

Mixture EMOS model for calibrating ensemble forecasts of wind speed

S. Baran^{a*} and S. Lerch^{b,c}

Ensemble model output statistics (EMOS) is a statistical tool for post-processing forecast ensembles of weather variables obtained from multiple runs of numerical weather prediction models in order to produce calibrated predictive probability density functions. The EMOS predictive probability density function is given by a parametric distribution with parameters depending on the ensemble forecasts. We propose an EMOS model for calibrating wind speed forecasts based on weighted mixtures of truncated normal (TN) and log-normal (LN) distributions where model parameters and component weights are estimated by optimizing the values of proper scoring rules over a rolling training period. The new model is tested on wind speed forecasts of the 50 member European Centre for Medium-range Weather Forecasts ensemble, the 11 member Aire Limitée Adaptation dynamique Développement International-Hungary Ensemble Prediction System ensemble of the Hungarian Meteorological Service, and the eight-member University of Washington mesoscale ensemble, and its predictive performance is compared with that of various benchmark EMOS models based on single parametric families and combinations thereof. The results indicate improved calibration of probabilistic and accuracy of point forecasts in comparison with the raw ensemble and climatological forecasts. The mixture EMOS model significantly outperforms the TN and LN EMOS methods; moreover, it provides better calibrated forecasts than the TN-LN combination model and offers an increased flexibility while avoiding covariate selection problems. © 2016 The Authors Environmetrics Published by John Wiley & Sons Ltd.

Keywords: continuous ranked probability score; ensemble calibration; ensemble model output statistics; truncated normal distribution; log-normal distribution

1. INTRODUCTION

In our industrialized world, several important applications require reliable and accurate wind speed forecasts. These include, but are not limited to agriculture, aviation or wind energy production. In particular, high wind speeds can cause severe damages to infrastructure, and their predictions are important parts of weather warnings. Wind speed forecasts are standard outputs of numerical weather prediction (NWP) models. NWP has traditionally been viewed as a deterministic problem; but over the last decades, a change towards probabilistic forecasts, that is, forecasts in the form of a full predictive distribution, can be observed. Probabilistic forecasts are important in the context of weather forecasting as they allow for a quantification of the associated uncertainty of the prediction and further allow for optimal point forecasting by using certain functionals of the predictive distribution (e.g., Gneiting, 2011).

Nowadays, weather services typically produce ensemble forecasts, which consist of multiple runs of NWP models that differ in initial conditions and/or the numerical representation of the atmosphere (Gneiting and Raftery, 2005). While the transition to ensemble forecasts is an important step towards probabilistic forecasting, ensembles are finite and do not provide full predictive densities. Further, ensemble forecasts are typically underdispersive and subject to systematic bias; they thus require some form of statistical post-processing (Gneiting and Raftery, 2005; Gneiting *et al.*, 2007).

State-of-the-art techniques for statistical post-processing of ensemble forecasts include Bayesian model averaging developed by (Raftery *et al.*, 2005) and ensemble model output statistics (EMOS) or non-homogeneous regression by (Gneiting *et al.*, 2005). The Bayesian model averaging approach uses weighted mixtures of parametric probability density functions (PDFs), which depend on the ensemble forecasts, with the mixture weights being determined based on the performance of the ensemble members in the training period. Possible component choices for wind speed are given by PDFs of Gamma distributions (Sloughter *et al.*, 2010) or truncated normal (TN) distributions (Baran,

* Correspondence to: S. Baran, Faculty of Informatics, University of Debrecen, Kassai út 26, Debrecen, H-4028, Hungary. E-mail: baran.sandor@inf.unideb.hu

a Faculty of Informatics, University of Debrecen, Debrecen, Hungary

b Heidelberg Institute for Theoretical Studies, Heidelberg, Germany

c Institute of Stochastics, Karlsruhe Institute of Technology, Karlsruhe, Germany

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

2014). By contrast, the predictive distribution of the EMOS approach is given by a single parametric distribution with parameters depending on the ensemble forecasts. Thorarinsdottir and Gneiting (2010) propose the use of truncated normal distributions for EMOS models of wind speed. Alternative choices are given by generalized extreme value distributions (Lerch and Thorarinsdottir, 2013), log-normal (LN) distributions (Baran and Lerch, 2015), and combinations thereof.

The article at hand builds on the EMOS framework and proposes models based on weighted mixtures of TN and LN distributions.

This new approach allows for combining the advantages of lighter and heavier-tailed distributions but avoids problems encountered in using previously proposed combination models. Apart from the flexibility, the mixture models further exhibit desirable properties from a theoretical perspective and provide well calibrated and skillful probabilistic forecasts.

The novel EMOS mixture approach is applied to forecasts of maximal wind speed of the 50 member ensemble of the European Centre for Medium-range Weather Forecasts (ECMWF; ECMWF Directorate 2012), and the eight-member University of Washington mesoscale ensemble (UWME; Eckel and Mass, 2005), and to instantaneous wind speed forecasts of the 11 member Limited Area Model Ensemble Prediction System of the Hungarian Meteorological Service (HMS) called Aire Limitée Adaptation dynamique Développement International-Hungary Ensemble Prediction System (ALADIN-HUNEPS; Hágel 2010; Horányi *et al.*, 2006). The three ensemble prediction systems (EPSs) differ in the generation of their members, which is accounted for in the model formulation. The TN model of Thorarinsdottir and Gneiting (2010) and the LN and TN–LN combination models of Baran and Lerch (2015) serve as benchmark models for the three case studies.

The remainder of this article is organized as follows. Section 2 contains a description of the ensembles and observational data. In Section 3, the EMOS technique is reviewed, and the novel TN–LN mixture models are introduced. Section 4 summarizes the results of the three case studies. The article concludes with a discussion in Section 5.

2. DATA

We consider three distinct data sets of ensemble forecasts and corresponding observations, which differ both in the stochastic properties of the ensemble as well as the observed wind quantities. The data sets coincide with those used in Baran and Lerch (2015). We thus limit our discussion here to a succinct summary of the data and refer to Baran and Lerch (2015) for a more detailed description.

2.1. European Centre for Medium-range Weather Forecasts ensemble

The ECMWF ensemble consists of 50 exchangeable ensemble members of 1-day ahead forecasts of 10-m daily maximum wind speed (given in m s^{-1}) along with corresponding validating observations from 228 synoptic observation stations over Germany. The observations are daily maxima of hourly observations of 10-min average wind speed measured over the 10 min before the hour, where the maxima are taken over the 24 hours corresponding to the time frame of the ensemble forecast. The results presented in Section 4.1 are based on a verification period from 1 May 2010 to 30 April 2011, consisting of 83 220 individual forecast cases.

Figure 1(a) shows the verification rank histogram of the raw ensemble, that is, the histogram of ranks of validating observations with respect to the corresponding ensemble forecasts computed from the ranks at all locations and dates considered (e.g., Wilks, 2011 Section 7.7.2). The strongly U-shaped histogram indicates a highly underdispersive character of the ECMWF ensemble. The range of the ECMWF ensemble contains the validating observation only in 43.40% of all cases (the nominal value of this coverage is 49/51, that is, 96.08%), verifying the need of statistical post-processing.

2.2. Aire Limitée Adaptation dynamique Développement International-Hungary Ensemble Prediction System ensemble

The ALADIN-HUNEPS system of the HMS (Horányi *et al.*, 2006; Descamps *et al.*, 2014) consists of 11 members, 10 exchangeable forecasts initialized from perturbed initial conditions, and one control member from the unperturbed analysis. The data base contains ensembles of 42-h forecasts for 10-m wind speed (given in m s^{-1}) for 10 major cities in Hungary, together with the corresponding validating observations for the 1-year period between 1 April 2012 and 31 March 2013.

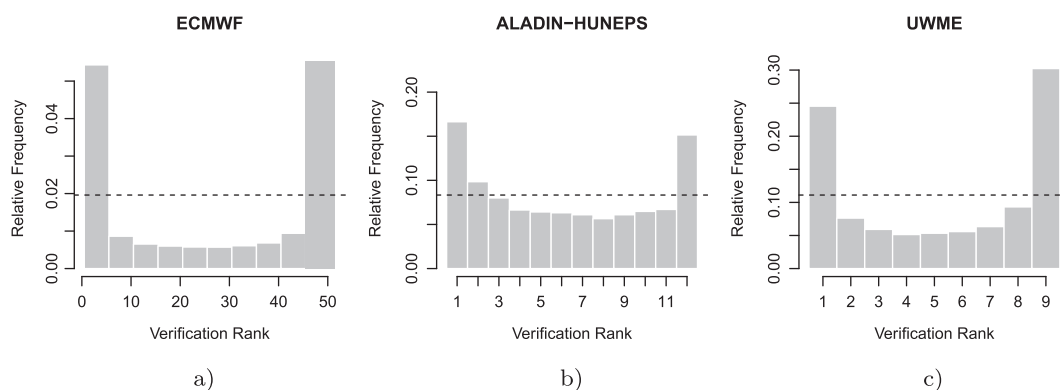


Figure 1. Verification rank histograms. (a) European Centre for Medium-range Weather Forecasts (ECMWF) ensemble for the period 1 May 2010–30 April 2011; (b) Aire Limitée Adaptation dynamique Développement International-Hungary Ensemble Prediction System (ALADIN-HUNEPS) ensemble for the period 1 April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

The validating wind speed measurements are considered as instantaneous values (valid at a given time), however, they are in fact mean values over the preceding 10 minutes. The model wind speed values are also considered as instantaneous, but they are representatives for a given model time step, which is 5 min in our case.

Similar to the ECMWF ensemble, the verification rank histogram of the raw ALADIN-HUNEPS ensemble is far from the desired uniform distribution (Figure 1(b)); however, it shows a much less underdispersive character. The better fit of the ensemble can also be observed on its coverage value of 61.21%, which should be compared with the nominal coverage of 83.33% (10/12).

2.3. University of Washington mesoscale ensemble

The UWME covering the Pacific Northwest region of western North America has eight members that are obtained from different runs of the fifth-generation Pennsylvania State University–National Center for Atmospheric Research mesoscale model (PSU-NCAR MM5; Grell *et al.*). Our data base contains ensembles of 48-h forecasts and corresponding validating observations of 10-m maximal wind speed (maximum of the hourly instantaneous wind speeds over the previous 12 h, given in $m s^{-1}$, e.g., Sloughter *et al.* (2010)) for 152 stations in the Automated Surface Observing Network (National Weather Service. 1998). The ensemble members are not exchangeable as they are generated with initial conditions from different sources.

In the present study, we investigate forecasts for calendar year 2008 with additional data from the last month of 2007 used for parameter estimation. After removing days and locations with missing data, 101 stations remain where the number of days for which forecasts and validating observations are available varies between 160 and 291.

Figure 1(c) shows the verification rank histogram of the raw ensemble where, similar to the previous cases, one can again observe a strongly underdispersive character. The ensemble coverage equals 45.24%, whereas the nominal coverage for eight ensemble members equals 7/9, that is 77.78%.

3. ENSEMBLE MODEL OUTPUT STATISTICS

As mentioned in the Introduction, the EMOS predictive distribution of a future weather quantity is a single parametric distribution, where the parameters depend on the ensemble. For example, a normal distribution provides a fairly good fit for temperature and pressure (Gneiting *et al.*, 2005), whereas wind speed requires a distribution with non-negative support.

In what follows, we consider three different types of EMOS models: standard EMOS models based on a single parametric family, combination models that select one of multiple parametric distributions based on the values of suitable covariates, and new mixture models. Models based on single parametric families for wind speed, which employ truncated normal, log-normal, or generalized extreme value distributions, as well as combination models selecting one of these distribution based on covariates have been explored in previous works (e.g., Thorarinsdottir and Gneiting, 2010; Lerch and Thorarinsdottir, 2013; Baran and Lerch, 2015) and are reviewed in Section 3.1. As an alternative choice, we propose new mixture models based on a weighted mixture of truncated normal and log-normal distributions in Section 3.2. The basic EMOS models from previous studies are used as benchmark models in order to assess the predictive performance of the novel mixture models in Section 4.

3.1. Basic ensemble model output statistics models

3.1.1. Models based on single parametric distributions

The EMOS model introduced by Thorarinsdottir and Gneiting (2010) is based on a TN distribution, that is, the predictive distribution is

$$\mathcal{N}_0(a_0 + a_1 f_1 + \dots + a_M f_M, b_0 + b_1 S^2) \quad \text{with} \quad S^2 := \frac{1}{M-1} \sum_{k=1}^M (f_k - \bar{f})^2 \tag{1}$$

where f_1, f_2, \dots, f_M denote the ensemble of distinguishable forecasts of wind speed for a given location and time, \bar{f} stands for the ensemble mean, and $\mathcal{N}_0(\mu, \sigma^2)$ denotes the TN distribution with location μ , scale $\sigma > 0$ and cut-off at zero having PDF

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

where φ and Φ are the PDF and the cumulative distribution function (CDF) of the standard normal distribution, respectively.

As an alternative to the TN distribution, Baran and Lerch (2015) propose the use of a LN distribution where the mean m and variance v are affine functions of the ensemble members and ensemble variance, respectively, that is,

$$m = \alpha_0 + \alpha_1 f_1 + \dots + \alpha_M f_M \quad \text{and} \quad v = \beta_0 + \beta_1 S^2. \tag{2}$$

Usually the PDF of the LN distribution $\mathcal{LN}(\mu, \sigma)$ is expressed using location μ and shape $\sigma > 0$ parameters and has the form

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

however, it can be easily expressed in terms of mean and variance with the help of transformations

$$\mu = \log \left(\frac{m^2}{\sqrt{v + m^2}} \right) \quad \text{and} \quad \sigma = \sqrt{\log \left(1 + \frac{v}{m^2} \right)}. \tag{3}$$

Lerch and Thorarinsdottir (2013) propose an EMOS approach based on a generalized extreme value (GEV) distribution; however, this model has the disadvantage of assigning positive probability to negative wind speed values (see e.g., Baran and Lerch, 2015).

Location and scale/shape parameters of models (1) and (2) can be estimated from the training data consisting of ensemble members and verifying observations from the preceding n days, by optimizing an appropriate verification score (Section 3.3).

Ensemble model output statistics models (1) and (2) are valid only in the cases when the sources of the ensemble members are clearly distinguishable, which is the case for the UWME described in Section 2.3 or for the Consortium for Small-scale Modeling Germany ensemble of the German Meteorological Service (Gebhardt *et al.*, 2011). However, in most of the currently used EPSs, some members are obtained with the help of perturbations of the initial conditions. These members are statistically indistinguishable and can be considered as exchangeable. This is the case for the ECMWF and ALADIN-HUNEPS ensembles described in Sections 2.1 and 2.2, respectively.

Suppose we have M ensemble members divided into m exchangeable groups, where the k th group contains $M_k \geq 1$ ensemble members such that $\sum_{k=1}^m M_k = M$. In such situations, the ensemble members within an exchangeable group should share the same parameters (Gneiting, 2014) resulting in a TN model

$$\mathcal{N}_0\left(a_0 + a_1 \sum_{\ell_1=1}^{M_1} f_{1,\ell_1} + \cdots + a_m \sum_{\ell_m=1}^{M_m} f_{m,\ell_m}, b_0 + b_1 S^2\right) \quad (4)$$

and a LN model with mean and variance

$$m = \alpha_0 + \alpha_1 \sum_{\ell_1=1}^{M_1} f_{1,\ell_1} + \cdots + \alpha_m \sum_{\ell_m=1}^{M_m} f_{m,\ell_m} \quad \text{and} \quad v = \beta_0 + \beta_1 S^2 \quad (5)$$

where $f_{k,\ell}$ denotes the ℓ th member of the k th group.

3.1.2. Combination models

Log-normal and GEV distributions have heavier upper tails than the TN distribution and are therefore more appropriate to model high wind speed values. To combine this advantage with the good performance of the TN model for low and medium wind speeds Lerch and Thorarinsdottir (2013) and Baran and Lerch (2015) also examined a regime-switching combination method, where either a TN or a heavy-tail distribution is used depending on the median value of the ensemble forecast. If the ensemble median is below a given threshold, wind speed is modeled by a TN distribution; otherwise, a GEV or LN distribution is employed. The optimal threshold value for a given EPS is determined during a preliminary study, and it is then fixed over the whole data set. The problem with this approach is that the threshold parameter is static (rarely updated) and cannot adapt to the changes in the ensemble. Baran and Lerch (2015) also consider a more adaptive method where the threshold parameter is re-estimated as a fixed quantile of the ensemble medians in the corresponding training period for each forecast date. However, this approach is computationally more demanding without yielding a significant improvement in predictive performance.

Ensemble model output statistics models based on combining two parametric families by exclusively selecting one of them at each forecast instance also suffer from the drawback that a suitable covariate has to be chosen as a selection criterion. This necessary step limits the flexibility of the combination models in practice as the adequacy of covariates might depend on the data set at hand. While the ensemble median works reasonably well in the data sets considered in this article, this observation might change for different EPSs.

3.2. Mixture models

In order to combine the advantages of lighter and heavier-tailed distributions and to avoid the aforementioned problems in the process, we introduce new EMOS models based on weighted mixtures of two parametric distributions.

In particular, we propose to model wind speed with a weighted mixture of models (1) and (2) (or (4) and (5) for exchangeable ensemble members) resulting in the predictive PDF

$$\psi(x | \mu_{TN}, \sigma_{TN}; \mu_{LN}, \sigma_{LN}; \omega) := \omega g(x | \mu_{TN}, \sigma_{TN}) + (1 - \omega)h(x | \mu_{LN}, \sigma_{LN}) \quad (6)$$

where the dependence of parameters μ_{TN} , σ_{TN} and μ_{LN} , σ_{LN} on the ensemble is given by (1) (or (4)), (2) (or (5)), and (3), respectively. In case of model (6), location and scale/shape parameters of the TN and LN models together with the weight $\omega \in [0, 1]$ are estimated simultaneously by optimizing some verification score over the training data.

Note that instead of a LN distribution, in (6), one can incorporate other non-negative laws with heavy right tails. A natural choice would be the generalized Pareto distribution used in extreme value theory (see, e.g., Bentzen and Friederichs, 2012), however, tests for the ensemble forecasts considered here indicate a worse predictive performance of the TN-generalized Pareto distribution model compared with the TN-LN mixture and the benchmark models.

In comparison with the basic EMOS models proposed in previous work, the new mixture models exhibit desirable properties from a theoretical perspective as they do not require the exclusive choice of one of multiple parametric families and are more flexible than models based on single parametric distributions. Their advantages from a practical perspective such as a significantly improved calibration will be demonstrated in Section 4.

3.3. Verification scores

The main aim of probabilistic forecasting is to access the maximal sharpness of the predictive distribution subject to calibration (Gneiting *et al.*, 2007). Calibration means a statistical consistency between the predictive distributions and the validating observations, whereas sharpness refers to the concentration of the predictive distribution. A straightforward way to check the calibration of a probabilistic forecast is the use of probability integral transform (PIT) histograms. The PIT is defined as the value of the predictive CDF evaluated at the verifying observations (Raftery *et al.*, 2005), and the closer the histogram to the uniform distribution, the better the calibration. PIT histograms are continuous analogues of verification rank histograms; a comparison of these histograms can thus be used as a measure of the possible improvements because of statistical post-processing.

Apart from the visual inspection of PIT histograms, formal statistical test of uniformity can be used to assess calibration. As the PIT values of multi-step ahead probabilistic forecast exhibit serial correlation (e.g., Diebold *et al.*, 1998) and the probabilistic forecasts cannot be assumed to be independent in space and time, we employ a moment-based test of uniformity proposed by Knüppel (2015), which accounts for dependence in the PIT values. In particular, we use the α_{1234}^0 test of Knüppel (2015) that has been demonstrated to have superior size and power properties compared with alternative choices. Because of the large sample size in case of the ECMWF and UWME data, the null hypothesis of uniformity is rejected for all post-processing models. However, as our focus lies on the comparative assessment of calibration, we report bootstrap estimates of the rejection rates of the α_{1234}^0 test based on 10,000 random samples of size 2500 each. If a model exhibits superior calibration, the null hypothesis of uniformity should be rejected in fewer cases compared with a model with inferior calibration.

Another approach to assess calibration is the investigation of the coverage of the $(1 - \alpha)100\%$, $\alpha \in (0, 1)$, central prediction interval, defined as the proportion of validating observations located between the lower and upper $\alpha/2$ quantiles of the predictive distribution, where α is chosen to match the nominal coverage of the raw ensemble (ECMWF: 96.08%; ALADIN-HUNEPS: 83.33%; UWME: 77.78%). The coverage of a calibrated predictive PDF should be around $(1 - \alpha)100\%$, and the proposed choices of α allow direct comparisons with the raw ensembles. Further, the average widths of these central prediction intervals provide information about the sharpness of the predictive distributions.

Calibration and sharpness can also be addressed simultaneously with the help of scoring rules, which measure the predictive performance by numerical values assigned to pairs of probabilistic forecasts and observations (Gneiting and Raftery, 2007). The most popular scoring rules are the logarithmic score (LogS), that is the negative logarithm of the predictive PDF $f(y)$ evaluated at the verifying observation,

$$\text{LogS}(F, x) = -\log(f(x))$$

and the continuous ranked probability score (CRPS; Gneiting and Raftery, 2007; Wilks, 2011). The CRPS of a CDF $F(y)$ and an observation x is defined as follows:

$$\text{CRPS}(F, x) := \int_{-\infty}^{\infty} (F(y) - 1_{\{y \geq x\}})^2 dy = \int_{-\infty}^x F^2(y) dy + \int_x^{\infty} (1 - F(y))^2 dy \tag{7}$$

where 1_H denotes the indicator of a set H . While both the CRPS and the logarithmic score are proper scoring rules (Gneiting and Raftery, 2007), the CRPS can be expressed in the same unit as the observation.

Further, point forecasts such as EMOS and ensemble medians and means are evaluated with the help of mean absolute errors (MAEs) and root mean squared errors (RMSEs). We remark that the former is optimal for the median, whereas the latter is for the mean (Gneiting, 2011).

Finally, to evaluate the goodness of fit of probabilistic forecasts to high wind speed values a useful tool to be considered is the threshold-weighted CRPS (twCRPS)

$$\text{twCRPS}(F, x) := \int_{-\infty}^{\infty} (F(y) - 1_{\{y \geq x\}})^2 \omega(y) dy$$

introduced by Gneiting and Ranjan (2011), where $\omega(y) \geq 0$ is a weight function. Obviously, $\omega(y) \equiv 1$ yields the traditional CRPS defined by (7), while one may set $\omega(y) = 1_{\{y \geq r\}}$ to address wind speeds above a given threshold r . Similar to Lerch and Thorarinsdottir (2013) and Baran and Lerch (2015), where the upper tail behaviors of regime-switching EMOS models are investigated, we consider threshold values approximately corresponding to the 90th, 95th, and 99th percentiles of the wind speed observations. One can also quantify the improvement in twCRPS with respect to some reference predictive CDF F_{ref} with the help of the threshold-weighted continuous ranked probability skill score (twCRPSS; e.g., Lerch and Thorarinsdottir, 2013) defined as follows:

$$\text{twCRPSS}(F, x) := 1 - \frac{\text{twCRPS}(F, x)}{\text{twCRPS}(F_{ref}, x)}$$

This score is obviously positively oriented, and in this study, the predictive CDF corresponding to the classical TN model is used as a reference.

In order to assess the statistical significance of observed score differences between the models, we use formal statistical tests of equal predictive performance. Diebold–Mariano (DM; Diebold and Mariano, 1995) tests allow to account for dependence in the forecast errors and are widely used in the econometric literature. Denote the mean values of a proper scoring rule S for two competing probabilistic forecasts F_t and G_t by $\bar{S}(F_t, x_t) = \frac{1}{N} \sum_{t \in T} S(F_t, x_t)$ and $\bar{S}(G_t, x_t) = \frac{1}{N} \sum_{t \in T} S(G_t, x_t)$, respectively, where $t \in T$ denotes the forecast cases in a test set T of size N . The test statistic of the DM test is given by the following:

$$t_N = \sqrt{N} \frac{\bar{S}(F_t, x_t) - \bar{S}(G_t, x_t)}{\hat{\sigma}_N} \quad (8)$$

where $\hat{\sigma}_N$ is a suitable estimator of the asymptotic standard deviation of the sequence of score differences $S(F_t, x_t) - S(G_t, x_t)$. Under some weak regularity assumptions, t_N asymptotically follows a standard normal distribution under the null hypothesis of equal predictive performance. Negative values of t_N indicate a better predictive performance of F , whereas G is preferred in case of positive values of t_N . The statistical significance of the observed values of the test statistic can be assessed by computing the corresponding p -values under the null hypothesis. Following suggestions of Diebold and Mariano (1995) and Gneiting and Ranjan (2011), we estimate the autocovariance of the sequence of score differences in (8) by the sample autocovariance up to lag $h - 1$ in case of h step ahead forecasts. The results of DM tests based on the LogS, the CRPS, and the twCRPS are discussed in Section 4 for the individual ensembles.

4. RESULTS

As mentioned in Section 1, the predictive skills of the mixture model (6) are tested on the 50 member ECMWF ensemble, the ALADIN-HUNEPS ensemble of the HMS, and the eight-member UWME. The three EPSs differ both in generation of the ensemble members and in the predicted wind speed quantity. Model performances are evaluated with the help of the verification scores given in Section 3.3.

The basic TN, LN, and TN–LN regime-switching combination EMOS models proposed in previous studies are used as benchmark models to assess the predictive performance of the new mixture models. For a detailed evaluation and comparison of the different basic EMOS models, we refer to Baran and Lerch (2015). We further compare the forecasts based on post-processing with the raw ensemble and with climatological forecasts where the observations of the training period are considered as an ensemble.

Following the ideas of Gneiting *et al.* (2005) and Thorarinsdottir and Gneiting (2010), the parameters of the TN, LN, and TN–LN mixture models are estimated by minimizing the mean CRPS of the predictive CDFs and corresponding validating observations over the training period.

However, in case of model (6), the CRPS can be evaluated only numerically, resulting in very long optimization procedures. Numerical minimization of the right-most part of Eq. 7, that is, $\text{CRPS}(F, x) = \int_{-\infty}^x F^2(y) dy + \int_x^{\infty} (1 - F(y))^2 dy$ leads to slightly lower computation times and better verification scores compared with minimizing alternative representations of the CRPS integral.

Because of the large computational costs of minimum CRPS estimation, we also investigate maximum likelihood (ML) estimation of the parameters. ML estimation corresponds to minimizing the mean logarithmic score, which has a simple and closed form. From a theoretical statistics perspective, both estimation approaches fit into a general framework of optimum score estimation and share asymptotic properties such as consistency; see Gneiting and Raftery (2007) for details. In applications to post-processing ensemble forecasts, the CRPS is often seen as the more appropriate scoring rule for parameter estimation because of the lower sensitivity to outliers and extreme events compared with the LogS (e.g., Gneiting *et al.*, 2005; Thorarinsdottir and Gneiting, 2010). In figures and tables, the corresponding mixture models are denoted by TN–LN mix. (CRPS) and TN–LN mix. (ML).

Finally, in order to ensure the comparability with the benchmark models, the same training period lengths and for the TN–LN regime-switching models the same thresholds and parameter estimation techniques as in Baran and Lerch (2015) are employed.

4.1. European Centre for Medium-Range Weather Forecasts ensemble

As the 50 members of the ECMWF ensemble are fully exchangeable, the dependencies of the parameters of the TN and LN distributions on the ensemble members are specified by (4) and (5), respectively, with $m = 1$ and $M = M_1 = 50$.

The preliminary study by Baran and Lerch (2015) suggested that the optimal training period length for this particular data set is 20 days, whereas the optimal value of the threshold parameter θ of the TN–LN regime-switching combination model equals 8 m s^{-1} , resulting in the use of an LN distribution in about 14% of the forecast cases. As mentioned in Section 2.1, model verification is performed on 83,220 forecast cases from the 1-year period between 1 May 2010 and 30 April 2011.

Figure 2 showing the verification rank histogram of the raw ensemble and the PIT histograms of the investigated EMOS models clearly indicates that statistical post-processing significantly improves the calibration of the raw ensemble. However, the histograms of TN and TN–LN regime-switching models are still biased, to a smaller extent the same applies for the LN model, whereas the PIT values of mixture model (6) with both parameter estimation methods suggest a better fit to the desired uniform distribution.

In order to quantify the observed differences in calibration, Table 1 shows rejection rates of the α_{1234}^0 test of uniformity based on random sub-samples. It can be observed that for the ECMWF data, the null hypothesis of uniformity is rejected in all of the cases for the TN, LN, and TN–LN combination model, whereas the novel mixture models show much lower rejection rates. This observation is clearly in line with the visual inspection of the PIT histograms in Figure 2.

The positive effect of post-processing can also be observed in Table 2 summarizing the verification scores for different probabilistic forecasts together with the average width and coverage of the 96.08% central prediction intervals. The improvement with respect to the raw ensemble and climatology is quantified in lower CRPS, twCRPS, MAE, and RMSE values, and the EMOS predictive PDFs result in calibrated central prediction intervals with coverages very close to the nominal value. The much wider central prediction intervals of the EMOS models compared with the ensemble are a natural consequence of the underdispersive character of the latter.

Among the competing post-processing methods, the TN–LN mixture and regime-switching models clearly outperform the TN and LN EMOS approaches in almost all scores investigated. The lowest CRPS value belongs to the mixture model with parameters estimated by optimizing the mean CRPS, whereas the regime-switching approach produces the best MAE, RMSE, and twCRPS scores. The two parameter estimation methods make only a very slight difference in model performance (ML estimation leads to slightly worse scores), and the TN–LN mixture EMOS models are fully able to keep up with the regime-switching approach. This ranking of models can also be observed

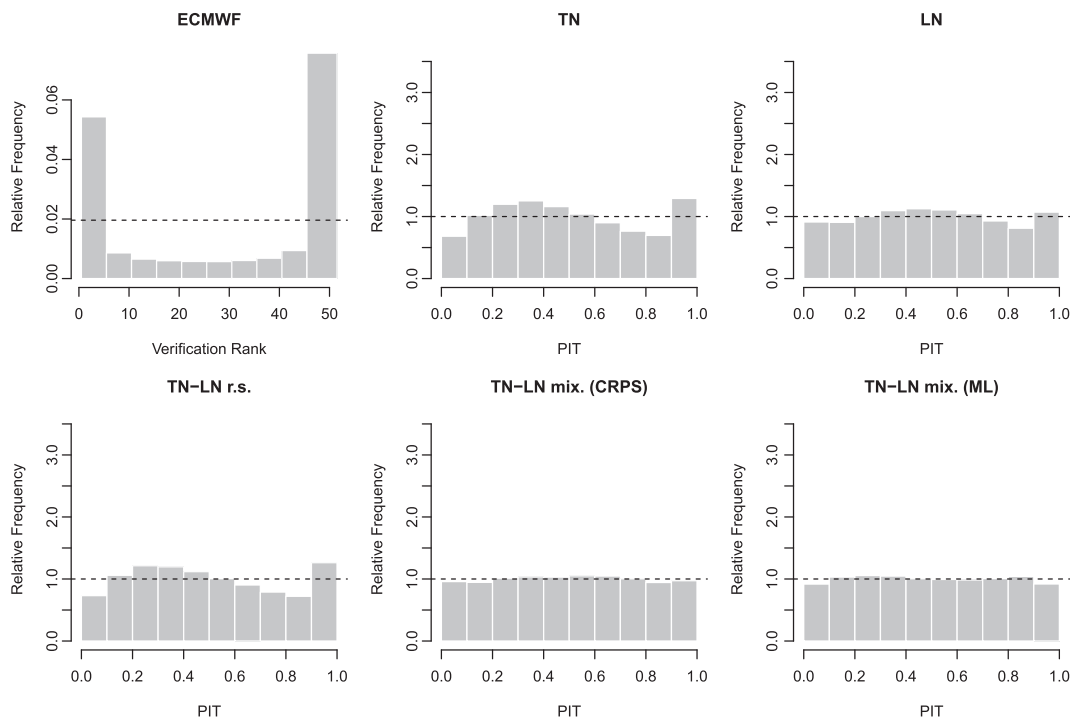


Figure 2. Verification rank histogram of the raw ensemble and probability integral transform (PIT) histograms of the ensemble model output statistics post-processed forecasts for the European Centre for Medium-range Weather Forecasts (ECMWF) ensemble. LN, log-normal; TN, truncated normal

Table 1. Bootstrap estimates of rejection rates of the α_{1234}^0 test of uniformity based on 10,000 random samples of size 2500 each at the 0.05 level for the different data sets. Lower rejection rates correspond to better calibrated forecasts with the null hypothesis of uniformity being rejected on fewer occasions

Ensemble	TN	LN	TN-LN r.s.	TN-LN mix. (CRPS)	TN-LN mix. (ML)
ECMWF	1	1	1	0.68	0.25
ALADIN-HUNEPS	1	1	1	0	0.01
UWME	1	1	0.68	0.47	0.31

ECMWF, European Centre for Medium-range Weather Forecasts; ALADIN-HUNEPS, Aire Limitée Adaptation dynamique Développement International-Hungary Ensemble Prediction System; UWME, University of Washington mesoscale ensemble; TN, truncated normal; LN, log-normal; CRPS, continuous ranked probability score; r.s., regime switching; mix., mixture.

Table 2. Mean CRPS, mean twCRPS for various thresholds r , MAE of median, and 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

Forecast	CRPS $m s^{-1}$	twCRPS $m s^{-1}$			MAE $m s^{-1}$	RMSE $m s^{-1}$	Cover. (%)	Av.w. $m s^{-1}$
		$r = 10$	$r = 12$	$r = 15$				
TN-LN mix. (CRPS)	1.030	0.194	0.106	0.041	1.384	2.135	94.34	7.71
TN-LN mix. (ML)	1.034	0.196	0.108	0.041	1.391	2.138	95.81	8.72
TN	1.045	0.200	0.110	0.042	1.388	2.148	92.19	6.39
LN	1.037	0.198	0.109	0.042	1.386	2.138	93.16	6.91
TN-LN r.s. ($\theta = 8.0$)	1.033	0.191	0.103	0.039	1.379	2.135	92.49	6.36
Ensemble	1.263	0.211	0.113	0.043	1.441	2.232	45.00	1.80
Climatology	1.550	0.251	0.128	0.045	2.144	2.986	95.84	11.91

TN, truncated normal; LN, log-normal; CRPS, continuous ranked probability score; av.w., average width; r.s., regime switching; mix., mixture; MAE, mean absolute error; RMSE, root mean squared error; twCRPS, threshold-weighted continuous ranked probability score.

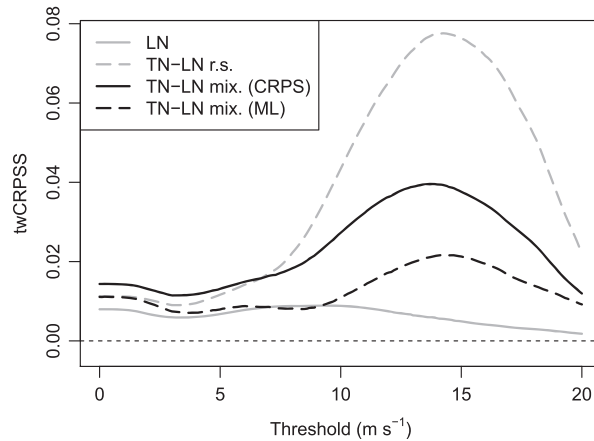


Figure 3. Threshold-weighted continuous ranked probability score (twCRPSS) values for the European Centre for Medium-range Weather Forecasts ensemble with truncated normal (TN) as reference model; ML, maximum likelihood; CRPS, continuous ranked probability score

Table 3. Values of the test statistics t_N of the two-sided DM test of equal predictive performance (8) for the comparison of the TN–LN mixture models and the benchmark models for the ECMWF ensemble

	TN	LN	TN–LN r.s.
<i>DM test based on LogS</i>			
TN–LN mix. (CRPS)	–36.67	–24.25	–32.79
TN–LN mix. (ML)	–36.43	–28.91	–32.86
<i>DM test based on CRPS</i>			
TN–LN mix. (CRPS)	–45.10	–27.10	–4.71
TN–LN mix. (ML)	–30.29	–11.42	0.32
<i>DM test based on twCRPS with threshold $r = 10$</i>			
TN–LN mix. (CRPS)	–21.55	–17.32	6.32
TN–LN mix. (ML)	–10.49	–2.83	9.92

Negative values indicate a superior predictive performance of the mixture model in the left column, and positive values indicate a better performance of the benchmark model. Values that are significant at the 0.05 level under the null hypothesis of equal predictive performance are printed in bold. DM, Diebold–Mariano; TN, truncated normal; LN, log-normal; mix., mixture; CRPS, continuous ranked probability score; twCRPS, threshold-weighted continuous ranked probability score; ML, maximum likelihood; r.s., regime switching; LogS, logarithmic score.

in Figure 3 displaying the twCRPSS values of the LN, TN–LN regime-switching, and TN–LN mixture (with CRPS and logarithmic score optimization) EMOS methods with respect to the reference TN EMOS model as functions of the threshold r . While the regime-switching approach outperforms the other models for all threshold values and has the best overall performance, it is clearly less flexible than the mixture models, which show the second best performance.

Table 3 summarizes the values of the test statistics of DM tests in order to assess the statistical significance of the observed score differences between the novel mixture models and the benchmark models. It can be observed that in terms of all employed proper scoring rules, the mixture models perform significantly better than the TN and LN models. The results of the comparison with the TN–LN combination model depend on the scoring rule. While the mixture models perform significantly better in terms of the LogS, the combination model is preferred in terms of the twCRPS.

Figure 4 shows the weights ω of the mixture model (6) estimated using optimizations with respect to the mean CRPS and the mean logarithmic score over the training data. Despite the similar predictive skills (Table 2), the two parameter-estimating methods result in completely different sets of weights having only a minor non-significant correlation of 0.063. However, having a closer look at the predictive

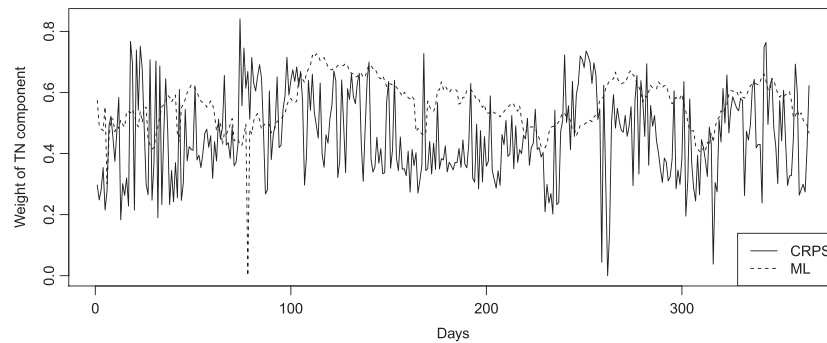


Figure 4. Weights of the truncated normal (TN) component for the European Centre for Medium-range Weather Forecasts ensemble. ML, maximum likelihood; CRPSS, continuous ranked probability score

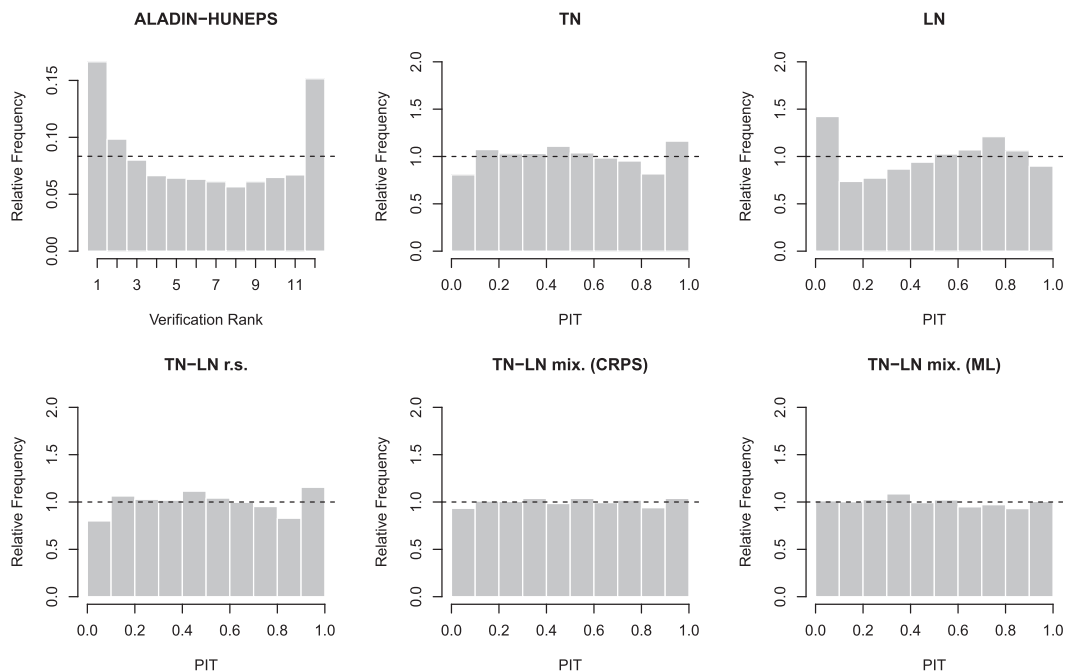


Figure 5. Verification rank histogram of the raw ensemble and probability integral transform (PIT) histograms of the ensemble model output statistics post-processed forecasts for the Aire Limitée Adaptation dynamique Développement International-Hungary Ensemble Prediction System (ALADIN-HUNEPS) ensemble. LN, log-normal; TN, truncated normal; CRPSS, continuous ranked probability score

PDFs, one can observe that the corresponding locations and scales/shapes of the TN and LN components produced by the two different estimation methods are strongly correlated, their correlations vary between 0.921, and 0.968, except for the scales of the TN component with a correlation of 0.283.

4.2. Aire Limitée Adaptation dynamique Développement International-Hungary Ensemble Prediction System ensemble

The ALADIN-HUNEPS ensemble consists of a control member and 10 exchangeable ensemble members (Section 2.2) inducing a natural splitting of the ensemble into two groups. The first group contains just the control, whereas the second group consists of the 10 exchangeable ensemble members. This results in TN and LN models (4) and (5), respectively, where $m = 2$, with $M_1 = 1$ and $M_2 = 10$.

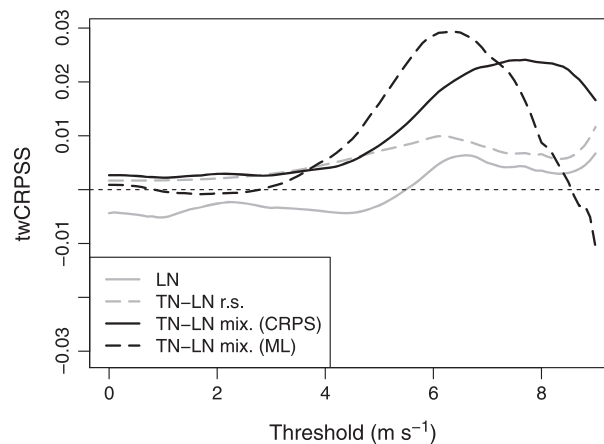
For this particular ensemble, Baran *et al.* (2014) and Baran and Lerch (2015) showed that a training period of length 43 days is optimal both for the TN and the LN EMOS models, whereas the optimal threshold for the TN–LN regime-switching combination model is $\theta = 6.9 \text{ m s}^{-1}$. For this threshold value, a LN distribution is used in 4% of the forecast cases.

Using a 43-day training period, one has ensemble forecasts and verifying observations for 315 calendar days (i.e., 3150 forecast cases) between 15 May 2012 and 31 March 2013. Figure 5 shows the PIT histograms of the various post-processing methods together with the verification rank histogram of the ALADIN-HUNEPS ensemble. Again, compared with the raw ensemble, one can clearly see the improvement in calibration of post-processed forecasts, whereas from the competing EMOS methods, the two variants of the mixture model (6) provide the most uniform PIT histograms.

Table 4. Mean CRPS, mean twCRPS for various thresholds r , MAE of median, and 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

Forecast	CRPS m s^{-1}	twCRPS m s^{-1}			MAE m s^{-1}	RMSE m s^{-1}	Cover. (%)	Av.w. m s^{-1}
		$r=6$	$r=7$	$r=9$				
TN-LN mix. (CRPS)	0.736	0.100	0.053	0.011	1.037	1.358	83.02	3.62
TN-LN mix. (ML)	0.737	0.100	0.053	0.012	1.040	1.360	83.14	3.58
TN	0.738	0.102	0.054	0.012	1.037	1.357	83.59	3.53
LN	0.741	0.102	0.054	0.011	1.038	1.362	80.44	3.57
TN-LN r.s. ($\theta = 6.9$)	0.737	0.101	0.054	0.011	1.035	1.356	83.59	3.54
Ensemble	0.803	0.112	0.059	0.013	1.069	1.373	68.22	2.88
Climatology	1.046	0.127	0.064	0.012	1.481	1.922	82.54	3.43

TN, truncated normal; LN, log-normal; CRPS, continuous ranked probability score; av.w., average width; r.s., regime switching; mix., mixture; MAE, mean absolute error; RMSE, root mean squared error; twCRPS, threshold-weighted continuous ranked probability score.

**Figure 6.** Threshold-weighted continuous ranked probability score (twCRPSS) values for the Aire Limitée Adaptation dynamique Développement International-Hungary Ensemble Prediction System ensemble with truncated normal (TN) as reference model. LN, log-normal; ML, maximum likelihood; CRPS, continuous ranked probability score

The significantly better calibration of the mixture models can also be observed from the rejection rates of the α_{1234}^0 test reported in Table 1. The null hypothesis of uniformity of the PIT values is not rejected for almost all of the random samples, whereas it is rejected on almost all occasions for the competing models based on single parametric distributions and the TN-LN regime-switching combination model.

In Table 4, the verification scores of probabilistic and point forecasts and coverage and average width of 83.33% central prediction intervals are given for the various EMOS models, the ALADIN-HUNEPS ensemble, and climatological forecasts. The raw ensemble outperforms climatology and produces sharp forecasts, however, at the cost of being uncalibrated. Post-processing substantially improves the calibration and predictive skill of the raw ensemble, which is in line with the shapes of histograms displayed in Figure 5. The TN-LN mixture and regime-switching combination models show some small improvements over the TN and LN models in terms of all scoring rules and display almost the same predictive performance. For small threshold values, the two versions of the mixture model slightly outperform the three benchmark EMOS approaches. However, above the 99th percentile of the validating observations (9 m s^{-1}), their performances decay quickly; see Figure 6.

In order to assess the statistical significance of the score differences, results of the DM tests of equal predictive performance are reported in Table 5. In terms of LogS and CRPS, the mixture models significantly outperform the LN model, and the same observation holds in terms of the twCRPS for the comparison with the TN model. In comparison with the TN-LN combination model, the preferred model depends on the employed scoring rule. The only significant score difference in this comparison is in terms of the LogS and favors the mixture model based on ML estimation.

Finally, similar to the previous case study, the weights belonging to the two parameter estimation methods for the TN-LN mixture model (Figure 7) are uncorrelated, whereas the correlations of the corresponding location and scale/shape parameters of the TN (μ_{TN} and σ_{TN}) and LN components (μ_{LN} and σ_{LN}) are 0.875, 0.660 and 0.747, 0.414, respectively.

Table 5. Values of the test statistics t_N of the two-sided DM test of equal predictive performance (8) for the comparison of the TN–LN mixture models and the benchmark models for the ALADIN-HUNEPS ensemble

	TN	LN	TN–LN r.s.
<i>DM test based on LogS</i>			
TN–LN mix. (CRPS)	0.71	–4.78	0.63
TN–LN mix. (ML)	–2.43	–5.36	–2.09
<i>DM test based on CRPS</i>			
TN–LN mix. (CRPS)	–2.03	–3.73	–0.71
TN–LN mix. (ML)	–0.58	–2.56	0.31
<i>DM test based on twCRPS with threshold $r = 6$</i>			
TN–LN mix. (CRPS)	–3.19	–1.90	–1.14
TN–LN mix. (ML)	–2.24	–1.86	–1.11

Negative values indicate a superior predictive performance of the mixture model in the left column, and positive values indicate a better performance of the benchmark model. Values that are significant at the 0.05 level under the null hypothesis of equal predictive performance are printed in bold. DM, Diebold–Mariano; TN, truncated normal; LN, log-normal; mix., mixture; CRPS, continuous ranked probability score; twCRPS, threshold-weighted continuous ranked probability score; ML, maximum likelihood; r.s., regime switching; LogS, logarithmic score.

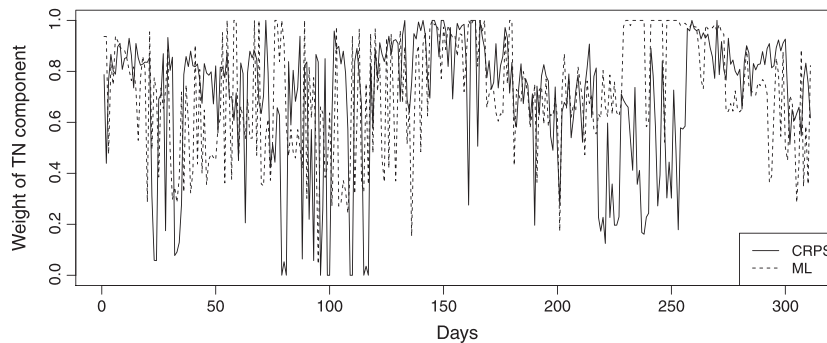


Figure 7. Weights of the truncated normal (TN) component for the Aire Limitée Adaptation dynamique Développement International-Hungary Ensemble Prediction System ensemble. LN, log-normal; ML, maximum likelihood; CRPS, continuous ranked probability score

4.3. University of Washington mesoscale ensemble

The members of the UWME are clearly distinguishable, as they are generated using initial conditions from eight different sources. Hence, location and scale/shape parameters of the TN and LN models are linked to the ensemble via (1) and (2), respectively, with $M = 8$. According to Baran and Lerch (2015), the optimal training period for this data set is of length 30 days, and the optimal threshold value θ of the TN–LN combination model is equal to 5.7 m s^{-1} . Ensemble forecasts for calendar year 2008 are calibrated using these parameters, and in case of the regimes-switching approach, a LN model is used in around one third of the 27,481 individual forecast cases.

Similar to the previous two sections, consider first the PIT histograms of the EMOS predictive distributions displayed in Figure 8. Compared with the verification rank histogram of the raw ensemble, all post-processing methods result in significant improvements in the goodness of fit to the uniform distribution, while, from the various calibration methods, the TN–LN mixture and regime-switching models have the best performance.

The rejection rates of the α_{1234}^0 tests reported in Table 1 indicate that the TN–LN mixture model based on ML estimation exhibits the best calibration followed by the mixture model based on minimum CPRS estimation and the TN–LN combination model.

Verification scores for probabilistic and point forecasts and the coverage and average width of 77.78% central prediction intervals are reported in Table 6. Compared with the raw ensemble and climatology, post-processed forecast exhibits the same behavior as before improved

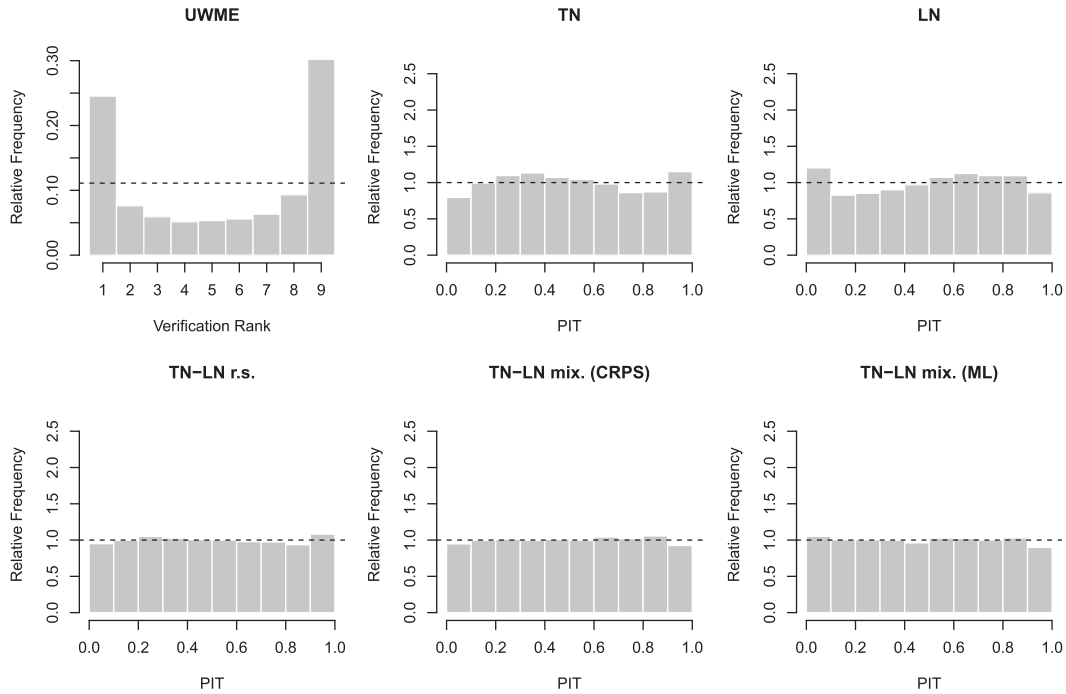


Figure 8. Verification rank histogram of the raw ensemble and probability integral transform (PIT) histograms of the ensemble model output statistics post-processed forecasts for the University of Washington mesoscale ensemble (UWME). LN, log-normal; TN, truncated normal; CRPS, continuous ranked probability score

Table 6. Mean CRPS, mean twCRPS for various thresholds r , MAE of median, and RMSE of mean forecasts and coverage and average width of 77.78% central prediction intervals for the University of Washington mesoscale ensemble

Forecast	CRPS $m s^{-1}$	twCRPS $m s^{-1}$			MAE $m s^{-1}$	RMSE $m s^{-1}$	Cover. (%)	Av. w. $m s^{-1}$
		$r = 9$	$r = 10.5$	$r = 14$				
TN-LN mix. (CRPS)	1.105	0.147	0.073	0.010	1.550	2.045	79.02	4.77
TN-LN mix. (ML)	1.108	0.147	0.073	0.010	1.560	2.062	78.12	4.78
TN	1.114	0.150	0.074	0.010	1.550	2.048	78.65	4.67
LN	1.114	0.147	0.073	0.010	1.554	2.052	77.29	4.69
TN-LN r.s. ($\theta = 5.7$)	1.105	0.149	0.073	0.010	1.550	2.050	77.73	4.64
Ensemble	1.353	0.175	0.085	0.011	1.655	2.169	45.24	2.53
Climatology	1.412	0.173	0.081	0.010	1.987	2.629	81.10	5.90

TN, truncated normal; LN, log-normal; CRPS, continuous ranked probability score; av.w., average width; r.s., regime switching; mix., mixture; MAE, mean absolute error; RMSE, root mean squared error; twCRPS, threshold-weighted continuous ranked probability score.

predictive skills and better calibration. In general, models based on combinations of both investigated distributions outperform the TN and LN methods; the smallest CRPS and MAE values and the best coverage, combined with a rather narrow central prediction interval, belong to the regime-switching approach, while the mixture model with parameters optimizing the mean CRPS provides the lowest twCRPS and RMSE scores.

The results of DM tests of equal predictive performance are summarized in Table 7. In terms of all employed scoring rules, the mixture models exhibit significantly better predictive performance compared with the TN and LN models. As before, the results for the comparisons with the TN-LN combination model are mixed and depend on the scoring rule. While the mixture models show significantly better results in terms of the LogS, the combination model is preferred in terms of the CPRS and its threshold-weighted version.

This desirable behavior of the mixture models for high wind speeds can also be observed in Figure 9 where the twCRPSS values of the LN, TN-LN regime-switching, and mixture models with respect to the TN EMOS reference model are plotted as functions of the threshold. Up to threshold $r = 9 m s^{-1}$, the regime-switching method slightly outperforms the mixture model, whereas above it, this advantage disappears.

In contrast to the ECMWF and ALADIN-HUNEPS ensembles, the weights of the TN component of the two versions of model (6) plotted in Figure 10 show a positive correlation of 0.214. Finally, for the UWME, the parameter estimates of μ_{LN} and σ_{LN} exhibit strong-

Table 7. Values of the test statistics t_N of the two-sided DM test of equal predictive performance (8) for the comparison of the TN–LN mixture models and the benchmark models for the UWME data

	TN	LN	TN–LN r.s.
<i>DM test based on LogS</i>			
TN–LN mix. (CRPS)	–7.63	–13.62	–4.68
TN–LN mix. (ML)	–18.44	–16.45	–12.24
<i>DM test based on CRPS</i>			
TN–LN mix. (CRPS)	–15.73	–16.70	–0.95
TN–LN mix. (ML)	–5.62	–6.52	2.24
<i>DM test based on twCRPS with threshold $r = 9$</i>			
TN–LN mix. (CRPS)	–7.68	–4.63	1.03
TN–LN mix. (ML)	–7.04	–5.11	1.06

Negative values indicate a superior predictive performance of the mixture model in the left column, and positive values indicate a better performance of the benchmark model. Values that are significant at the 0.05 level under the null hypothesis of equal predictive performance are printed in bold. DM, Diebold–Mariano; TN, truncated normal; LN, log-normal; mix., mixture; CRPS, continuous ranked probability score; twCRPS, threshold-weighted continuous ranked probability score; ML, maximum likelihood; r.s., regime switching; LogS, logarithmic score.

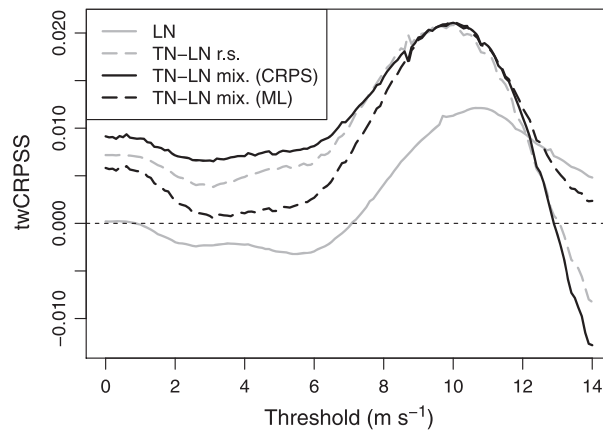


Figure 9. Threshold-weighted continuous ranked probability score (twCRPSS) values for the University of Washington mesoscale ensemble with truncated normal (TN) as reference model. LN, log-normal; CRPS, continuous ranked probability score; ML, maximum likelihood

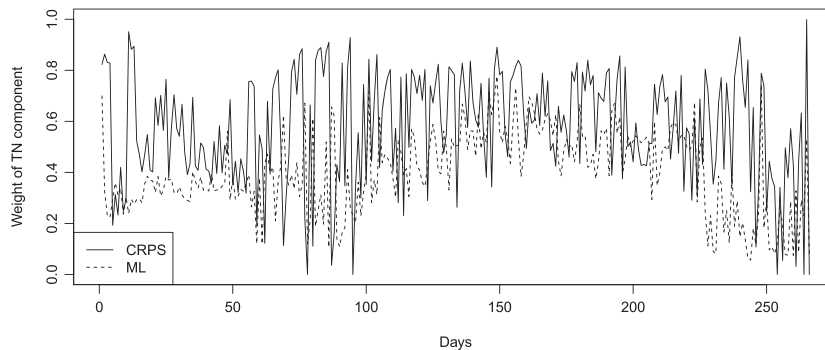


Figure 10. Weights of the truncated normal (TN) component for the University of Washington mesoscale ensemble. CRPS, continuous ranked probability score; ML, maximum likelihood

er correlations than the estimated location and scale parameters μ_{TN} and σ_{TN} of the TN component; the corresponding values are 0.858, 0.826 and 0.427, 0.259, respectively.

5. CONCLUSIONS

A new EMOS model for post-processing ensemble forecasts of wind speed is introduced, where the predictive PDF is a weighted mixture of a truncated normal and a log-normal distribution with location and scale/shape parameters depending on the ensemble. Model parameters and mixture weight are estimated simultaneously by optimizing either the mean continuous ranked probabilistic score or the mean logarithmic score (ML estimation) of the predictive distribution over the training data.

The mixture models are tested on three data sets of wind speed forecasts, which differ in the generation of the ensemble members and the predicted wind quantities. The predictive skills of the new model are compared with those of the TN-based EMOS method (Thorarindottir and Gneiting, 2010), the LN and the TN–LN regime-switching EMOS models (Baran and Lerch, 2015), the raw ensemble and the climatological forecasts with the help of graphical tools, appropriate verification scores and formal statistical tests of calibration, and equal predictive performance. The presented case studies clearly show that compared with the raw ensemble and climatology, statistical post-processing results in a significant improvement in calibration of probabilistic and accuracy of point forecasts.

In a comparative view of the different EMOS models, it can be observed that the TN–LN regime-switching combination model and the new mixture models significantly outperform the simple EMOS models based on single TN and LN distributions and provide much better calibrated probabilistic forecasts as demonstrated by formal statistical tests. The novel mixture models are able to keep up with the TN–LN regime-switching combination method in terms of the various verification scores. Further, the results of formal tests of uniformity indicate a superior calibration of the forecasts produced by the mixture models compared with the combination model. No substantial difference can be observed between the results corresponding to the two parameter estimation methods of the mixture model; ML estimation results in slightly worse verification scores but provides better calibrated forecasts.

Compared with the TN–LN regime-switching combination model, the proposed mixture models exhibit desirable properties from both a theoretical and an applied perspective. They are more flexible in that they do not require the exclusive choice of one of the parametric families as forecast distribution. Further, it is not necessary to determine suitable covariates for the model selection, or to estimate the model selection threshold over a training period.

The article at hand suggests that the choice of a suitable parametric EMOS model for post-processing ensemble forecasts is a non-trivial exercise and depends on the employed performance measure as well as the data. Following the suggestion of an anonymous reviewer, we note that an alternative approach might be given by suitably adapted smoothing techniques applied to the empirical CDF of the ensemble. Recent methods such as the Bernstein polynomial smoother proposed by Turnbull and Ghosh (2014) allow for preserving the desired unimodal shape of the distribution. A detailed investigation of this approach is left as a topic for future research.

Acknowledgements

Essential part of this work was made during a visit of Sándor Baran at the Heidelberg Institute for Theoretical Studies. The research stay in Heidelberg was funded by the DAAD program “Research Stays for University Academics and Scientists, 2015”. Sándor Baran was also supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. Sebastian Lerch gratefully acknowledges support by the Volkswagen Foundation within the program “Mesoscale Weather Extremes – Theory, Spatial Modelling and Prediction (WEX-MOP)”, and by the Klaus Tschira Foundation. The authors thank Tilmann Gneiting, Alexander Jordan, and Fabian Krüger for helpful discussions, and Fabian Krüger for providing R code for the α_{1234}^0 test. The comments of the associate editor and two reviewers greatly improved the quality of this article. The authors further thank the University of Washington MURI group for providing the UWME data and Mihály Szűcs from the HMS for providing the ALADIN-HUNEPS data.

REFERENCES

- Baran S. 2014. Probabilistic wind speed forecasting using Bayesian model averaging with truncated normal components. *Computational Statistics and Data Analysis* **75**:227–238.
- Baran S, Lerch S. 2015. Log-normal distribution based EMOS models for probabilistic wind speed forecasting. *Quarterly Journal of the Royal Meteorological Society* **141**:2289–2299.
- 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.
- Bentzen S, Friederichs P. 2012. Generating and calibrating probabilistic quantitative precipitation forecasts from the high-resolution NWP model COSMO-DE. *Weather and Forecasting* **27**:988–1002.
- Descamps L, Labadie C, Joly A, Bazile E, Arbogast P, Cébron P. 2014. PEARP, the Météo-France short-range ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society* **141**:1671–1685.
- Diebold FX, Gunther T, Tay A. 1998. Evaluating density forecasts, with applications to financial risk management. *International Economic Review* **39**: 863–883.
- Diebold FX, Mariano RS. 1995. Comparing predictive accuracy. *Journal of Business & Economic Statistics* **13**:253–263.
- Eckel FA, Mass CF. 2005. Effective mesoscale, short-range ensemble forecasting. *Weather and Forecasting* **20**:328–350.
- ECMWF Directorate. 2012. Describing ECMWF’s forecasts and forecasting system. *ECMWF Newsletter* **133**:11–13.
- Gebhardt C, Theis SE, Paulat M, Bouallégué ZB. 2011. Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries. *Atmospheric Research* **100**:168–177.
- 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. (Available from: http://old.ecmwf.int/publications/library/ecpublications/_pdf/tm/701-800/tm719.pdf.) [Accessed on 27 July 2015]
- Gneiting T, Raftery AE. 2005. Weather forecasting with ensemble methods. *Science* **310**:248–249.
- Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association* **102**:359–378.
- 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, Balabdaoui F, Raftery A E. 2007. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B* **69**:243–268.
- Gneiting T, Raftery AE, Westveld AH, Goldman T. 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review* **133**:1098–1118.
- Grell GA, Dudhia J, Stauffer DR. 1995: A description of the fifth-generation Penn state/NCAR mesoscale model (MM5). Technical Note NCAR/TN-398+STR. National Center for Atmospheric Research, Boulder. (Available from: <http://nldr.library.ucar.edu/repository/assets/technotes/TECH-NOTE-000-000-000-214.pdf>) [Accessed on 27 July 2015]
- Hágel E. 2010. *Időjárás* **114**:121–133.
- 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.
- Knüppel M. 2015. Evaluating the calibration of multi-step-ahead density forecasts using raw moments. *Journal of Business & Economic Statistics* **33**: 270–281.
- Lerch S, Thorarinsdottir TL. 2013. Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus A* **65**:21206.
- National Weather Service. 1998: Automated Surface Observing System (ASOS) User Guide. (Available from: <http://www.weather.gov/asos/aum-toc.pdf>) [Accessed on 27 July 2015]
- Raftery AE, Gneiting T, Balabdaoui F, Polakowski M. 2005. Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review* **133**:1155–1174.
- Sloughter JM, Gneiting T, Raftery AE. 2010. Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *Journal of the American Statistical Association* **105**:25–37.
- Thorarinsdottir TL, 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* **173**:371–388.
- Turnbull BC, Ghosh SK. 2014. Unimodal density estimation using Bernstein polynomials. *Computational Statistics & Data Analysis* **72**:13–29.
- Wilks DS. 2011. *Statistical Methods in the Atmospheric Sciences* 3rd ed. Elsevier: Amsterdam.