations $m = e^{\mu+\sigma^2/2}$ and $v = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$, respectively. In the log-normal EMOS proposed by [1] m and v are affine functions of the ensemble members and the ensemble variance, respectively:

$$m = \beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m \quad \text{and} \quad v = \gamma_0 + \gamma_1 S^2, \tag{4}$$

where $\beta_0 \in \mathbb{R}$ and $\beta_1, \ldots, \beta_m, \gamma_0, \gamma_1 \geq 0$. Model parameters $\beta_0, \ldots, \beta_m$ and γ_0, γ_1 may be estimated by optimising the log-score or the CRPS over the training data. Here, we show the results of a simulation study based on $M = 5000$ Monte Carlo replications, with $B = 200$ bootstrap samples for calibration. The sample size, that is the length of the sliding window of training observations, is $n = 25$ with $m = 10$ ensemble members. We have simulated 5025 outcomes of the ensemble, using a multivariate normal distribution for the log-transformed ensemble members, with mean 0 and variance 1 for each component, and pairwise correlation $\rho = 0.75$. The same number of observations for a weather variable following the log-normal EMOS have been generated with regression coefficients set to $\beta_j = j + 1$, $j = 0, \ldots, 10$, and $\gamma_0 = 100$, $\gamma_1 = 100$. We report the PIT histograms (Figure 1), the mean values of the log-score and the CRPS (Table 1) and the coverage probabilities of upper limits of level $\alpha = 0.9$, 0.95, and 0.99 (Table 2), for the estimative distributions obtained with the MLE

and the minimum CRPS estimator, and the corresponding calibrated versions. All the results show the improvement of the calibrated procedures over the estimative ones. We have repeated the simulation study using different correlations between the ensemble members. The results, not reported here, are not affected by this choice and always show the improvement of the bootstrap calibrated procedure over the estimative one. We have also repeated the study varying the sample size and the number of ensemble members. The results, not presented here, show a better improvement when the sample size n is small with respect to the number of ensembles m.

Table 1 Log-normal EMOS: Average log-score and CRPS values for the four predictive distributions. Standard errors in brackets. Est log denotes the estimative EMOS with MLE estimates and Est CRPS the estimative EMOS with CRPS estimates, while Cal log and Cal CRPS are the respective calibrated counterparts.

	Est log	Cal log	Est CRPS	Cal CRPS
Log-score	10.45 (0.81)	4.84 (0.16)	14.65 (1.27)	4.90 (0.12)
CRPS	14.34 (0.33)	12.79 (0.27)	15.09 (0.35)	13.15 (0.28)

Table 2 Log-normal EMOS: coverage probabilities of upper prediction limits for the four predictive distributions. Standard errors in brackets. Est log denotes the estimative EMOS with MLE estimates and Est CRPS the estimative EMOS with CRPS estimates, while Cal log and Cal CRPS are the respective calibrated counterparts.

α	Est log	Cal log	Est CRPS	Cal CRPS
0.90	0.739 (0.006)	0.887 (0.004)	0.715 (0.006)	0.894 (0.004)
0.95	0.797 (0.006)	0.936 (0.003)	0.772 (0.006)	0.937 (0.003)
0.99	0.872 (0.005)	0.977 (0.002)	0.848 (0.005)	0.975 (0.002)

3.2 Truncated normal EMOS

[34] propose a truncated normal model to model wind speed. The truncated normal distribution with location μ, scale $\sigma > 0$, and lower truncation at 0, has density function

$$f(z;\ \mu, \sigma) = \frac{1}{\sigma}\phi\left(\frac{z-\mu}{\sigma}\right) / \Phi\left(\frac{\mu}{\sigma}\right), \quad z > 0,$$

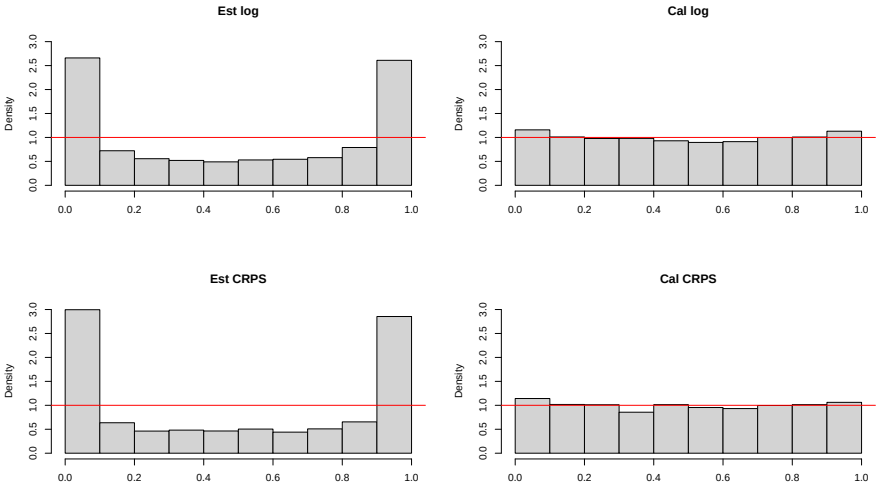


Fig. 1 Log-normal EMOS: PIT histograms of the four predictive distributions. Est log denotes the estimative EMOS with MLE estimates and Est CRPS the estimative EMOS with CRPS estimates, while Cal log and Cal CRPS are the respective calibrated counterparts.

where $\phi(\cdot)$ is the density function and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. In the truncated normal EMOS, the location and scale are linked to the ensemble members through the following formulas

$$\mu = \beta_0 + \beta_1 X_1 + \ldots + \beta_m X_m \quad \text{and} \quad \sigma^2 = \gamma_0 + \gamma_1 S^2, \tag{5}$$

where $\beta_0 \in \mathbb{R}$ and $\beta_1, \ldots, \beta_m, \gamma_0, \gamma_1 \geq 0$. Again model parameters $\beta_0, \beta_1, \ldots, \beta_m$ and γ_0, γ_1 can be estimated by optimising the log-score and the CRPS over the training data.

In order to assess and compare the performance of the estimative and the calibrated predictive distributions we have performed several experiments with simulated ensembles. The ensemble members are drawn from a 10-variate truncated normal distribution with location 0 and scale 1 for each component, and correlation $\rho = 0.75$ between pairs of the ensemble members. The observations are generated from a truncated normal random variable with parameters specified in (5) with $\beta_j = j + 1$, $j = 0, \ldots, 10$, and $\gamma_0 = 0$, $\gamma_1 = 1$. The sample size is $n = 25$ and the bootstrap calibrating procedure is based on 200 bootstrap samples. The number of Monte Carlo replications is 5000. We evaluate the estimative and calibrated predictive distributions in terms of coverage probabilities, PIT histograms, and also using the mean log-score and CRPS.

Table 3 provides the results of the simulation study for comparing coverage probabilities of upper limits of level $\alpha = 0.9$, 0.95, and 0.99 obtained from the estimative and the calibrated distributions with minimum CRPS and

maximum likelihood estimates. It can be noted that the coverage probabilities associated with the calibrated quantiles are very accurate, being almost equal to the nominal values. The same conclusions can be drawn from the PIT histograms (Figure 2). We also assess the improvement of the calibrated predictive distributions over the estimative ones by computing the log-score and the CRPS, averaged over the 5000 replicates, as shown in Table 4. The superior performance of the calibrated distributions is evident. Indeed, average values of the scores for estimative distributions are higher with respect to their calibrated counterparts. As in the previous example, we do not report the results of other simulation studies performed by using different settings. However, these results indicate that when the sample size is small with respect to the number of ensembles, the improvement of the calibrated predictive distribution on the estimative one is more evident.

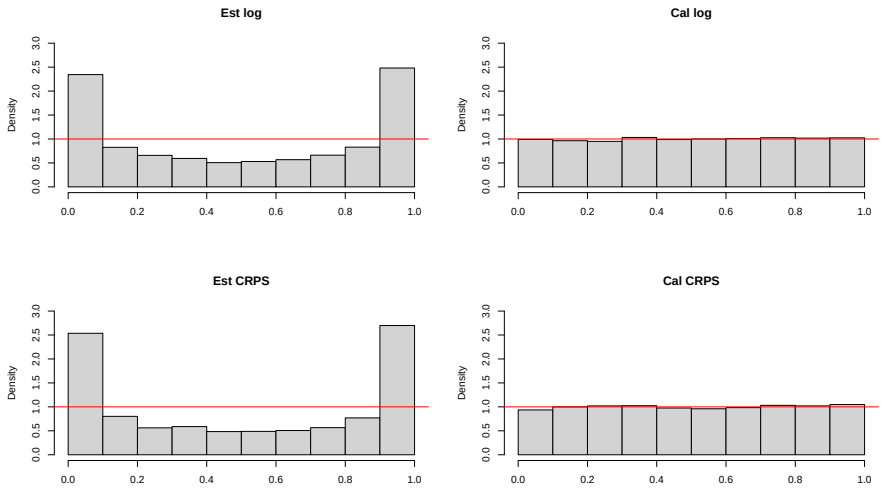


Fig. 2 Truncated normal EMOS: PIT histograms of the four predictive distributions. Est log denotes the estimative EMOS with MLE estimates and Est CRPS the estimative EMOS with CRPS estimates, while Cal log and Cal CRPS are the respective calibrated counterparts.

4 Temperature forecasts in Veneto

In order to assess and compare the performance of different EMOS predictive distributions, we analyse maximum daily temperatures for stations located throughout the Veneto region in the northeast of Italy, see Figure 3.

Table 3 Truncated normal EMOS: coverage probabilities of upper prediction limits for the four predictive distributions. Standard errors in brackets. Est log denotes the estimative EMOS with MLE estimates and Est CRPS the estimative EMOS with CRPS estimates, while Cal log and Cal CRPS are the respective calibrated counterparts.

α	Est log	Cal log	Est CRPS	Cal CRPS
0.90	0.752 (0.006)	0.898 (0.004)	0.730 (0.006)	0.895 (0.004)
0.95	0.805 (0.006)	0.948 (0.003)	0.785 (0.006)	0.945 (0.003)
0.99	0.889 (0.005)	0.987 (0.002)	0.866 (0.005)	0.987 (0.002)

Table 4 Truncated normal EMOS: Average log-score and CRPS values of the four predictive distributions. Standard errors in brackets. Est log denotes the estimative EMOS with MLE estimates and Est CRPS the estimative EMOS with CRPS estimates, while Cal log and Cal CRPS are the respective calibrated counterparts.

	Est log	Cal log	Est CRPS	Cal CRPS
Log-score	1.88 (0.05)	1.02 (0.02)	2.37 (0.07)	1.05 (0.02)
CRPS	0.39 (0.005)	0.36 (0.004)	0.40 (0.005)	0.37 (0.004)

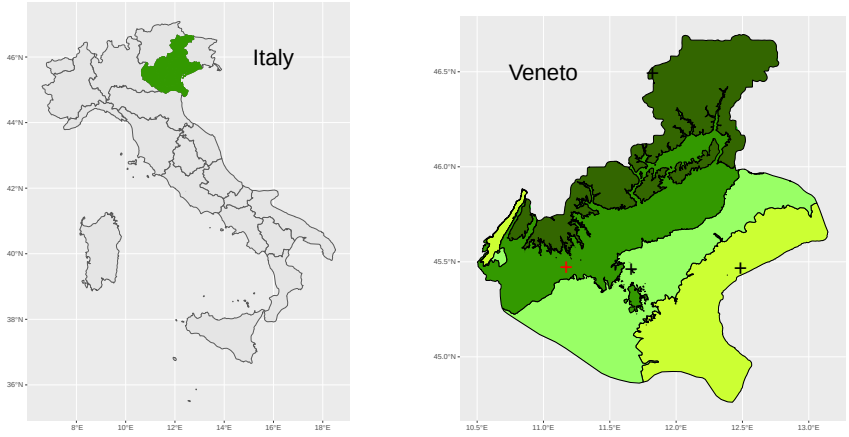
4.1 Data description

Two sources of information about maximum daily temperatures are used in this application: ground measurements and numerical forecasts. The first includes historical maximum daily temperatures provided by the Italian national system for the collection, processing, and dissemination of climate data, created by ISPRA (http://www.scia.isprambiente.it/). The second source consists of numerical forecasts (the ensemble predictions) available from the Earth System Grid Federation (https://esgf-node.llnl.gov/search/cmip6/, last accessed on February 2022). We use the World Climate Research Programme's Coupled Model Intercomparison Project Phase 6 system (CMIP6). The project delivers a huge number of simulations from global climate models at high spatial resolution; in fact, it comprises over 120 global climate models and approximately 45 universities and organizations globally (https://pcmdi.llnl.gov/CMIP6). One of the scientific focuses of the CMIP6 experiment is to understand past, present, and future climate changes ([12]). The CMIP6 models used for this study are given in the Supplementary Material. Although some CMIP6 models have a large number of members, we use a single member for each CMIP6 model as in [24]. Therefore in this application, each CMIP6 model is considered a single member of our ensemble. Thus, all ensemble members have individually distinguishable physical features and are not exchangeable. The ISPRA historical datasets are available from 1850, but for

evaluation purposes, this study focuses on the period 2009-2012 to match the timespan of CMIP6 numerical simulations. Ground measurement data from ISPRA are used as benchmarks and are collected at different meteorological stations in the Veneto region.

The region, which is located in northern Italy, is characterised by large elevation variations, with a mountainous area in the northwestern part, an intermediate hill zone in the middle, and a broad flat area in the southeastern part. Its elevation varies from sea level (and also below sea level) to around 3,300 meters, resulting in a wide range of temperatures. The elevation is used in Figure 3 (right panel) to classify the various zones of the Veneto region based on its quartile division, where higher elevation areas are represented in darker tones. Numerical predictions are then interpolated to the station level using elevation as a reference.

Fig. 3 Left panel: Geographical location of studied area in Italy. Right panel: Location of the meteorological stations in the Veneto region. Crosses represent stations, while colors outline the four elevation zones (elevation quartile base division). The darker the tone, the higher the elevation. The red cross represents the Illasi station.

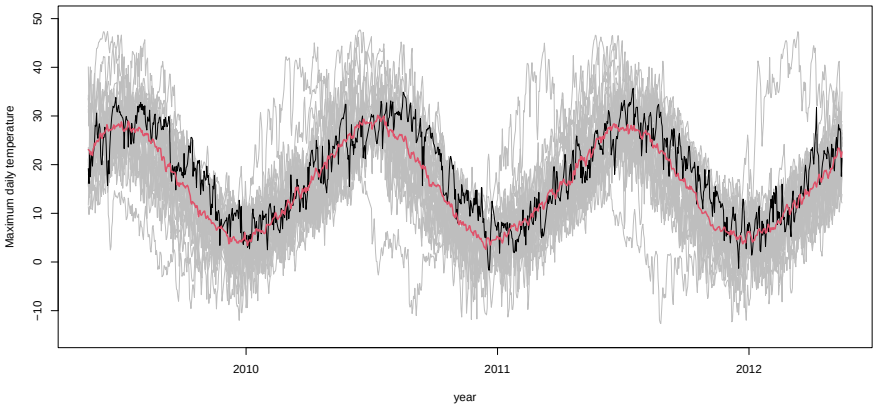


4.2 Analysis and results for the Illasi station

We have considered four stations, see crosses in Figure 3 (right panel), one for each of the four zones identified by the elevation quartile-based division. We report here only the analysis of the Illasi station (Longitude: 11.17178°, Latitude: 45.45954°) represented by the red cross in Figure 3 (right panel), since for all the considered stations we have observed similar behaviors. Data from Cavallino Treporti station have already been used in [17] as a real case application in the context of confidence predictive distributions.

All CMIP6 climate models considered in this study show bias, namely systematic difference between historical ground measurements and numerical simulations, as can be observed in Figure 4. In this figure, the black line is

Fig. 4 Temperature case study: differences between historical observations and numerical forecasts. Time series of each numerical forecast from one CMIP6 model (gray lines) and the corresponding ISPRA historical observations (black line) together with the numerical forecasts obtained by the ensemble mean (red line).



the time series of the true historical maximum daily temperatures at Illasi station collected from the ISPRA website and used as benchmarks. Each grey line represents the time series of numerical forecasts from one CMIP6 model (the list of which is given in the Supplementary Material). The red line is the time series of the numerical forecast obtained by averaging the forecasts from the CMIP6 models (ensemble mean). The data cover a period of 3 years from 16 May 2009 to 15 May 2012. After removing missing observations from the selected station, the sample contains 1079 daily temperature observations and 26 ensemble members. Following [18], we consider a sliding window of 40 observations as the training set. The remaining 1039 days will serve as a test set. A short time period of training observations allows us to avoid considering seasonality, since in such a short temporal window the generating process can be assumed as stationary. The classic EMOS with normal distribution often provides a reasonable model for temperatures ([18]). This is also the case for the considered data.

The EMOS parameters are estimated by optimising both the log-score and the CRPS over the sliding training period. Then the performance of the two estimative distributions derived from the log-score and the CRPS, as well as their bootstrap calibrated counterparts computed as in (3) are evaluated from 1 day up to 10 days ahead on the test set.

The different predictive models are compared at each lead time in terms of the log-score and the CRPS. Figure 5 shows average log-score (left) and CRPS (right) values at each lead time for the predictive EMOS models (the smaller the better). The two calibrated EMOS result in the lowest average log-score and CRPS values, for all lead times, significantly outperforming their estimative competitors. We also evaluate the performances of the four predictive

models in terms of the coverage probability of central intervals of level 0.67 and the coverage probabilities of upper prediction limits of levels 0.90, 0.95, and 0.99; see Figure 6. It can be seen that the two calibrated EMOS result in the best coverage for each target nominal level. They are much closer to the nominal coverage level than the estimative EMOS. The PIT histograms of calibrated EMOS forecasts, not presented here, show the positive effect of calibration, already shown in Figure 6. They are much closer to uniformity than the PIT histograms of the estimative EMOS confirming the results obtained with the coverage probabilities.

The normal EMOS models have been estimated with the R package `ensembleMOS` [35]. For the optimization of the log-score and of the CRPS over the training data, we have used the optimisation algorithm BFGS ([9, 13, 21, 33]).

Fig. 5 Temperature case study: Log-score (left) and CRPS average values (right) for the four predictive distributions on different days. Est log denotes the estimative EMOS with MLE estimates and Est CRPS the estimative EMOS with CRPS estimates, while Cal log and Cal CRPS are the respective calibrated counterparts.

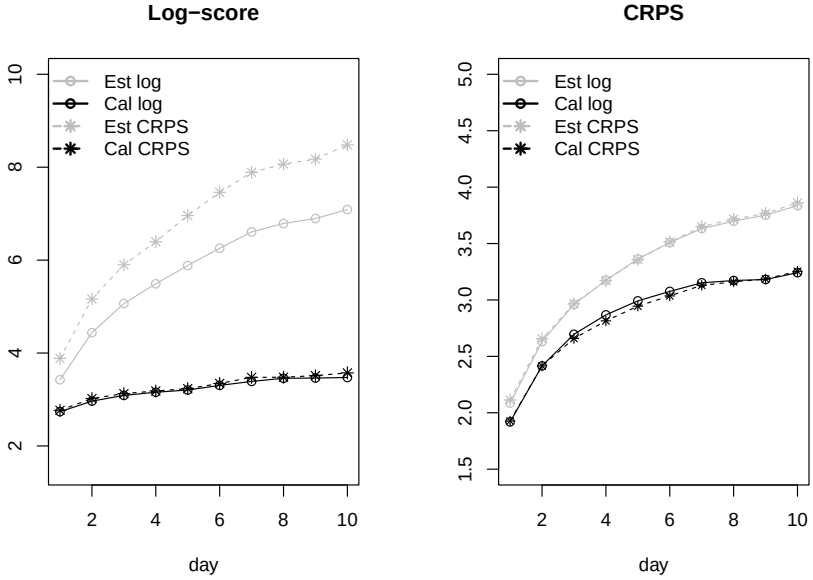
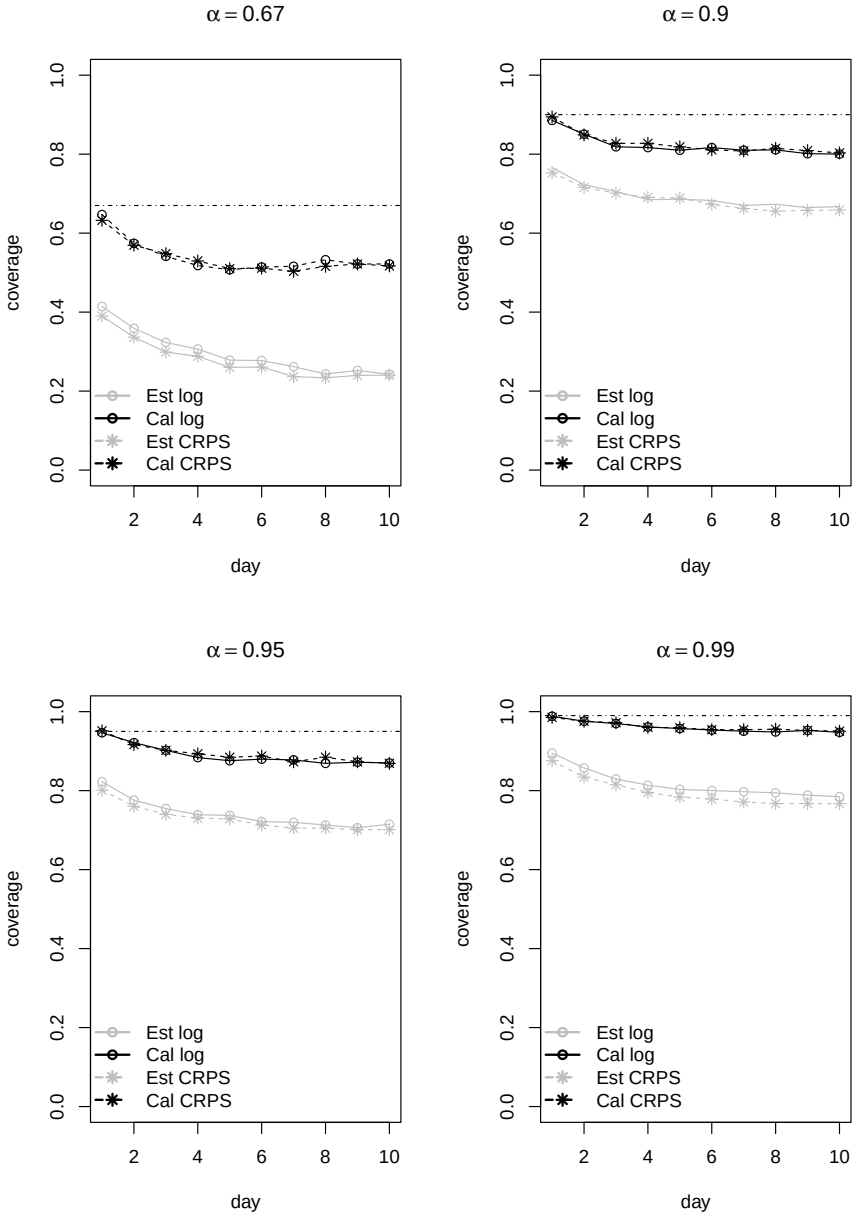


Fig. 6 Temperature case study: Coverage probabilities for the four predictive distributions on different days for different target nominal levels $\alpha = 0.67, 0.90, 0.95, 0.99$. The ideal coverage is indicated by the horizontal dashed-dotted line in each plot. Est log denotes the estimative EMOS with MLE estimates and Est CRPS the estimative EMOS with CRPS estimates, while Cal log and Cal CRPS are the respective calibrated counterparts.



5 Wind speed forecasts in Germany

We examine wind speed data for stations in Germany in order to more thoroughly evaluate and contrast the performance of various extended EMOS based on non-normal distributions.

5.1 Data description

This dataset has been recently studied by [11] and is available from https://doi.org/10.6084/m9.figshare.19453622. It consists of forecasts of daily 10-meter wind speed in 198 weather stations located all over Germany, produced by the 50-member ensemble of the European Center for Medium-range Weather Forecasts (ECMWF). The dataset also contains historical observations from the Climate Data Center of the German weather service. In contrast with the previous case-study, ensemble members in this application can be thought of as exchangeable because they lack distinguishing physical characteristics. The mean and standard deviation of the ensemble forecasts are then determined and used to specify model parameters. A total of 10 years of daily forecasts and observations ranging from the 2007 to the 2016 are available. This provides a rich dataset for investigating the performance of the ensemble forecasting methods. See [11] for further details about the data.

5.2 EMOS models for wind speed

In the following, we apply the calibration procedure presented in Section 2.2 to different extended EMOS for daily wind speed forecast in Germany. We consider extended EMOS already proposed in the literature, such as those based on the normal distribution left-truncated at zero [34], the logistic distribution left-truncated at zero [28, 29], the log-normal distribution [1], and the generalized extreme value distribution (GEV) [5, 25].

Usually, for all these models the unknown parameters are linked to the ensemble members, as for instance in equations (4) and (5); however, in the present case study, ensemble members are exchangeable, so unknown parameters are written as functions of the ensemble mean $\bar{X}$ and the ensemble variance S^2. In particular, for the normal and the logistic distributions left-truncated at zero we have modeled the location parameter as

$$\mu = \beta_0 + \beta_1 \bar{X}, \tag{6}$$

where $\beta_0 \in \mathbb{R}$, $\beta_1 \geq 0$. The variance is a linear function of the ensemble variance, as already specified in (5). Similarly, for the log-normal distribution, the mean has been modeled as a linear function of the ensemble mean $\bar{X}$, and the variance as in (4). In the extended EMOS with the GEV distribution, the location parameter is specified as in (6), the logarithm of the scale parameter is considered as a linear function of the logarithm of the ensemble mean, that is $\log \sigma = \gamma_0 + \gamma_1 \log \bar{X}$, with $\gamma_0, \gamma_1 \in \mathbb{R}$, and the shape parameter is an unknown constant. Different ways for modelling the scale parameter have also

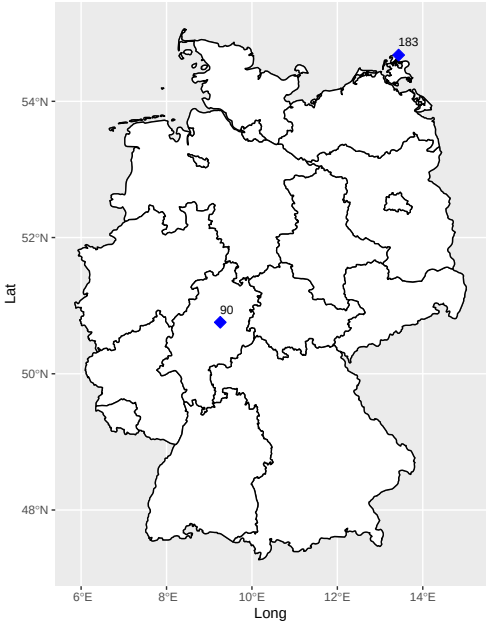
been considered, but the final results do not seem to be much affected by this choice, as also mentioned in [5]. In this case, it is possible to obtain non-zero probabilities of negative wind speed. However, this rarely happens in this dataset.

In the following sections, the truncated logistic and the truncated normal EMOS models are estimated using the R package `crch` [27, 30]. Instead, the log-normal and the GEV EMOS models are estimated using the R package `ensembleMOS` [35], with the L-BFGS-B optimisation algorithm, and `extRemes` [15], respectively.

5.3 Wind speed forecasts for stations 90 and 183

Here we only report the analysis of the two stations shown in Figure 7: station 90, located in the center of Germany (Longitude: 9.2583, Latitude: 50.7557), and station 183, located in the north of Germany (Longitude: 13.4343, Latitude: 54.6792). These two stations have been selected as examples of different behavior in the distribution of wind speed.

Fig. 7 Location in Germany of stations 90 and 183 considered in this study.



5.3.1 Station 90

The sample consists of 3576 observations of daily wind speed. The training set is a sliding window with 25 observations, and the test set consists of the remaining days. Here, we take into account extended EMOS with the GEV

distribution ([5, 25]), and with the log-normal distribution ([1]). For both the EMOS models, the parameters are calculated by maximising the log-score over a training set of 25 observations. The performance of the two estimative distributions obtained with the log-score — one based on the EMOS with the log-normal distribution and the other with the GEV distribution — as well as the corresponding calibrated counterparts obtained by the bootstrap procedure (3) are then assessed using coverage probabilities for each of the days available in the test set. In particular, we consider coverage probabilities of central intervals of level 0.67 (Table 5) to assess the calibration and sharpness in the central part of the predictive distributions. The log-normal and the GEV distributions show similar results. Additionally, we also consider the coverage probabilities of upper prediction limits of levels 0.90, 0.95, and 0.99 (Table 5). The findings indicate that when compared to estimative models, calibrated predictive models have better coverage probabilities for both central intervals and upper prediction limits. The PIT histograms in Figure 8, with the log-normal distribution at the top and the GEV distribution at the bottom, further support the superior performance of the calibrated models. For this station, we have also considered extended EMOS with the truncated normal and truncated logistic distributions, but the results are unsatisfactory because the two distributions do not adequately fit the data. In fact, the proposed calibration procedure is only effective under a good model specification.

Fig. 8 Wind case study for station 90. a) Log-normal distribution and b) GEV distribution. PIT histograms of the estimative EMOS with MLE estimates (Est log), and the respective calibrated counterpart (Cal log).

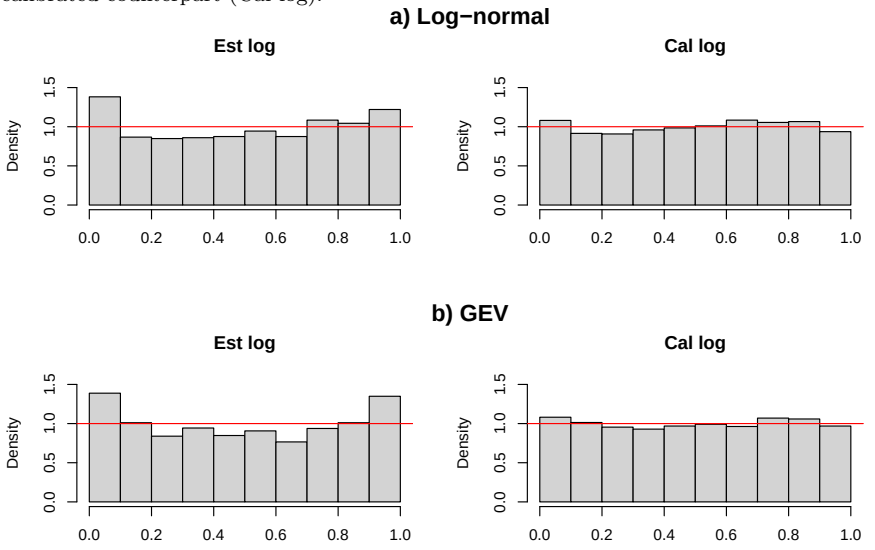


Table 5 Wind case study for station 90. a) Log-normal distribution and b) GEV distribution. Coverage probabilities of the central interval of level 0.67 and upper prediction limits for the estimative EMOS with MLE estimates (Est log), and the respective calibrated counterpart (Cal log). Standard errors in brackets.

a) Log-normal

α	Est log	Cal log
0.67	0.613	0.667
	(0.008)	(0.008)
0.90	0.878	0.906
	(0.005)	(0.005)
0.95	0.935	0.956
	(0.004)	(0.003)
0.99	0.981	0.992
	(0.002)	(0.002)

b) GEV

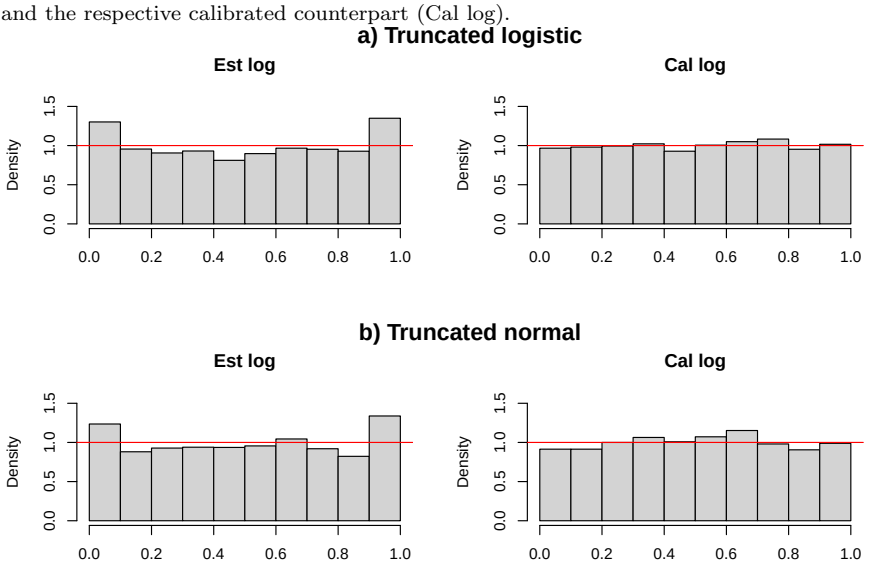
α	Est log	Cal log
0.67	0.593	0.652
	(0.008)	(0.008)
0.90	0.865	0.903
	(0.006)	(0.005)
0.95	0.916	0.953
	(0.005)	(0.004)
0.99	0.959	0.979
	(0.003)	(0.002)

5.3.2 Station 183

The sample contains 3610 observations of daily wind speed. We use a sliding window of 25 observations as a training set, with the remaining days available as a test set. Here, we consider extended EMOS with the normal distribution left-truncated at zero [34], and with the logistic distribution left-truncated at zero [28, 29]. The EMOS parameters for both models are estimated by optimising the log-score over the sliding training period. The performance of the two estimative distributions obtained with the log-score — one based on the EMOS with the normal distribution left-truncated at zero and the other with the logistic distribution left-truncated at zero — as well as the corresponding calibrated distributions obtained using the bootstrap procedure (3) are evaluated in terms of coverage probabilities of central intervals of level 0.67 (Table 6), and upper prediction limits of levels 0.90, 0.95, and 0.99 (Table 6). The truncated normal and the truncated logistic distributions show similar results. It is important to remark that the coverage probabilities for calibrated

predictive models for both the truncated logistic and the truncated normal distributions are much closer to the nominal values than those for the corresponding estimative models. The PIT histograms for the four investigated predictive models are finally shown in Figure 9, with the truncated logistic distribution at the top and the truncated normal distribution at the bottom. The U-shaped histograms of the estimative models are due to the excessive underdispersion. Instead, the effect of calibration results in a flat PIT histogram, very close to the uniform one.

Fig. 9 Wind case study for station 183. a) Truncated logistic distribution and b) Truncated normal distribution. PIT histograms of the estimative EMOS with MLE estimates (Est log), and the respective calibrated counterpart (Cal log).



6 Conclusions

In this work, we compare the estimative EMOS with the bootstrap calibrated EMOS. We present some simulation studies and two real case applications to temperature forecast in the Veneto region (Italy) and to wind speed forecast in Germany. Appropriate verification measures such as CRPS, log-score, and coverage probabilities of central and upper prediction intervals are used for assessing the calibration and sharpness of the predictive models. From the results of the analyses, one can conclude that calibrated EMOS remarkably improves on estimative EMOS in terms of all the most commonly used measures of goodness.

Table 6 Wind case study for station 183. a) Truncated logistic distribution and b) Truncated normal distribution. Coverage probabilities of central intervals of level 0.67 and upper prediction limits for the estimative EMOS with MLE estimates (Est log), and the respective calibrated counterpart (Cal log). Standard errors in brackets.

a) Truncated logistic

α	Est log	Cal log
0.67	0.613 (0.008)	0.669 (0.008)
0.90	0.865 (0.006)	0.898 (0.005)
0.95	0.916 (0.005)	0.948 (0.004)
0.99	0.971 (0.003)	0.989 (0.002)

b) Truncated normal

α	Est log	Cal log
0.67	0.630 (0.008)	0.690 (0.008)
0.90	0.866 (0.006)	0.901 (0.005)
0.95	0.915 (0.005)	0.945 (0.004)
0.99	0.962 (0.003)	0.984 (0.002)

The analysis of maximum temperatures in Veneto and the analysis of wind speed in Germany does not include either the temporal or the spatial component in the model. As noticed in [18, 22], the temporal component can be disregarded by using a short enough window of training observations. Indeed, in a short period, the process can be assumed to be stationary, and in the presence of few observations, the need for calibrating estimative solutions is more compelling. The spatial structure could be included by allowing the coefficients to depend on the location, as in geo-statistical output perturbation models. This will be investigated in future work.

We would also like to note that, for the wind speed data, other stations have been analysed. It has been observed that the underlying distribution used to fit the data has a significant impact on the results. Indeed, the proposed calibrating procedure strongly relies on a good model specification. If the chosen model does not fit the data well, the bootstrap calibrated solution is unable

to correct the estimative solution. Further research is needed to develop a calibration method that is more robust to model misspecifications.

Funding. No funding was received for conducting this study.

Conflict of interest/Competing interests. The authors have no relevant financial or non-financial interests to disclose.

Supplementary information. The Supplementary Material contains the CMIP6 models used for this study.

Acknowledgments. We acknowledge the World Climate Research Programme that coordinated and promoted CMIP6, the Earth System Grid Federation (ESGF) for archiving the data and providing access. We appreciate all the organisations listed in the Supplementary Material for implementing and making their models available. Furthermore, we are very grateful to Sebastian Lerch for providing us with the wind speed dataset, useful for applying calibration to extended EMOS.

References

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

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

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

[4] S. Baran, D. Nemoda, Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting. Environmetrics, 27 (2016) 280–292.

[5] S. Baran, P. Szokol, M. Szabó, Truncated generalized extreme value distribution-based ensemble model output statistics model for calibration of wind speed ensemble forecasts. Environmetrics (2021), e2678.

[6] O.E. Barndorff-Nielsen, D.R. Cox, Prediction and asymptotics. Bernoulli, 2 (1996) 319–340.

[7] P. Bauer, A. Thorpe, G. Brunet, The quiet revolution of numerical weather prediction. Nature, 525 (2015) 47–55.

[8] R.H. Byrd, P. Lu, J. Nocedal, C.Y. Zhu, A limited memory algorithm for bound constrained optimization. SIAM Journal of Scientific Computing, 16 (1995) 1190–1208.

[9] C.G. Broyden, The convergence of a class of double-rank minimization algorithms. IMA Journal of Applied Mathematics, 6 (1970) 76–90.

[10] R. Buizza, Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system. Monthly Weather Review, 125 (1997) 99–119.

[11] Chen, J., Janke, T., Steinke, F., Lerch, S., Generative machine learning methods for multivariate ensemble post-processing. https://publikationen.bibliothek.kit.edu/1000151932 (2022).

[12] V. Eyring, S. Bony, G.A. Meehl, C.A. Senior, B. Stevens, R.J. Stouffer, K.E. Taylor, Overview of the Coupled Model Intercomparison Project phase 6 (CMIP6) experimental design and organization. Geoscientific Model Development, 9 (2016) 1937–1958.

[13] R. Fletcher, A new approach to variable metric algorithms computer. The computer Journal, 13 (1970) 317–322.

[14] G. Fonseca, F. Giummolè, P. Vidoni, Calibrating predictive distributions. Journal of Statistical Computation and Simulation, 84 (2014) 373–383.

[15] E. Gilleland, R. W. Katz. extRemes 2.0: An Extreme Value Analysis Package in R. Journal of Statistical Software, 72 (2016), 1–39.

[16] F. Giummolè, V. Mameli, Comparing predictive distributions in EMOS. Book of short papers SIS 2020, 823–828, Pearson, 2020.

[17] F. Giummolè, V. Mameli, Confidence predictive distributions: an application to temperature forecasting in Veneto. Book of short papers GRASPA 2023.

[18] T. Gneiting, A.E. Raftery, A.H. Westveld III, T. Goldman, Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. Monthly Weather Review, 133 (2005) 1098–1118.

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

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

[21] A. Goldfarb, A family of variable metric methods derived by variational means. Mathematics and Computation, 24 (1970) 23–26.

[22] L.E.S. Gomes, T.C.O. Fonseca, K.C.M. Gonçalves, R. Ruiz-Cárdenas, Space-time calibration of wind speed forecasts from regional climate models. Environmental and Ecological Statistics, 28 (2021) 631–665.

[23] T. Haiden, M. Janousek, F. Vitart, L. Ferranti, F. Prates, Evaluation of ECMWF forecasts, including the 2019 upgrade. ECMWF Tech. Memo. 588 (2019).

[24] Y.H. Kim, S.K. Min, X. Zhang, J. Sillmann, M. Sandstad, Evaluation of the CMIP6 multi-model ensemble for climate extreme indices. Weather and Climate Extremes, 29 (2020) 1–15.

[25] S. Lerch, T. L. Thorarinsdottir. Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. Tellus A: Dynamic Meteorology and Oceanography, 65:1 (2013) 21206.

[26] S. Lerch, S. Baran, Similarity-based semilocal estimation of postprocessing models. Royal Statistical Society. Series C, 66 (2017) 29–51.

[27] J.W. Messner, A. Zeileis, J. Broecker, G.J. Mayr. Probabilistic Wind Power Forecasts with an Inverse Power Curve Transformation and Censored Regression. Wind Energy, 17 (2013) 1753–1766.

[28] J.W. Messner, G.J. Mayr, A. Zeileis, D.S. Wilks. Heteroscedastic extended logistic regression for postprocessing of ensemble guidance. Monthly Weather Review, 142 (2014) 448–456.

[29] J.W. Messner, G.J. Mayr, D.S. Wilks, A. Zeileis, Extending Extended Logistic Regression: Extended versus Separate versus Ordered versus Censored. Monthly Weather Review, 142 (2014) 3003–3014,

[30] J.W. Messner, G.J. Mayr, A. Zeileis, Heteroscedastic Censored and Truncated Regression with crch. The R-Journal, 8 (2016), 173–181.

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

[32] A.E. Raftery, T. Gneiting, F. Balabdaoui, M. Polakowski, The ECMWF ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years. Quarterly Journal of the Royal Meteorological Society, 145 (2019) 12–24.

[33] J. Schanno, Conditions of quasi-Newton methods for function minimization. Mathematics of Computation, 24 (1970) 647–650.

[34] T.L. Thorarinsdottir, T. Gneiting, 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 (2010) 371–388.

[35] R.A. Yuen, S. Baran, C. Fraley, T. Gneiting, S. Lerch, M. Scheuerer, T. Thorarinsdottir, ensembleMOS: Ensemble Model Output Statistics. R package version 0.8.2 (2018).