Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting

S. Baran[a] and D. Nemoda[a,b]

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

[b]Faculty of Mechanical Engineering and Informatics, University of Miskolc, Hungary

Abstract

Recently all major weather prediction centres provide forecast ensembles of different weather quantities which are obtained from multiple runs of numerical weather prediction models with various initial conditions and model parametrizations. However, ensemble forecasts often show an underdispersive character and may also be biased, so that some post-processing is needed to account for these deficiencies. Probably the most popular modern post-processing techniques are the ensemble model output statistics (EMOS) and the Bayesian model averaging (BMA) which provide estimates of the density of the predictable weather quantity.

In the present work an EMOS method for calibrating ensemble forecasts of precipitation accumulation is proposed, where the predictive distribution follows a censored and shifted gamma (CSG) law with parameters depending on the ensemble members. The CSG EMOS model is tested on ensemble forecasts of 24 h precipitation accumulation of the eight-member University of Washington mesoscale ensemble and on the 11 member ensemble produced by the operational Limited Area Model Ensemble Prediction System of the Hungarian Meteorological Service. The predictive performance of the new EMOS approach is compared with the fit of the raw ensemble, the generalized extreme value (GEV) distribution based EMOS model and the gamma BMA method. According to the results, the proposed CSG EMOS model slightly outperforms the GEV EMOS approach in terms of calibration of probabilistic and accuracy of point forecasts and shows significantly better predictive skill that the raw ensemble and the BMA model.

Key words: Continuous ranked probability score, ensemble calibration, ensemble model output statistics, gamma distribution, left censoring.

1 Introduction

Reliable and accurate prediction of precipitation is of great importance in agriculture, tourism, aviation and in some other fields of economy as well. In order to represent the

uncertainties of forecasts based on observational data and numerical weather prediction (NWP) models one can run these models with different initial conditions or change model physics, resulting in a forecast ensemble (Leith, 1974). In the last two decades this approach has became a routinely used technique all over the world and recently all major weather prediction centres have their own operational ensemble prediction systems (EPS), e.g. the Consortium for Small-scale Modelling (COSMO-DE) EPS of the German Meteorological Service (DWD; Gebhardt *et al.*, 2011; Bouallègue *et al.*, 2013), the Prévision d'Ensemble ARPEGE (PEARP) EPS of Méteo France (Descamps *et al.*, 2014) or the EPS of the independent intergovernmental European Centre for Medium-Range Weather Forecasts (ECMWF Directorate, 2012). With the help of a forecast ensemble one can estimate the distribution of the predictable weather quantity which opens up the door for probabilistic forecasting (Gneiting and Raftery, 2005). By post-processing the raw ensemble the most sophisticated probabilistic methods result in full predictive cumulative distribution functions (CDF) and correct the possible bias and underdispersion of the original forecasts. The underdispersive character of the ensemble has been observed with several ensemble prediction systems (Buizza *et al.*, 2005) and this property also leads to the lack of calibration. Using predictive CDFs one can easily get consistent estimators of probabilities of various meteorological events or calculate different prediction intervals.

Recently, probably the most widely used ensemble post-processing methods leading to full predictive distributions (for an overview see e.g. Gneiting, 2014; Williams *et al.*, 2014) are the Bayesian model averaging (BMA; Raftery *et al.*, 2005) and the non-homogeneous regression or ensemble model output statistics (EMOS; Gneiting *et al.*, 2005), as they are partially implemented in the `ensembleBMA` and `ensembleMOS` packages of `R` (Fraley *et al.*, 2011).

The BMA predictive probability density function (PDF) of the future weather quantity is the mixture of individual PDFs corresponding to the ensemble members with mixture weights determined by the relative performance of the ensemble members during a given training period. To model temperature or sea level pressure a normal mixture seems to be appropriate (Raftery *et al.*, 2005), wind speed requires non-negative and skewed component PDFs such as gamma (Sloughter *et al.*, 2010) or truncated normal (Baran, 2014) distributions, whereas for surface wind direction a von Mises distribution (Bao *et al.*, 2010) is suggested. However, in some situations BMA post-processing might result, for instance, in model overfitting (Hamill, 2007) or over-weighting climatology (Hodyss *et al.*, 2015).

In contrast to BMA, the EMOS technique uses a single parametric PDF with parameters depending on the ensemble members. Again, for temperature and sea level pressure the EMOS predictive PDF is normal (Gneiting *et al.*, 2005), whereas for wind speed truncated normal (Thorarinsdottir and Gneiting, 2010), generalized extreme value (GEV; Lerch and Thorarinsdottir, 2013), censored logistic (Messner *et al.*, 2014), truncated logistic, gamma (Scheuerer and Möller, 2015) and log-normal (Baran and Lerch, 2015) distributions are suggested.

However, statistical calibration of ensemble forecasts of precipitation is far more difficult than the post-processing of the above quantities. As pointed out by Scheuerer and Hamill (2015), precipitation has a discrete-continuous nature with a positive probability of being zero and larger expected precipitation amount results in larger forecast uncertainty.

Sloughter *et al.* (2007) introduced a BMA model where each individual predictive PDF consists of a discrete component at zero and a gamma distribution modelling the case of positive precipitation amounts. Wilks (2009) uses extends logistic regression to provide full probability distribution forecasts, whereas Scheuerer (2014) suggests an EMOS model based on a censored GEV distribution. Finally, Scheuerer and Hamill (2015) propose a more complex three step approach where they first fit a censored and shifted gamma (CSG) distribution model to the climatological distribution of observations, then after adjusting the forecasts to match this climatology derive three ensemble statistics, and with the help of a nonhomogeneus regression model connect these statistics to the CSG model.

Based on the idea of Scheuerer and Hamill (2015) we introduce a new EMOS approach which directly models the distribution of precipitation accumulation with a censored and shifted Gamma predictive PDF. The novel EMOS approach is applied to 24 hour precipitation accumulation forecasts of the eight-member University of Washington mesoscale ensemble (UWME; Eckel and Mass, 2005) and the 11 member operational EPS of the Hungarian Meteorological Service (HMS) called Aire Limitée Adaptation dynamique Dévelopment International - Hungary EPS (ALADIN-HUNEPS; Horányi *et al.*, 2006, 2011). In these case studies the performance of the proposed EMOS model is compared to the forecast skills of the GEV EMOS method of Scheuerer (2014) and to the gamma BMA approach of Sloughter *et al.* (2007) serving as benchmark models.

2 Ensemble Model Output Statistics

As mentioned in the Introduction, the EMOS predictive PDF of a future weather quantity is a single parametric distribution with parameters depending on the ensemble members. Due to the special discrete-continuous nature of precipitation one should think only of non-negative predictive distributions assigning positive mass to the event of zero precipitation. Mixing a point mass at zero and a separate non-negative distribution does the job (see e.g. the BMA model of Sloughter *et al.*, 2007), but left censoring of an appropriate continuous distribution at zero can also be a reasonable choice. The advantage of the latter approach is that the probability of zero precipitation can directly be derived from the corresponding original (uncensored) cumulative distribution function (CDF), so the cases of zero and positive precipitation can be treated together. The EMOS model of Scheuerer (2014) utilizes a censored GEV distribution with shape parameter ensuring a positive skew and finite mean, whereas our EMOS approach is based on a CSG distribution appearing in the more complex model of Scheuerer and Hamill (2015).

2.1 Censored and shifted gamma EMOS model

Consider a gamma distribution $\Gamma(k,\theta)$ with shape $k>0$ and scale $\theta>0$ having PDF

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

where $\Gamma(k)$ denotes value of the gamma function at k. A gamma distribution can also be parametrized by its mean $\mu > 0$ and standard deviation $\sigma > 0$ using expressions

$$k = \frac{\mu^2}{\sigma^2} \qquad \text{and} \qquad \theta = \frac{\sigma^2}{\mu}.$$

Now, let $\delta > 0$ and denote by $G_{k,\theta}$ the CDF of the $\Gamma(k, \theta)$ distribution. Then the shifted gamma distribution left censored at zero (CSG) $\Gamma^0(k, \theta, \delta)$ with shape k, scale θ and shift δ can be defined with CDF

$$G^0_{k,\theta,\delta}(x) := \begin{cases} G_{k,\theta}(x + \delta), & x \geq 0, \\ 0, & x < 0. \end{cases} \tag{2.1}$$

This distribution assigns mass $G_{k,\theta}(\delta)$ to the origin and has generalized PDF

$$g^0_{k,\theta,\delta}(x) := \mathbb{I}_{\{x=0\}} G_{k,\theta}(\delta) + \mathbb{I}_{\{x>0\}} \big(1 - G_{k,\theta}(\delta)\big) g_{k,\theta}(x + \delta),$$

where $\mathbb{I}_A$ denotes the indicator function of the set A. Short calculation shows that the mean κ of $\Gamma^0(k, \theta, \delta)$ equals

$$\kappa = \theta k \big(1 - G_{k,\theta}(\delta)\big)\big(1 - G_{k+1,\theta}(\delta)\big) - \delta\big(1 - G_{k,\theta}(\delta)\big)^2,$$

whereas the p-quantile q_p $(0 < p < 1)$ of (2.1) equals 0 if $p \leq G_{k,\theta}(\delta)$, and the solution of $G_{k,\theta}(q_p + \delta) = p$, otherwise.

Now, denote by $f_1, f_2, \ldots, f_m$ the ensemble of distinguishable forecasts of precipitation accumulation for a given location and time. This means that each ensemble member can be identified and tracked, which holds for example for the UWME (see Section 3.1) or for the COSMO-DE EPS of the DWD. In the proposed CSG EMOS model the ensemble members are linked to the mean μ and variance σ^2 of the underlying gamma distribution via equations

$$\mu = a_0 + a_1 f_1 + \cdots + a_m f_m \qquad \text{and} \qquad \sigma^2 = b_0 + b_1 \overline{f}, \tag{2.2}$$

where $\overline{f}$ denotes the ensemble mean. Mean parameters $a_0, a_1, \ldots, a_m \geq 0$ and variance parameters $b_0, b_1 \geq 0$ of model (2.2) 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 (see Section 2.2).

However, most of the currently used EPSs produce ensembles containing groups of statistically indistinguishable ensemble members which are obtained with the help of random perturbations of the initial conditions. This is the case for the ALADIN-HUNEPS ensemble described in Section 3.2 or for the 51 member ECMWF ensemble. The existence of several exchangeable groups is also a natural property of some multi-model EPSs such as the the THORPEX Interactive Grand Global Ensemble (Swinbank *et al.*, 2015) or the GLAMEPS ensemble (Iversen *et al.*, 2011).

Suppose we have M ensemble members divided into m exchangeable groups, where the kth group contains $M_k \geq 1$ ensemble members, such that $\sum_{k=1}^m M_k = M$. Further, we denote by $f_{k,\ell}$ the ℓth member of the kth group. In this situation ensemble members

within a given group should share the same parameters (Gneiting, 2014) resulting in the exchangeable version

$$\mu = a_0 + a_1 \sum_{\ell_1=1}^{M_1} f_{1,\ell_1} + \cdots + a_m \sum_{\ell_m=1}^{M_m} f_{m,\ell_m}, \qquad \sigma^2 = b_0 + b_1 \overline{f}, \tag{2.3}$$

of model (2.2).

Note, that the expression of the mean (or location) as an affine function of the ensemble is general in EMOS post-processing (see e.g. Thorarinsdottir and Gneiting, 2010; Scheuerer, 2014; Baran and Lerch, 2015), whereas the dependence of the variance parameter on the ensemble mean is similar to the expression of the variance in the gamma BMA model of Sloughter *et al.* (2007), and it is in line with the relation of forecast uncertainty to the expected precipitation amount mentioned in the Introduction. Moreover, practical tests show that, at least for the UWME and ALADIN-HUNEPS ensemble considered in the case studies of Section 5, models (2.2) and (2.3), respectively, significantly outperform the corresponding CSG EMOS models with variance parameters

$$\sigma^2 = b_0 + b_1 S^2 \qquad \text{and} \qquad \sigma^2 = b_0 + b_1 \mathrm{MD},$$

where

$$S^2 := \frac{1}{m-1} \sum_{k=1}^{m} \left(f_k - \overline{f}\right)^2 \qquad \text{and} \qquad \mathrm{MD} := \frac{1}{m^2} \sum_{k,\ell=1}^{m} \left|f_k - f_\ell\right|$$

are the ensemble variance and the more robust ensemble mean difference (Scheuerer, 2014), respectively. Further, compared to the proposed models, natural modifications

$$\sigma^2 = b_0 + b_1 S^2 + b_2 \overline{f} \qquad \text{or} \qquad \sigma^2 = (b_0 + b_1 \overline{f})^2$$

in the CSG EMOS variance structure do not result in improved forecasts skills.

2.2 Parameter estimation

The main aim of probabilistic forecasting is to access the maximal sharpness of the predictive distribution subject to calibration (Gneiting *et al.*, 2007). The latter means a statistical consistency between the predictive distributions and the validating observations, whereas the former refers to the concentration of the predictive distribution. This goal can be addressed 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). In atmospheric sciences the most popular scoring rules for evaluating predictive distributions are the logarithmic score, i.e. the negative logarithm of the predictive PDF evaluated at the verifying observation (Gneiting and Raftery, 2007), and the continuous ranked probability score (CRPS; Gneiting and Raftery, 2007; Wilks, 2011). For a predictive CDF $F(y)$ and an observation x the CRPS is defined as

$$\mathrm{CRPS}\left(F, x\right) := \int_{-\infty}^{\infty} \left(F(y) - \mathbb{1}_{\{y \geq x\}}\right)^2 \mathrm{d}y = \mathsf{E}|X - x| - \frac{1}{2}\mathsf{E}|X - X'|, \tag{2.4}$$

where $\mathbb{1}_H$ denotes the indicator of a set H, while X and X' are independent random variables with CDF F and finite first moment. The CRPS can be expressed in the same units as the observation and one should also note that both scoring rules are proper (Gneiting and Raftery, 2007) and negatively oriented, that is the smaller the better.

For a CSG distribution defined by (2.1) the CRPS can be expressed in a closed form, Scheuerer and Hamill (2015) showed that

$$\begin{aligned}\mathrm{CRPS}\left(G^0(k,\theta,\delta),x\right) = {} & (x+\delta)\Big(2G_{k,\delta}(x+\delta)-1\Big) - \frac{\theta k}{\pi}B\big(1/2,k+1/2\big)\Big(1-G_{2k,\delta}(2\delta)\Big) \\ & + \theta k\Big(1+2G_{k,\delta}(\delta)G_{k+1,\delta}(\delta) - G^2_{k,\delta}(\delta) - 2G_{k+1,\delta}(y+\delta)\Big) - \delta G^2_{k,\delta}(\delta).\end{aligned}$$

Following the ideas of Gneiting *et al.* (2005) and Scheuerer (2014), the parameters of models (2.2) (and (2.3) as well) are estimated by minimizing the mean CRPS of predictive distributions and validating observations corresponding to forecast cases of the training period. We remark that optimization with respect to the mean logarithmic score, that is, maximum likelihood (ML) estimation of parameters, has also been investigated. Obviously, in terms of CRPS this model cannot outperform the one fit via CRPS minimization, however, in our test cases the ML method results in a reduction of the predictive skill of the CSG EMOS model in terms of almost all verification scores considered, so the corresponding values are not reported.

3 Data

3.1 University of Washington mesoscale ensemble

The eight-member UWME covers the Pacific Northwest region of North America and operates on a 12 km grid. The ensemble members 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.*, 1995) with initial and boundary conditions from various weather centres. We consider 48 h forecasts and corresponding validating observations of 24 h precipitation accumulation for 152 stations in the Automated Surface Observing Network (National Weather Service, 1998) in five US states. The forecasts are initialized at 0 UTC (5 PM local time when daylight saving time (DST) is in use and 4 PM otherwise) and we investigate data for calendar year 2008 with additional forecasts and observations from the last three months of 2007 used for parameter estimation. After removing days and locations with missing data 83 stations remain resulting in 20 522 forecast cases for 2008.

Figure 1a 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 for all forecast cases (see e.g. Wilks, 2011, Section 7.7.2), where zero observations are randomized among all zero forecasts. This histogram is far from the desired uniform distribution as in many cases the ensemble members overestimate the validating observation. The ensemble range contains the observed precipitation accumulation in 67.82 % of the cases,

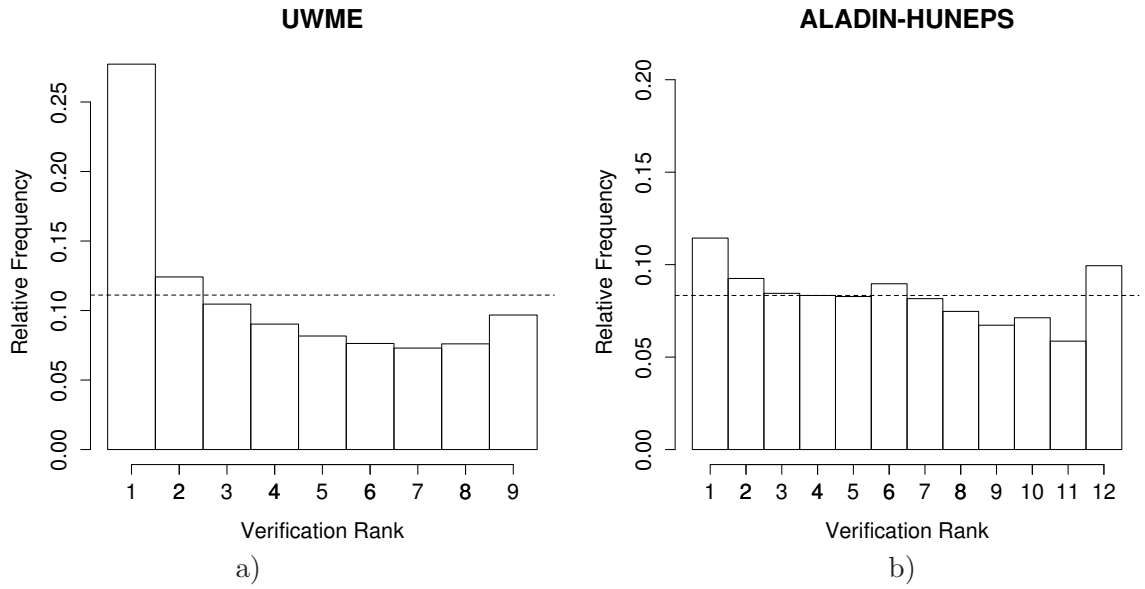

Figure 1: Verification rank histograms. a) UWME for the calendar year 2008; ALADIN-HUNEPS ensemble for the period 1 October 2010 – 25 March 2011.

whereas the nominal coverage of the ensemble equals 7/9, i.e 77.78 %. Hence, the UWME is uncalibrated, and would require statistical post-processing to yield an improved forecast probability density function.

3.2 ALADIN-HUNEPS ensemble

The ensemble forecasts produced by the operational ALADIN-HUNEPS system of the HMS are obtained with dynamical downscaling of the global PEARP system of Météo France by the ALADIN limited area model with an 8 km horizontal resolution. The EPS covers a large part of continental Europe and has 11 ensemble members, 10 exchangeable forecasts from perturbed initial conditions and one control member from the unperturbed analysis (Horányi *et al.*, 2011). The data base at hand contains ensembles of 42 h forecasts (initialized at 18 UTC, i.e. 8 pm local time when DST operates and 7 pm otherwise) for 24 h precipitation accumulation for 10 major cities in Hungary (Miskolc, Sopron, Szombathely, Győr, Budapest, Debrecen, Nyíregyháza, Nagykanizsa, Pécs, Szeged) together with the corresponding validating observations for the period between 1 October 2010 and 25 March 2011. The data set is fairly complete since there are only two dates when three ensemble members are missing for all sites. These dates are excluded from the analysis.

The verification rank histogram of the raw ensemble, displayed in Figure 1b, shows far better calibration, than that of the UWME. The coverage of the ALADIN-HUNEPS ensemble equals 84.20 %, which is very close to the nominal value of 83.33 % (10/12).

4 Results

As mentioned earlier, the predictive performance of the CSG EMOS model is tested on ensemble forecasts produced by the UWME and ALADIN-HUNEPS EPSs, and the results are compared with the fits of the GEV EMOS and gamma BMA models investigated by Scheuerer (2014) and Sloughter *et al.* (2007), respectively, and the verification scores of the raw ensemble. We remark that according to the suggestions of Scheuerer (2014) for estimating the parameters of the GEV EMOS model for a given day, the estimates for the preceding day serve as initial conditions for the box constrained Broyden-Fletcher-Goldfarb-Shanno (Byrd *et al.*, 1995) optimization algorithm. Compared with the case of fixed initial conditions this approach results in a slight increase of the forecast skills of the GEV EMOS model, whereas for the CSG EMOS method, at least in our case studies, fixed initial conditions are preferred. Further, we consider regional (or global) EMOS approach (see e.g. Thorarinsdottir and Gneiting, 2010) which is based on ensemble forecasts and validating observations from all available stations during the rolling training period and consequently results in a single universal set of parameters across the entire ensemble domain.

4.1 Diagnostics

To get the first insight about the calibration of EMOS and BMA post-processed forecasts we consider probability integral transform (PIT) histograms. Generally, the PIT is the value of the predictive CDF evaluated at the verifying observation (Raftery *et al.*, 2005), however, for our discrete-continuous models in the case of zero observed precipitation a random value is chosen uniformly from the interval between zero and the probability of no precipitation (Sloughter *et al.*, 2007). Obviously, the closer the histogram to the uniform distribution, the better the calibration. In this way the PIT histogram is the continuous counterpart of the verification rank histogram of the raw ensemble and provides a good measure about the possible improvements in calibration.

The predictive performance of probabilistic forecasts is quantified with the help of the mean CRPS over all forecast cases, where for the raw ensemble the predictive CDF is replaced by the empirical one. Further, as suggested by Gneiting and Ranjan (2011), Diebold-Mariano (DM; Diebold and Mariano, 1995) tests are applied for investigating the significance of the differences in scores corresponding to the various post-processing methods. The DM test takes into account the dependence in the forecasts errors and for this reason it is widely used in econometrics.

Besides the CRPS we also consider Brier scores (BS; Wilks, 2011, Section 8.4.2) for the dichotomous event that the observed precipitation amount x exceeds a given threshold y. For a predictive CDF $F(y)$ the probability of this event is $1-F(y)$, and the corresponding Brier score is given by

$$\mathrm{BS}\left(F, x; y\right) := \left(F(y) - \mathbb{1}_{\{y \geq x\}}\right)^2, \tag{4.1}$$

see e.g. Gneiting and Ranjan (2011). Obviously, the BS is negatively oriented and the CRPS (2.4) is the integral of the BSs over all possible thresholds. In our case studies we consider 0 mm precipitation, 5, 15, 25, 30 mm and 1, 5, 7, 9 mm threshold values for the UWME

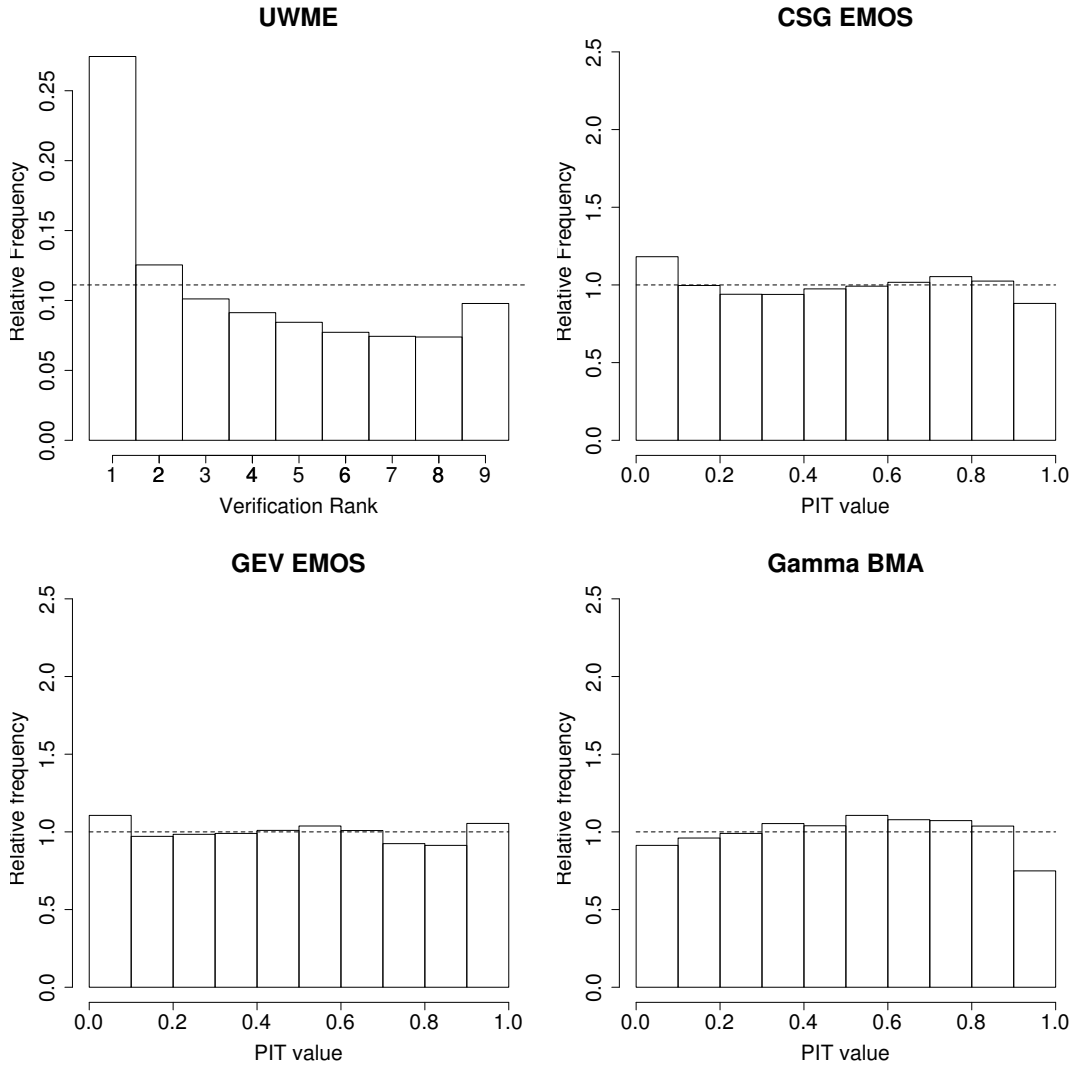

Figure 2: Verification rank histogram of the raw ensemble and PIT histograms of the EMOS and BMA post-processed forecasts for the UWME for the calendar year 2008.

Table 1: p-values of Kolmogorov-Smirnov tests for uniformity of PIT values for the UWME. Means of 10000 random samples of sizes 2500 each.

Model	CSG EMOS	GEV EMOS	Gamma BMA
Mean p-value	0.154	0.310	0.044

and ALADIN-HUNEPS ensemble, respectively, corresponding approximately to the 45th, 75th, 85th and 90th percentiles of the observed non-zero precipitation accumulations, and compare the mean BSs of the pairs of predictive CDFs and verifying observations over all forecast cases.

The improvement in CRPS and BS with respect to a reference predictive distribution

F_{ref} can be measured with the help of the continuous ranked probability skill score (CRPSS) and the Brier skill score (BSS) defined as

$$\mathrm{CRPSS}\left(F,x\right) := 1 - \frac{\mathrm{CRPS}\left(F,x\right)}{\mathrm{CRPSS}\left(F_{ref},x\right)} \quad \text{and} \quad \mathrm{BSS}\left(F,x;y\right) := 1 - \frac{\mathrm{BS}\left(F,x;y\right)}{\mathrm{BS}\left(F_{ref},x;y\right)},$$

respectively (Wilks, 2011; Friedrichs and Thorarinsdottir, 2012). These scores are positively oriented and in our two case studies we use the raw ensemble as a reference.

To compare the calibration of probabilities of a dichotomous event of exceeding a given threshold calculated from the raw ensemble and the EMOS and BMA predictive distributions, we make use of reliability diagrams (Wilks, 2011, Section 8.4.4). The reliability diagram plots the a graph of the observed frequency of the event against the binned forecast frequencies and in the ideal case this graph should lie on the main diagonal of the unit square. In the case studies of Sections 4.1 and 4.2 we consider the same thresholds as for the BSs (UWME: 5, 15, 25, 30 mm; ALADIN-HUNEPS: 1, 5, 7, 9 mm;), whereas the unit interval is divided into 11 bins with break points $0.05, 0.15, 0.25, \ldots, 0.95$. Following Bröcker and Smith (2007) and Scheuerer (2014), the observed relative frequency of a bin is plotted against the mean of the corresponding probabilities, and we also add inset histograms displaying the frequencies of the different bins on log 10 scales.

Further, one can investigate calibration and sharpness of a predictive distribution with the help of the coverage and average width of the $(1-\alpha)100\,\%$, $\alpha \in (0,1)$, central prediction interval. By coverage we mean the proportion of validating observations located between the lower and upper $\alpha/2$ quantiles of the predictive CDF and level α should be chosen to match the nominal coverage of the raw ensemble, i.e. 77.78 % for the UWME and 83.33 % for the ALADIN-HUNEPS. As the coverage of a calibrated predictive distribution should be around $(1-\alpha)100\,\%$, the suggested choices of α allow direct comparisons with the raw ensembles, whereas the average width of the central prediction interval assesses the sharpness of the forecast.

Finally, point forecasts such as EMOS, BMA and ensemble medians are evaluated with the help of mean absolute errors (MAEs) and DM tests for the forecast errors are applied to check whether the differences are significant.

4.2 Verification results for the UWME

The eight members of the UWME are generated using initial and boundary conditions from different sources, implying that the ensemble members are clearly distinguishable. Hence, the mean and the variance of the underlying gamma distribution of the CSG EMOS model are linked to the ensemble members according to (2.2) with $m = 8$. Obviously, the reference GEV EMOS and gamma BMA models are also formulated under the assumption of non-exchangeable ensemble members.

A detailed study of CRPS and MAE values of the CSG EMOS and gamma BMA models corresponding to training period lengths of $20, 25, \ldots, 100$ days indicates that both scores have global minima at 70 days. Hence, in our analysis we calibrate the UWME forecasts for calendar year 2008 using this training period length.

Table 2: Mean CRPS of probabilistic forecasts, MAE of median forecasts and coverage and average width of 77.78 % central prediction intervals for the UWME.

Forecast	CRPS (mm)	MAE (mm)	Coverage (%)	Av.width (mm)
CSG EMOS	2.252	3.019	80.46	8.350
GEV EMOS	2.283	3.033	79.91	8.683
Gamma BMA	2.357	3.220	83.44	9.515
Ensemble	2.929	3.708	67.95	8.599

Table 3: Values of the test statistics of the DM test for equal predictive performance based on the CRPS (*upper triangle*) and the prediction error of the median forecast (*lower triangle*) for the UWME. Negative/positive values indicate a superior predictive performance of the forecast given in the row/column label, bold numbers correspond to tests with p values under 0.05 level of significance.

Forecast	CSG EMOS	GEV EMOS	Gamma BMA	Ensemble
CSG EMOS	–	**-5.237**	**-4.909**	**-29.265**
GEV EMOS	1.631	–	**-3.688**	**-26.845**
Gamma BMA	**5.648**	**5.892**	–	**-15.556**
Ensemble	**21.967**	**20.504**	**9.076**	–

Table 4: CRPSS and BSS values with respect to the raw UWME.

Forecast	CRPSS	Brier Skill Score				
		0 mm	5 mm	15 mm	25 mm	30 mm
CSG EMOS	0.231	0.393	0.243	0.268	0.248	0.237
GEV EMOS	0.221	0.403	0.219	0.252	0.239	0.235
Gamma BMA	0.196	0.419	0.231	0.240	0.196	0.188

Figure 2 showing the verification rank histogram of the raw ensemble and the PIT histograms of the CSG EMOS, GEV EMOS and gamma BMA models clearly illustrates the advantage of statistical post-processing. Unfortunately, the Kolmogorov-Smirnov (KS) test rejects the uniformity of the PIT values for all models, the highest p-value of 5.562×10^{-3} corresponds to the GEV EMOS approach. However, the small p-values are consequences of numerical problems caused by the large sample size (see e.g. Baran *et al.*, 2013) and the mean p-values of 10000 random samples of PITs of sizes 2500 each, given in Table 1, nicely follow the shapes of the histograms of Figure 2.

In Table 2 the mean CRPS of probabilistic forecasts, the MAE of median forecasts and the coverage and average width of 77.78 % central prediction intervals for the two EMOS approaches, the gamma BMA model and the raw ensemble are reported, whereas Table 3 shows the results of DM tests for equal predictive performance based on the CRPS values and

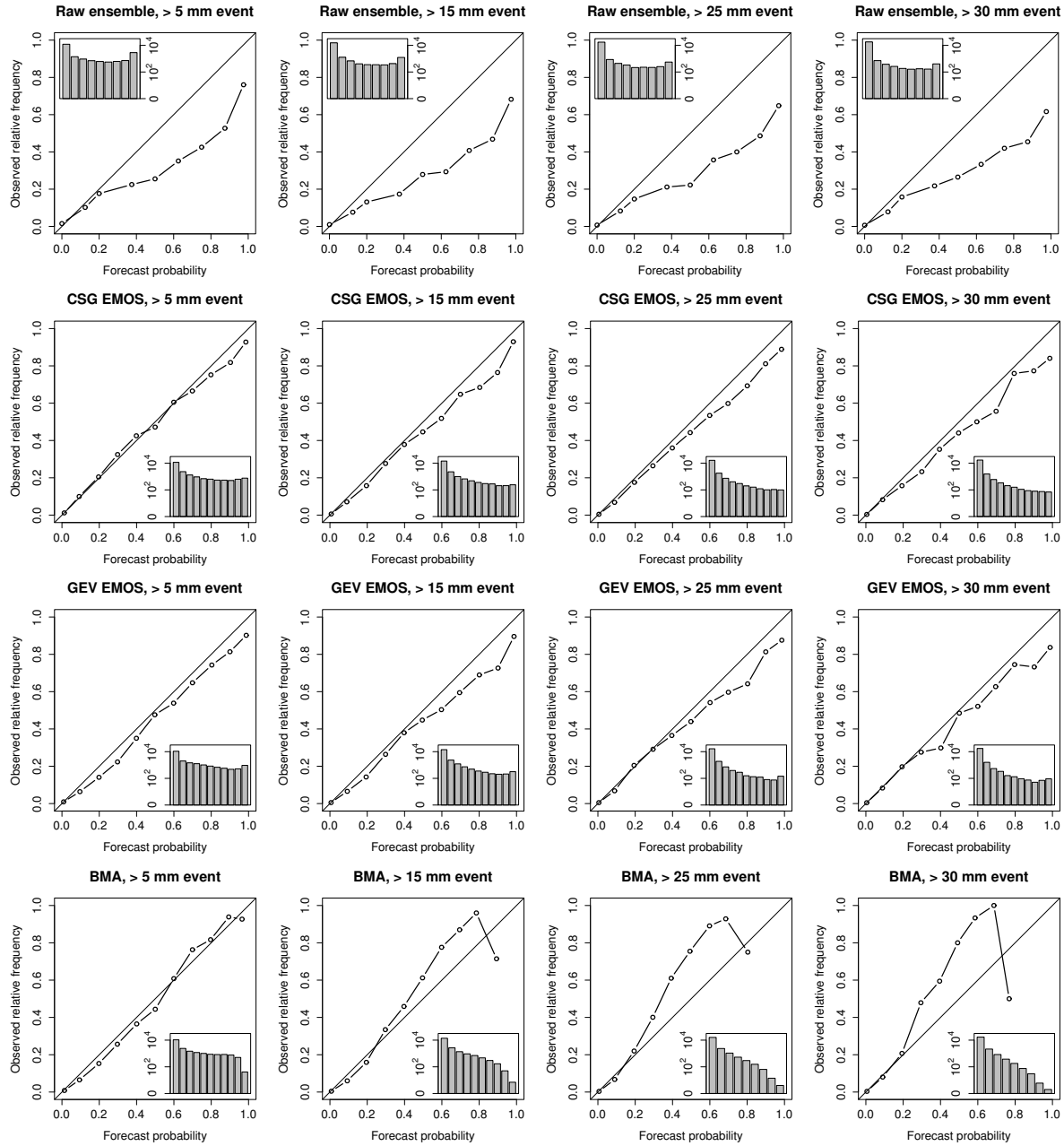

Figure 3: Reliability diagrams of the raw ensemble and EMOS and BMA post-processed forecasts for the UWME for the calendar year 2008. The inset histograms display the log-frequency of cases within the respective bins.

the prediction errors of the median. By examining these results, one can clearly observe the obvious advantage of post-processing with respect to the raw ensemble, which is quantified in the significant decrease of CRPS and MAE values and in a substantial improvement in coverage. Further, the CSG EMOS model results in the lowest CRPS value, whereas in terms MAE there is no difference between the two EMOS methods, which significantly outperform the gamma BMA approach both in calibration of probabilistic and accuracy of

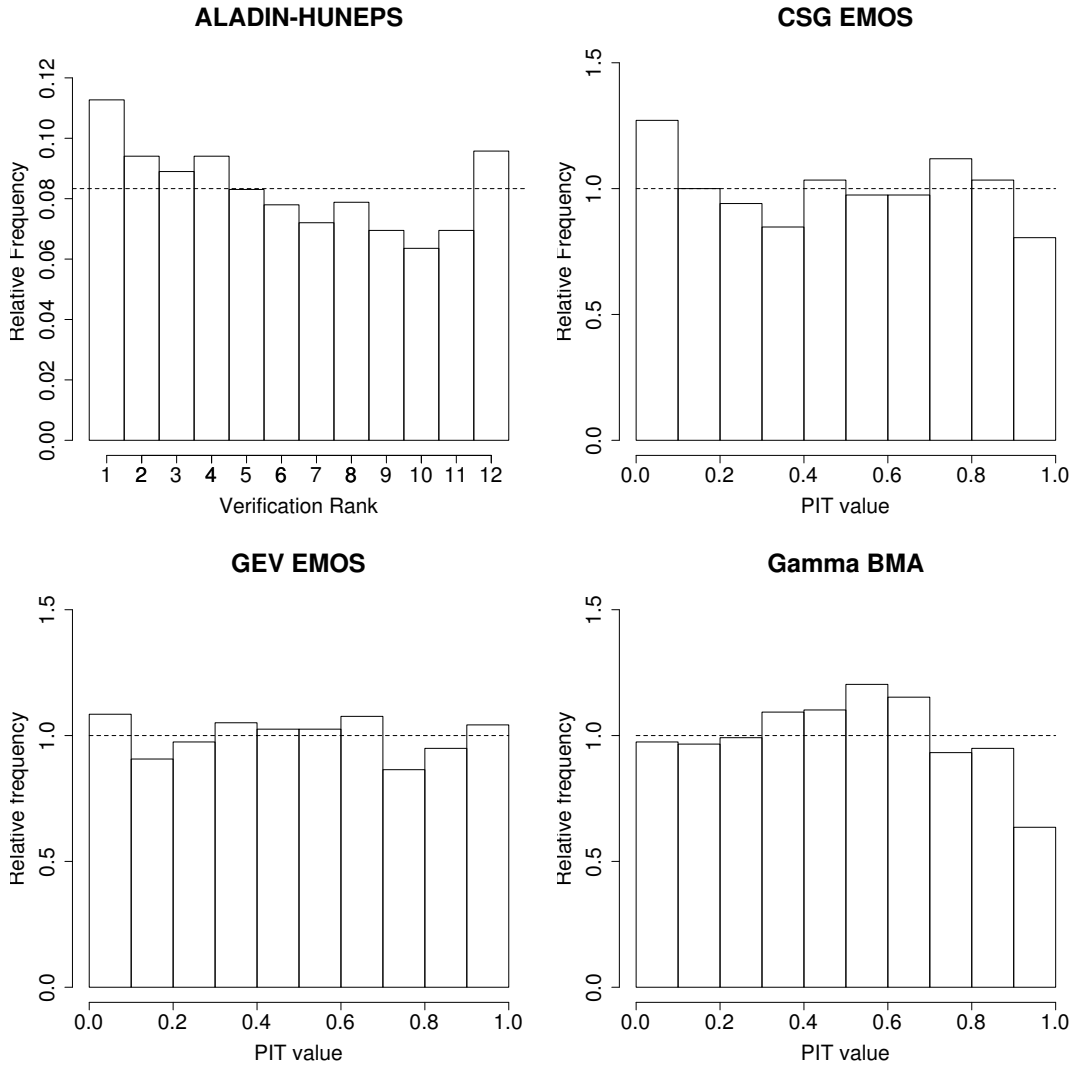

Figure 4: Verification rank histogram of the raw ensemble and PIT histograms of the EMOS and BMA post-processed forecasts for the ALADIN-HUNEPS ensemble for the period 27 November 2010 – 25 March 2011.

point forecasts. The CSG EMOS model results in the sharpest central prediction interval combined with a rather fair coverage, whereas the central prediction intervals corresponding to the other two calibration methods are slightly wider than that of the raw ensemble.

The improvement in calibration caused by statistical post-processing can also be observed in skill scores reported in Table 4 and reliability diagrams displayed in Figure 3. Gamma BMA method performs well in predicting the probability of positive precipitation and exceeding the 5 mm threshold, whereas for higher threshold values it is behind the two EMOS approaches, where the CSG EMOS model presents slightly better forecast skills. Hence, one can conclude, that in case of the UWME the EMOS approaches outperform both the raw ensemble and the gamma BMA model and the proposed CSG EMOS model slightly outperforms the GEV EMOS method.

Table 5: p-values of Kolmogorov-Smirnov tests for uniformity of PIT values for the ALADIN-HUNEPS ensemble.

Model	CSG EMOS	GEV EMOS	Gamma BMA
p-value	0.119	0.921	0.003

Table 6: Mean CRPS of probabilistic forecasts, MAE of median forecasts and coverage and average width of 83.33 % central prediction intervals for the ALADIN-HUNEPS ensemble.

Forecast	CRPS (mm)	MAE (mm)	Coverage (%)	Av.width (mm)
CSG EMOS	0.465	0.636	89.15	2.185
GEV EMOS	0.477	0.641	86.53	2.192
Gamma BMA	0.532	0.708	93.73	2.854
Ensemble	0.485	0.640	84.24	2.436

4.3 Verification results for the ALADIN-HUNEPS ensemble

As a contrast to the UWME, the way the ALADIN-HUNEPS ensemble is generated (see Section 3.2) induces a natural grouping of the ensemble members. The first group contains the control, whereas the second group consists of the 10 exchangeable ensemble members. This splitting results in the GEV EMOS model (2.3) with $m = 2$, $M_1 = 1$ and $M_2 = 10$ and the same grouping is considered for the benchmarking GEV EMOS and gamma BMA models (Fraley *et al.*, 2010).

Again, in order to determine the appropriate length of the rolling training period the mean CRPS and MAE values of the various models for training periods of lengths $20, 25, \ldots, 100$ calendar days are investigated. In order to ensure the comparability of the results corresponding to different training period lengths, verification scores from 10 January to 25 March 2011 are considered. The corresponding curves of the CRPS and MAE scores plotted against the training period lengths (not shown) have global minima at 85 days, however they have elbows at 55 days, that is, up to this training period length the decrease is rather steep then the values stabilize. Hence, as in general shorter training periods are preferred, for calibrating the ALADIN-HUNEPS ensemble a training period of length 55 days is used. This means that ensemble members, validating observations, and predictive PDFs are available for the period from 27 November 2010 to 25 March 2011 having 119 calendar days (just after the first 55 day training period) and 1180 forecast cases, since on 15 February 2011 three ensemble members are missing and this date is excluded from the analysis. This time interval starts more than 6 weeks earlier than the one used for determination of the optimal training period length.

Compared with the verification rank histogram of the raw ensemble the PIT histograms of the post-processed forecasts displayed in Figure 4 show a substantial improvement in calibration. For the two EMOS models the KS test accepts the uniformity of the PIT values (see Table 5 and note the extremely high p-value for the GEV EMOS), whereas the histogram

Table 7: Values of the test statistics of the DM test for equal predictive performance based on the CRPS (*upper triangle*) and the prediction error of the median forecast (*lower triangle*) for the ALADIN-HUNEPS ensemble. Negative/positive values indicate a superior predictive performance of the forecast given in the row/column label, bold numbers correspond to tests with p values under 0.05 level of significance.

Forecast	CSG EMOS	GEV EMOS	Gamma BMA	Ensemble
CSG EMOS	–	**-2.758**	**-3.978**	**-2.928**
GEV EMOS	0.799	–	**-3.586**	-1.177
Gamma BMA	**2.682**	**2.697**	–	**2.498**
Ensemble	0.246	-0.078	**-2.109**	–

Table 8: CRPSS and BSS values with respect to the raw ALADIN-HUNEPS ensemble.

Forecast	CRPSS	Brier Skill Score				
		0 mm	1 mm	5 mm	7 mm	9 mm
CSG EMOS	0.042	0.094	0.057	-0.011	-0.025	0.019
GEV EMOS	0.017	0.166	0.008	-0.022	-0.030	0.027
Gamma BMA	-0.098	0.151	-0.070	-0.265	-0.136	-0.023

of the Gamma BMA model is hump shaped indicating some overdispersion.

Concerning the two EMOS approaches, the verification scores of Table 6 together with the results of the corresponding DM tests for equal predictive performance (see Table 7) display similar behavior as in the case of the UWME. There is no significant difference between the MAE values of the CSG and GEV EMOS methods and the former results in the lowest CRPS and the sharpest 83.33 % central prediction interval. Further, the EMOS models significantly outperform both the raw ensemble and the gamma BMA approach, despite the raw ensemble is rather well calibrated and has far better predictive skill than the BMA calibrated forecast. Note that the large mean CRPS and coverage of the BMA predictive distribution is totally in line with the shape of the corresponding PIT histogram of Figure 4.

The good predictive performance of the ALADIN-HUNEPS ensemble can also be observed on the large amount of negative skill scores reported in Table 8 and on the reliability diagrams of Figure 5. Similar to the case of the UWME, for 0 mm threshold the gamma BMA model has good predictive performance, whereas for higher threshold values it underperforms the CSG and GEV EMOS models and the raw ensemble. However, in connection with the reliability diagrams one should also note that the hectic behavior of the graphs (compared to the rather smooth diagrams of Figure 3) is a consequence of the shortage of data, as the verification period contains only 394 observations of positive precipitation, which is around one third of the forecast cases.

Taking into account both the uniformity of the PIT values and the verification scores

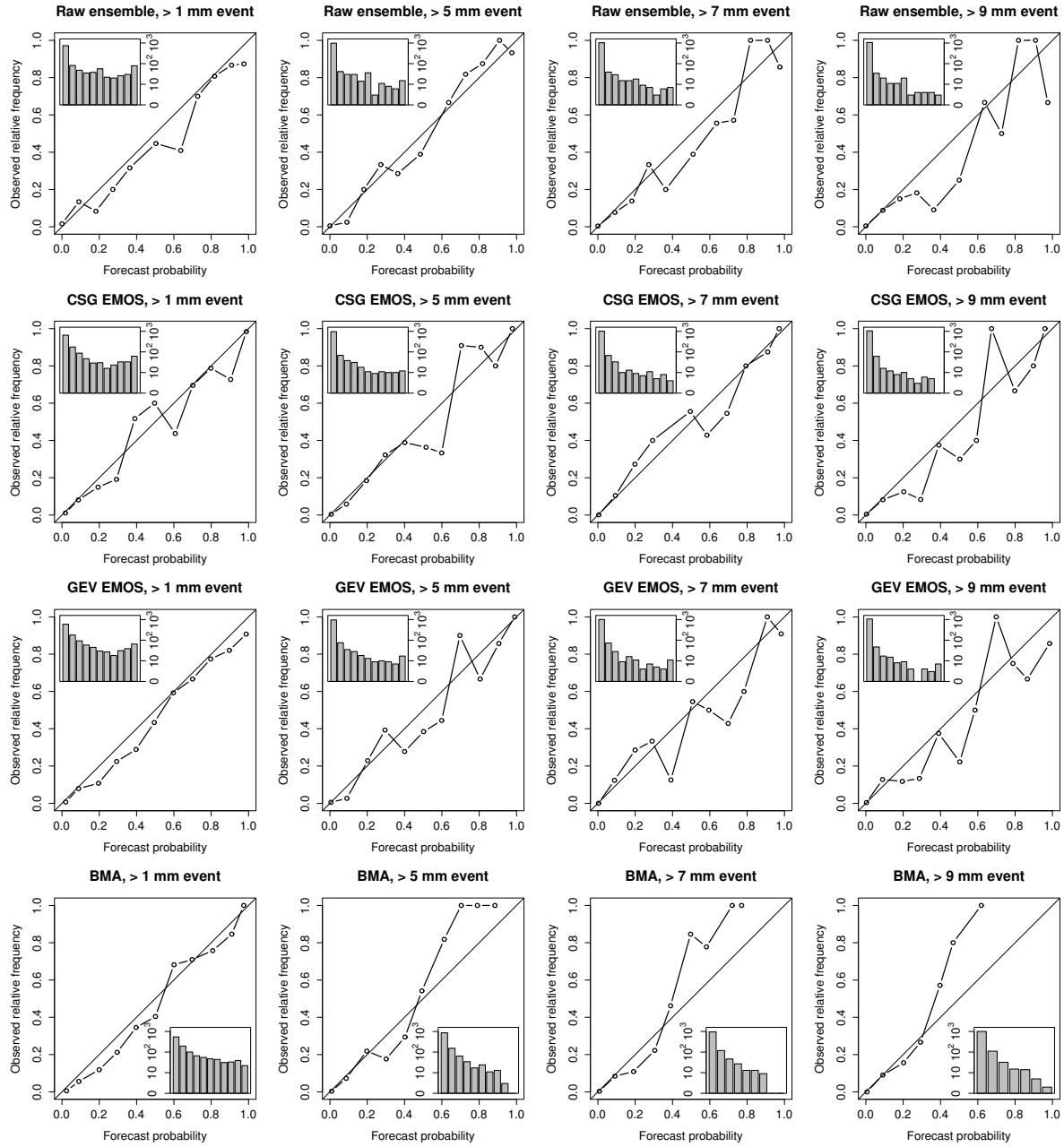

Figure 5: Reliability diagrams of the raw ensemble and EMOS and BMA post-processed forecasts for the ALADIN-HUNEPS ensemble for the period 27 November 2010 – 25 March 2011. The inset histograms display the log-frequency of cases within the respective bins.

in Tables 6 and 8 it can be said that the proposed CSG EMOS model has the best overall performance in calibration of the raw ALADIN-HUNEPS ensemble forecasts of precipitation accumulation.

5 Conclusions

A new EMOS model for calibrating ensemble forecasts of precipitation accumulation is proposed where the predictive distribution follows a censored and shifted gamma distribution, with mean and variance of the underlying gamma law being affine functions of the raw ensemble and the ensemble mean, respectively. The CSG EMOS method is tested on ensemble forecasts of 24 h precipitation accumulation of the eight-member University of Washington mesoscale ensemble and on the 11 member ALADIN-HUNEPS ensemble of the Hungarian Meteorological Service. These ensemble prediction systems differ both in the climate of the covered area and in the generation of the ensemble members. By investigating the uniformity of the PIT values of predictive distributions, the mean CRPS of probabilistic forecasts, the Brier scores and reliability diagrams for various thresholds, the MAE of median forecasts and the average width and coverage of central prediction intervals corresponding to the nominal coverage, the predictive skill of the new approach is compared with that of the GEV EMOS method (Scheuerer, 2014), the gamma BMA model (Sloughter *et al.*, 2007) and the raw ensemble. From the results of the presented case studies one can conclude that in terms of calibration of probabilistic and accuracy of point forecasts the proposed CSG EMOS model significantly outperforms both the raw ensemble and the BMA model and shows slightly better forecast skill than the GEV EMOS approach.

Acknowledgments. Sándor Baran is supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. Dóra Nemoda partially carried out her research in the framework of the Center of Excellence of Mechatronics and Logistics at the University of Miskolc. The authors are indebted to Michael Scheuerer for his useful suggestions and remarks and for providing the `R` code for the GEV EMOS model. The authors further thank the University of Washington MURI group for providing the UWME data and Mihály Szűcs from the HMS for the ALADIN-HUNEPS data.

References

Bao L, Gneiting T, Raftery AE, Grimit EP, Guttorp P. 2010. Bias correction and Bayesian model averaging for ensemble forecasts of surface wind direction. *Monthly Weather Review* **138**:1811–1821, DOI: 10.1175/2009mwr3138.1.

Baran S. 2014. Probabilistic wind speed forecasting using Bayesian model averaging with truncated normal components. *Computational Statistics and Data Analysis* **75**:227–238, DOI: 10.1016/j.csda.2014.02.013.

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, DOI: 10.1002/qj.2521.

Baran S, Sikolya K, Veress L. 2013. Estimating the risk of a Down's syndrome term pregnancy using age and serum markers: Comparison of various methods. *Communications in Statistics – Simulation and Computation* **42**:1654–1672, DOI: 10.1080/03610918.2012.674596.

Bouallègue ZB, Theis SE, Gebhardt C. 2013. Enhancing COSMO-DE ensemble forecasts by inexpensive techniques. *Meteorologische Zeitschrift* **22**:49–59, DOI: 10.1127/0941-2948/2013/0374.

Buizza R, Houtekamer PL, Toth Z, Pellerin G, Wei M, Zhu Y. 2005. A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Monthly Weather Review* **133**:1076–1097, DOI: 10.1175/mwr2905.1.

Broĉker J, Smith LA. 2007. Increasing the reliability of reliability diagrams. *Weather and Forecasting* **22**:651–661, DOI: 10.1175/waf993.1.

Byrd RH, Lu P, Nocedal J, Zhu C. 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing* **16**:1190–1208, DOI: 10.1137/0916069.

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, DOI: 10.1002/qj.2469.

Diebold FX, Mariano, RS. 1995. Comparing predictive accuracy. *Journal of Business & Economic Statistics* **13**:253–263, DOI: 10.1198/073500102753410444.

Eckel FA, Mass CF. 2005. Effective mesoscale, short-range ensemble forecasting. *Weather and Forecasting* **20**:328–350, DOI: 10.1175/waf843.1.

ECMWF Directorate 2012. Describing ECMWF's forecasts and forecasting system. *ECMWF Newsletter* **133**:11–13.

Fraley C, Raftery AE, Gneiting T. 2010. Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Monthly Weather Review* **138**:190–202, DOI: 10.1175/2009mwr3046.1.

Fraley C, Raftery AE, Gneiting T, Sloughter JM, Berrocal VJ. 2011. Probabilistic weather forecasting in R. *The R Journal* **3**:55–63.

Friederichs P, Thorarinsdottir TL. 2012. Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics* **23**:579–594, DOI: 10.1002/env.2176.

Gebhardt C, Theis SE, Paulat M, Bouallègue ZB. 2011. Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries. *Atmospheric Research* **100**:168–177, DOI: 10.1016/j.atmosres.2010.12.008.

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 24 November 2015]

Gneiting T, Balabdaoui F, Raftery AE. 2007. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B* **69**:243–268, DOI: 10.1111/j.1467-9868.2007.00587.x.

Gneiting T, Raftery AE. 2005. Weather forecasting with ensemble methods. *Science* **310**:248–249, DOI: 10.1126/science.1115255.

Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association* **102**:359–378, DOI: 10.1198/016214506000001437.

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, DOI: 10.1175/mwr2904.1.

Gneiting T, Ranjan R. 2011. Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics* **29**:411–422, DOI: 10.1198/jbes.2010.08110.

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://www2.mmm.ucar.edu/mm5/documents/mm5-desc-doc.html) [Accessed on 24 November 2015]

Hamill TM. 2007. Comments on "Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian Model Averaging." *Monthly Weather Review* **135**:4226–4230, DOI: 10.1175/2007mwr1963.1.

Hodyss D, Satterfield E, McLay J, Hamill TM, Scheuerer M. 2015. Inaccuracies with multimodel post-processing methods involving weighted, regression-corrected forecasts. *Monthly Weather Review* DOI: 10.1175/mwr-d-15-0204.1.

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.

Horányi A, Mile M, Szűcs M. 2011. Latest developments around the ALADIN operational short-range ensemble prediction system in Hungary. *Tellus A* **63**:642–651, DOI: 10.1111/j.1600-0870.2011.00518.x.

Iversen T, Deckmin A, Santos C, Sattler K, Bremnes JB, Feddersen H, Frogner I-L. 2011. Evaluation of 'GLAMEPS' – a proposed multimodel EPS for short range forecasting. *Tellus A* **63**:513–530, DOI: 10.1111/j.1600-0870.2010.00507.x.

Leith CE. 1974. Theoretical skill of Monte-Carlo forecasts. *Monthly Weather Review* **102**:409–418, DOI: 10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2.

Lerch S, Thorarinsdottir TL. 2013. Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus A* **65**:21206, DOI: 10.3402/tellusa.v65i0.21206.

Messner JW, Mayr GJ, Zeileis A, Wilks DS. 2014. Heteroscedastic extended logistic regression for postprocessing of ensemble guidance. *Monthly Weather Review* **142**:448–456, DOI: 10.1175/mwr-d-13-00271.1.

National Weather Service. 1998. *Automated Surface Observing System (ASOS) Users Guide.* (Available from: http://www.weather.gov/asos/aum-toc.pdf) [Accessed on 24 November 2015]

Raftery AE, Gneiting T, Balabdaoui F, Polakowski M. 2005. Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review* **133**:1155–1174, DOI: 10.1175/mwr2906.1.

Scheuerer M. 2014. Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society* **149**:1086–1096, DOI: 10.1002/qj.2183.

Scheuerer M, Hamill TM. 2015. Statistical post-processing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions *Monthly Weather Review* **143**:4578–4596, DOI: 10.1175/mwr-d-15-0061.1.

Scheuerer M, Möller D. 2015. Probabilistic wind speed forecasting on a grid based on ensemble model output statistics. *Annals of Applied Statistics* **9**:1328–1349, DOI: 10.1214/15-aoas843

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, DOI: 10.1198/jasa.2009.ap08615.

Sloughter JM, Raftery AE, Gneiting T, Fraley C. 2007. Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review* **135**:3209–3220, DOI: 10.1175/mwr3441.1.

Swinbank R, Kyouda M, Buchanan P, Froude L, Hamill TM, Hewson TD, Keller JH, Matsueda M, Methven J, Pappenberger F, Scheuerer M, Titley HA, Wilson L, Yamaguchi M. 2015. The TIGGE project and its achievements. *Bulletin of the American Meteorological Society*, DOI: 10.1175/bams-d-13-00191.1.

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, DOI: 10.1111/j.1467-985X.2009.00616.x.

Williams RM, Ferro CAT, Kwasniok F. 2014. A comparison of ensemble post-processing methods for extreme events. *Quarterly Journal of the Royal Meteorological Society* **140**:1112–1120, DOI: 10.1002/qj.2198.

Wilks DS. 2009. Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteorological Applications* **16**:361–368, DOI: 10.1002/met.134.

Wilks DS. 2011. *Statistical Methods in the Atmospheric Sciences* (3rd ed.) Elsevier: Amsterdam.