Meteorologische Zeitschrift, Vol. 22, No. 3, 1–10 (August 2013)

Article

Statistical post-processing of probabilistic wind speed forecasting in Hungary

SÁNDOR BARAN[1,*], ANDRÁS HORÁNYI[2] and DÓRA NEMODA[1]

[1]Faculty of Informatics, University of Debrecen, Hungary
[2]Hungarian Meteorological Service, Hungary

(Manuscript received August 28, 2012; in revised form February 11, 2013; accepted March 6, 2013)

Abstract

Weather forecasting is mostly based on the outputs of deterministic numerical weather forecasting models. Multiple runs of these models with different initial conditions result in a forecast ensemble which is applied for estimating the future distribution of atmospheric variables. However, as these ensembles are usually under-dispersive and uncalibrated, post-processing is required. In the present work, Bayesian Model Averaging (BMA) is applied for calibrating ensembles of wind speed forecasts produced by the operational Limited Area Model Ensemble Prediction System of the Hungarian Meteorological Service (HMS). We describe two possible BMA models for wind speed data of the HMS and show that BMA post-processing significantly improves the calibration and accuracy of point forecasts.

Keywords: Bayesian Model Averaging, gamma distribution, continuous ranked probability score.

1 Introduction

The aim of weather forecasting is to give a reliable prediction of the future states of the atmosphere on the basis of present observations, prior forecasts and mathematical models describing the dynamics (physical behaviour) of the atmosphere. These models consist of sets of non-linear partial differential equations which can only be solved numerically. The problem with these numerical weather prediction models is that the solutions depend strongly on the initial conditions and also on other uncertainties related to the numerical weather prediction process. Therefore, the results of such models are never fully accurate. A possible solution to address this problem is to run the model with different initial conditions and produce an ensemble of forecasts. With the help of an ensemble, one can estimate the probability distribution of future weather variables which allows probabilistic weather forecasting (GNEITING and RAFTERY, 2005), where not only the future atmospheric states are predicted, but also the related uncertainty information. The ensemble prediction method was proposed by LEITH (1974) and since its first operational implementation (BUIZZA, et al., 1993; TOTH and KALNAY, 1997), it has became a widely used technique all over the world. However, although, e.g. the ensemble mean on average yields better forecasts of a meteorological quantity than any of the individual ensemble members, the ensemble is usually under-dispersive and in this way, uncalibrated. This characteristic has been observed with several operational ensemble prediction systems. For an overview, see e.g. BUIZZA et al. (2005).

The Bayesian Model Averaging (BMA) method for post-processing ensembles in order to calibrate them was introduced by RAFTERY et al. (2005). The basic idea of BMA is that for each ensemble member forecast, there is a corresponding conditional probability density function (PDF) that can be interpreted as the conditional PDF of the future weather quantity provided the considered forecast is the best one. The BMA predictive PDF of the future weather quantity is then the weighted sum of the individual PDFs corresponding to the ensemble members and the weights are based on the relative performances of the ensemble members during a given training period. In RAFTERY et al. (2005), the BMA method was successfully applied to obtain 48 hour forecasts of surface temperature and sea level pressure in the North American Pacific Northwest based on the 5 members of the University of Washington Mesoscale Ensemble (GRIMIT and MASS, 2002). These weather quantities can be modeled by normal distributions, so the predictive PDF is a Gaussian mixture. Later, SLOUGHTER et al. (2007) developed a discrete-continuous BMA model for precipitation forecasting, where the discrete part corresponds to the event of no precipitation, while the cubic root of the precipitation amount (if it is positive) is modeled by a gamma distribution. In SLOUGHTER et al. (2010), the BMA method was used for wind speed forecasting and the component PDFs follow a gamma distribution. Finally, using a von Mises distribution to model angular data, BAO et al. (2010) introduced a BMA scheme to predict surface wind direction.

In the present work, we apply the BMA method for calibrating ensemble forecasts of wind speed produced

*Corresponding author: Sándor Baran, Faculty of Informatics, University of Debrecen Kassai út 26, 4028 Debrecen, Hungary,
e-mail: baran.sandor@inf.unideb.hu

DOI 10.1127/0941-2948/2013/0428
0941-2948/2013/0428 $ 4.50

by the operational Limited Area Model Ensemble Prediction System (LAMEPS) of the Hungarian Meteorological Service (HMS) called ALADIN-HUNEPS (HÁGEL, 2010; HORÁNYI et al., 2011). ALADIN-HUNEPS covers a large part of Continental Europe with a horizontal resolution of 12 km and it is obtained by dynamical downscaling (by the ALADIN limited area model) of the global ARPEGE based PEARP system of Météo France (HORÁNYI et al., 2006; DESCAMPS et al., 2009). The ensemble consists of 11 members, 10 initialized from perturbed initial conditions and one control member from the unperturbed analysis. As this construction implies that the ensemble contains groups of exchangeable forecasts (the ensemble members cannot be distinguished), for post-processing one has to use the modification of BMA as suggested by FRALEY et al. (2010).

2 Data

As was mentioned in the introduction, BMA post-processing of ensemble predictions was applied for wind speed data obtained from the ALADIN-HUNEPS system. The data base contains 11 member ensembles (10 forecasts started from perturbed initial conditions and one control) of 42 hour forecasts for 10 meter wind speed (given in m/s) for 10 major cities in Hungary (Miskolc, Szombathely, Győr, Budapest, Debrecen, Nyíregyháza, Nagykanizsa, Pécs, Kecskemét, Szeged) produced by the ALADIN-HUNEPS system of the HMS, together with the corresponding validating observations for the period between October 1, 2010 and March 25, 2011 (176 days, or 1760 data points). The forecasts are initialized at 18 UTC. The startup speed of the anemometers measuring the validating observations is 0.1 m/s. The data set is fairly complete since there are only two days (18.10.2010 and 15.02.2011) where three ensemble members are missing for all sites and one day (20.11.2010) when no forecasts are available.

Fig. 1 shows the verification rank histogram of the raw ensemble. This is the histogram of ranks of validating observations with respect to the corresponding ensemble forecasts computed from the ranks at all stations and over the whole verification period (see e.g. WILKS, 2006, Section 7.7.2). This histogram is far from the desired uniform distribution as in many cases the ensemble members either underestimate or overestimate the validating observations (the ensemble range contains the observed wind speed in 61.21% of the cases, while its nominal value equals 10/12, i.e 83.33%). Hence, the ensemble is under-dispersive and in this way it is uncalibrated. Therefore, statistical post-processing is required to improve the forecasted probability density function.

3 The model and diagnostics

To obtain a probabilistic wind speed forecast, the modification of the BMA gamma model of SLOUGHTER et al. (2010) for an ensemble with exchangeable members (FRALEY et al., 2010) was used. The first idea is to have two exchangeable groups. One contains the control denoted by f_c while in the other are 10 ensemble members corresponding to the different perturbed initial conditions denoted by $f_{\ell,1}, \ldots, f_{\ell,10}$. In this way we assume that the probability density function (PDF) of the forecasted wind speed, x equals:

$$p(x|f_c, f_{\ell,1}, \ldots, f_{\ell,10}; b_0, b_1, c_0, c_1) = \omega g(x|f_c, b_0, b_1, c_0, c_1) + \frac{1-\omega}{10} \times \sum_{j=1}^{10} g(x|f_{\ell,j}, b_0, b_1, c_0, c_1), \quad (3.1)$$

where $\omega \in [0, 1]$, and g is the conditional PDF corresponding to the ensemble members. As we are working with wind speed data, $g(x|f, b_0, b_1, c_0, c_1)$ is a gamma PDF with mean $b_0 + b_1 f$ and standard deviation $c_0 + c_1 f$. Here, both mean and standard deviation parameters are chosen to be the same for all ensemble members, which reduces the number of parameters and simplifies calculations. The mean parameters b_0 and b_1 are estimated by linear regression, while the weight parameter, ω and the standard deviation parameters c_0 and c_1 are estimated by the maximum likelihood method using training data consisting of ensemble members and verifying observations from the preceding n days (the training period). In order to handle the problem that wind speed values under 0.1 m/s are considered to be zero, the maximum likelihood (ML) method for gamma distributions suggested by WILKS (1990) is applied, while the maximum of the likelihood function is found with the help of EM algorithm (MCLACHLAN and KRISHNAN, 1997). For more details, see SLOUGHTER et al. (2010) and FRALEY et al. (2010). Once the estimated parameters for a given day are avail-

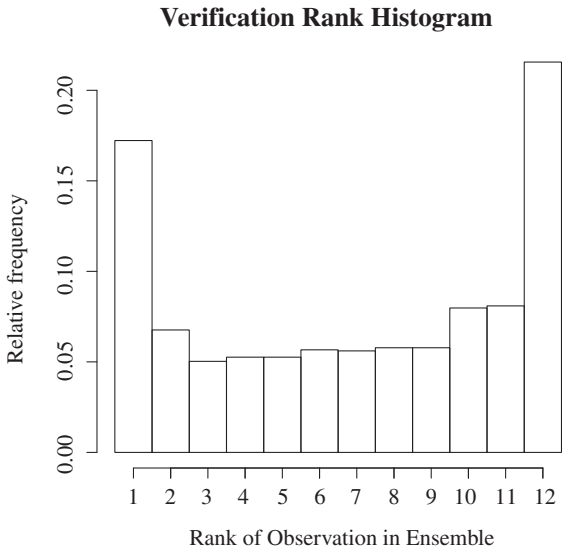

Figure 1: Verification rank histogram of the 11-member ALADIN-HUNEPS ensemble. Period: October 1, 2010 – March 25, 2011.

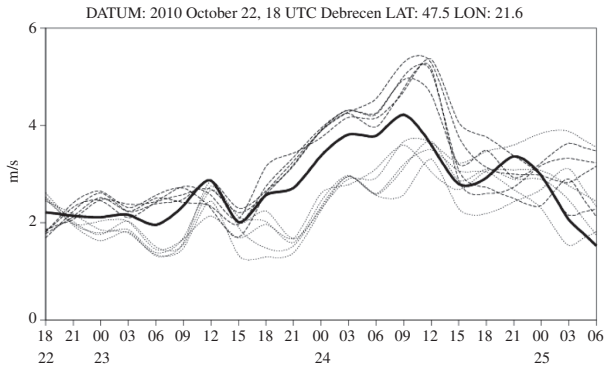

Figure 2: Plume diagram of ensemble forecast of 10 meter wind speed for Debrecen initialized at 18 UTC, 22.10.2010 (solid line: control; dotted line: odd numbered members; dashed line: even numbered members).

able, one can use either the mean or the median of the predictive PDF (3.1) as a point forecast.

Having a more careful look over the ensemble, one can notice that there are some differences in the generation of the ten exchangeable ensemble members. To obtain them, only five perturbations are calculated and then they are added to (odd numbered members) and subtracted from (even numbered members) the unperturbed initial conditions (HORÁNYI et al., 2011). Fig. 2 shows the plume diagram of the ensemble forecast of 10 meter wind speed for Debrecen initialized at 18 UTC, 22.10.2010 (solid line: control; dotted line: odd numbered members; dashed line: even numbered members). This diagram clearly illustrates that the behaviour of ensemble member groups $\{f_{\ell,1}, f_{\ell,3}, f_{\ell,5}, f_{\ell,7}, f_{\ell,9}\}$ and $\{f_{\ell,2}, f_{\ell,4}, f_{\ell,6}, f_{\ell,8}, f_{\ell,10}\}$ significantly differ from each other. Therefore, in this way one can also consider a model with three exchangeable groups: control, odd numbered exchangeable members and even numbered exchangeable members. This idea leads to the following PDF of the forecasted wind speed x:

$$\begin{aligned} q(x|f_c, f_{\ell,1}, \ldots, f_{\ell,10}; b_0, b_1, c_0, c_1) \\ = \omega_c g(x|f_c, b_0, b_1, c_0, c_1) \\ + \sum_{j=1}^{5} (\omega_o g(x|f_{\ell,2j-1}, b_0, b_1, c_0, c_1) \\ + \omega_e g(x|f_{\ell,2j}, b_0, b_1, c_0, c_1)), \end{aligned} \tag{3.2}$$

where for weights $\omega_c, \omega_o, \omega_e \in [0,1]$ we have $\omega_c + 5\omega_o + 5\omega_e = 1$, while the definition of the PDF, g and the parameters b_0, b_1, c_0, c_1 remains the same as for the model (3.1). Obviously, both the weights and the parameters can be estimated in the same way as before.

As an illustration, we consider the data and forecasts for Debrecen for two different dates, 30.12.2010 and 17.03.2011, for models (3.1) and (3.2). Figs. 3a and 3b show the PDFs of the two groups in model (3.1), the overall PDFs, the median forecasts, the verifying observations, the first and last deciles and the ensemble members. The same functions and quantities can be seen in Figs. 3c and 3d, where besides the overall PDF, we have the three component PDFs and three groups of ensemble members. On 30.12.2010, the spread of the ensemble members is reasonable and the ensemble range contains the validating observation (3.2 m/s). In this case, the ensemble median (3.77 m/s) overestimates, while BMA median forecasts corresponding to the two- and three-group models (3.29 m/s and 3.22 m/s, respectively) are quite close to the true wind speed. A different situation is illustrated in Figs. 3b and 3d where the spread of the ensemble is even larger and all ensemble members underestimate the validating observation (6.1 m/s). Obviously, the same holds for the ensemble median (3.3 m/s) and the BMA median forecasts corresponding to models (3.1) and (3.2), as they also give inaccurate results (3.34 m/s and 3.08 m/s, respectively).

In order to check the overall performance of the probabilistic forecasts (based on (3.1) and (3.2)) in terms of a probability distribution function, the mean continuous ranked probability scores (CRPS; WILKS, 2006; GNEITING and RAFTERY, 2007) and average widths of 66.7% and 90% central prediction intervals are computed and compared for the corrected and raw ensemble. In the latter case, the ensemble of forecasts corresponding to a given location and time is considered as a statistical sample and the sample quantiles are calculated according to HYNDMAN and FAN (1996, Definition 7). Additionally, the ensemble mean and median are used to consider point forecasts, which are evaluated with the use of mean absolute errors (MAE) and root mean square errors (RMSE). We remark that for MAE and RMSE, the optimal point forecasts are the median and the mean, respectively (GNEITING, 2011; PINSON and HAGEDORN, 2012). Further, given a cumulative distribution function (CDF) $F(y)$ and a real number x, the CRPS is defined as

$$\mathrm{crps}(F,x) := \int_{-\infty}^{\infty} (F(y) - \mathbb{1}_{\{y \geq x\}})^2 \mathrm{d}y.$$

The mean CRPS of a probability forecast is the average of the CRPS values of the predictive CDFs and corresponding validating observations taken over all locations and time points considered. For the raw ensemble, the empirical CDF of the ensemble replaces the predictive CDF. The coverage of a $(1-\alpha)100\%$, $\alpha \in (0,1)$ central prediction interval is the proportion of validating observations located between the lower and upper $\alpha/2$ quantiles of the predictive distribution. For a calibrated predictive PDF this value should be around $(1-\alpha)100\%$.

4 Results

The data analysis provided below was performed using the ensembleBMA package in R (FRALEY et al., 2009, 2011). As a first step, the length of the appropriate training period was determined, then the performances of the BMA post-processed ensemble forecasts corresponding to models (3.1) and (3.2) were analyzed.

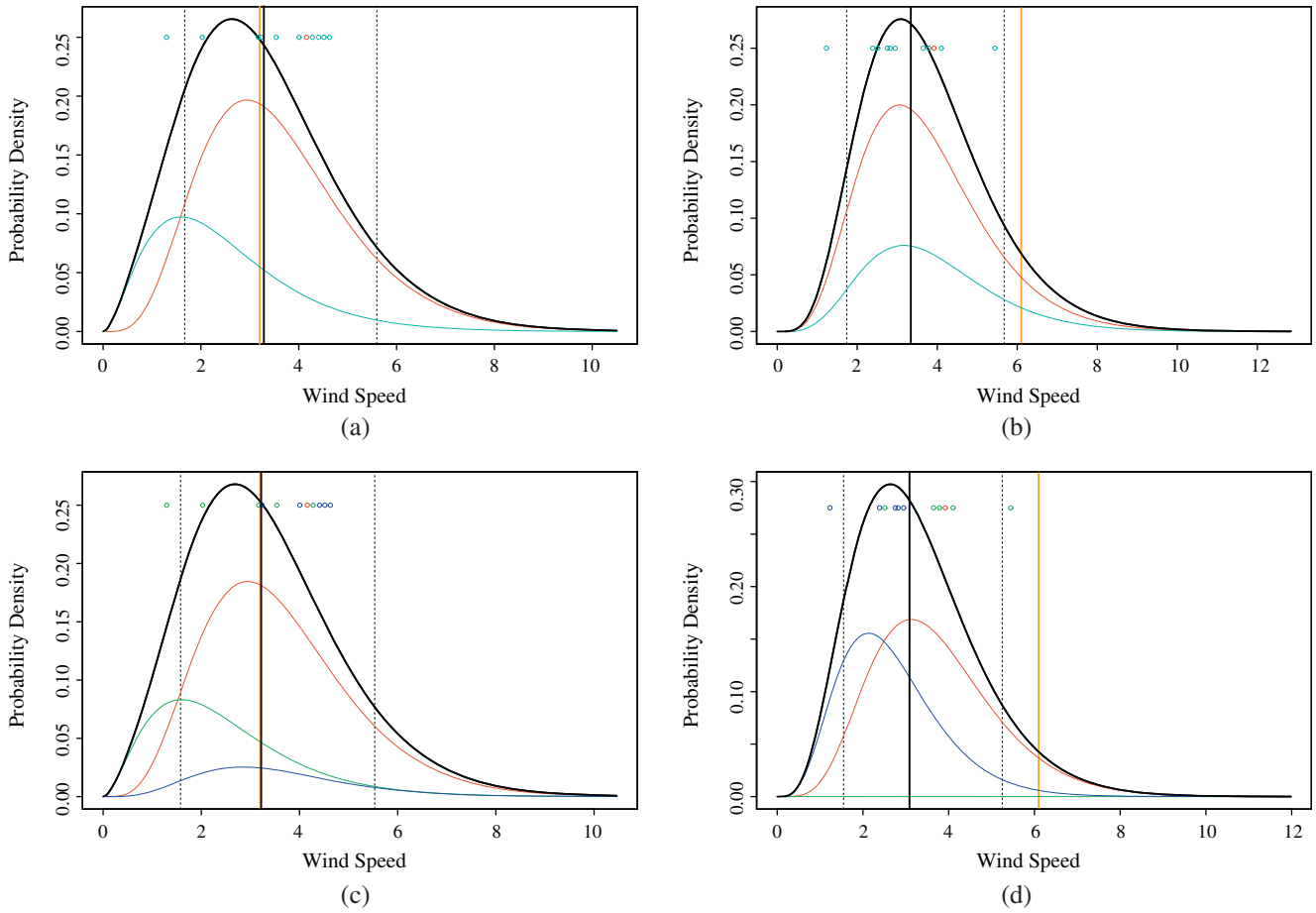

Figure 3: Ensemble BMA PDFs (overall: thick black line; control: red line; sum of exchangeable members in (a) and (b): light blue line; in (c) and (d): green (odd members) and blue (even members) lines), ensemble members (circles with the same colours as the corresponding PDFs), ensemble BMA median forecasts (vertical black line), verifying observations (vertical orange line) and the first and last deciles (vertical dashed lines) for wind speed in Debrecen for models (3.1): (a) 30.12.2010, (b) 17.03.2011; and (2): (c) 30.12.2010, (d) 17.03.2011.

4.1 Training period

We proceed in the same way as RAFTERY et al. (2005) and determine the length of the training period to be used by comparing the MAE values of BMA median forecasts, the RMSE values of BMA mean forecasts, the CRPS values of BMA predictive distributions and the coverage and average widths of 90% and 66.7% BMA central prediction intervals for training periods of length 10,11,...,60 calendar days. In order to ensure the comparability of the results, we consider verification results from 02.12.2010 to 25.03.2011 (114 days).

Consider first the two-group model (3.1). In Fig. 4, the average widths and coverage of 66.7% and 90% BMA central prediction intervals are plotted against the length of the training period. As the average widths of the central prediction intervals show an increasing trend, shorter training periods yield sharper forecasts. On the other hand, the coverage of 66.7% and 90% central prediction intervals also increase, but not monotonously. For short training periods, the coverage of the 66.7% central prediction interval oscillates around the correct 66.7%, but for training periods greater than around 20 days, it stays above this level. The coverage of the 90% central prediction interval stabilizes above the correct 90% for training periods longer than approximately 25 days. Hence, to have calibrated forecasts, one should not choose a training period less than 25 days, while training period lengths much higher than 25 days should be also avoided as the increasing coverage of the central prediction intervals, away from the nominal value, results in overdispersion (so this diagnostic would rather suggest to use values around 25).

Fig. 5 shows CRPS values of the BMA predictive distribution, MAE values of the BMA median forecasts and RMSE values of the BMA mean forecasts as function of the training period length. The CRPS, MAE and RMSE take their minima at around day 28–30. The corresponding values are 0.7388, 1.0472 and 1.3675, respectively. This means that for model (3.1), a 28 day training period seems to be reasonable (choosing the smallest value of the interval mentioned above), while an extension of the number of training days beyond 30 leads to inferior results.

Similar conclusions can be drawn from Figs. 6 and 7 for the three-group model (3.2). In this case, the 66.7%

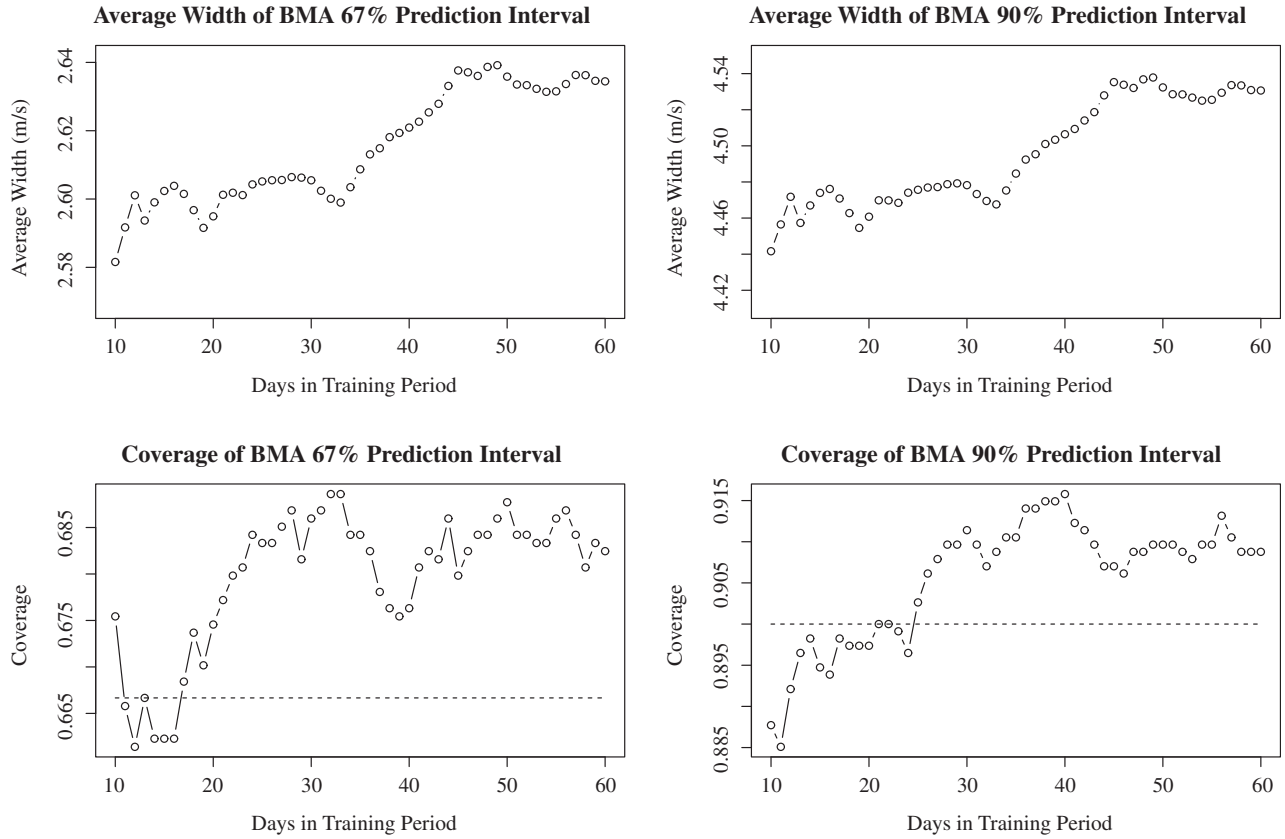

Figure 4: Average widths and coverages of 66.7% and 90% BMA central prediction intervals corresponding to the two-group model (3.1) for various training period lengths.

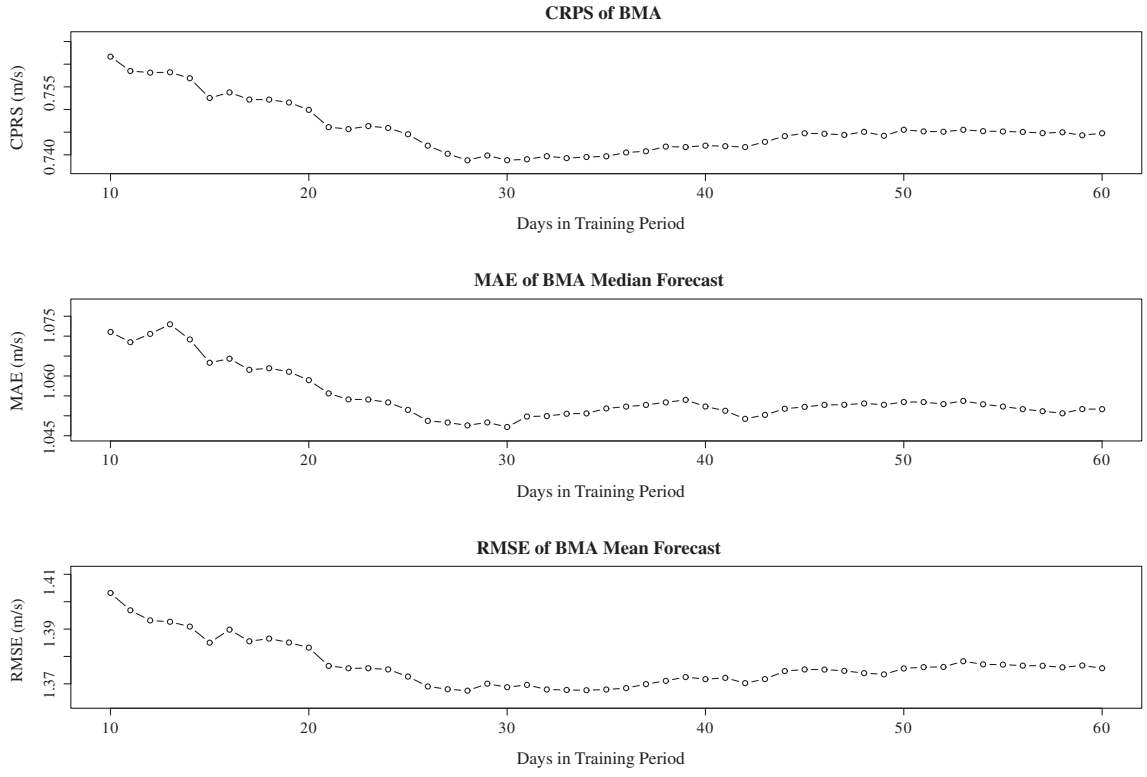

Figure 5: CRPS of the BMA predictive distribution, MAE values of the BMA median and RMSE values of the BMA mean forecasts corresponding to the two-group model (3.1) for various training period lengths.

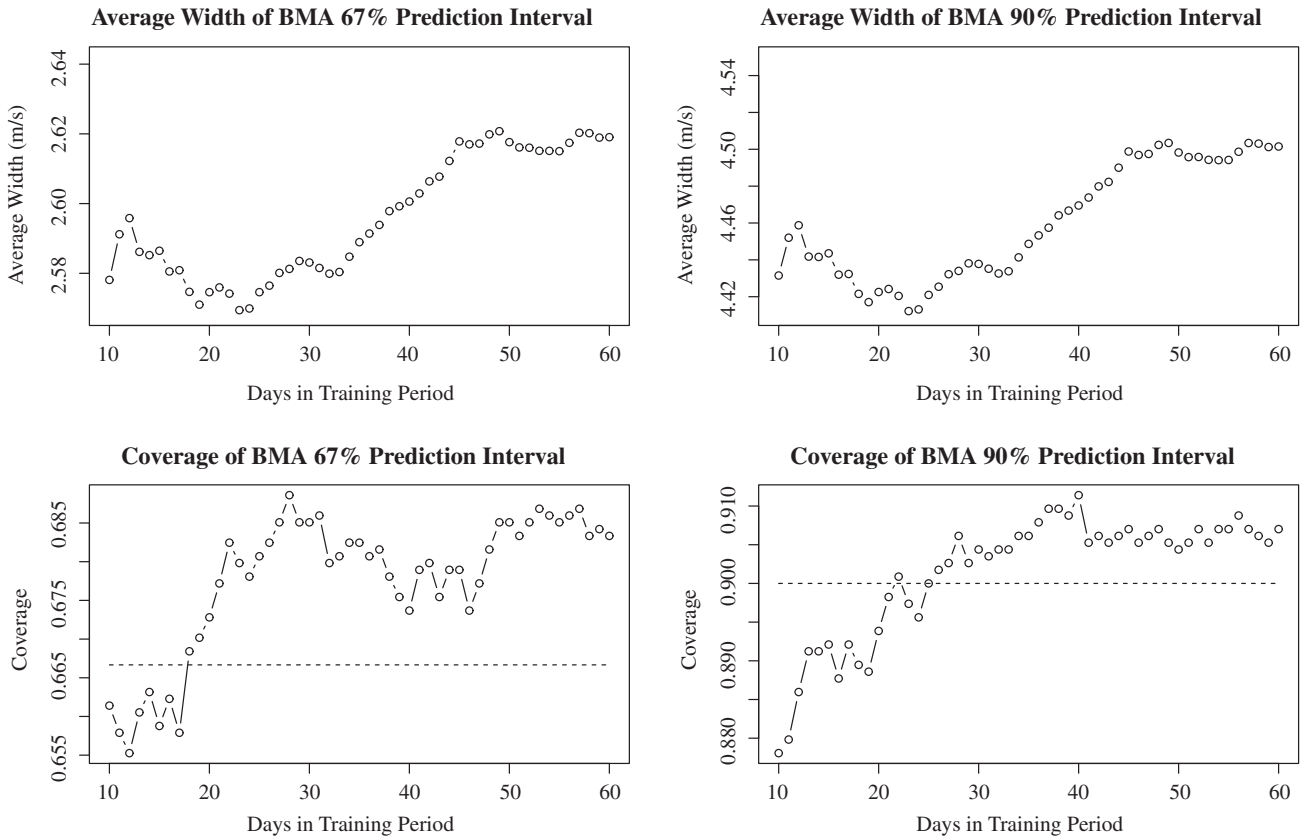

Figure 6: Average widths and coverages of 66.7% and 90% BMA central prediction intervals corresponding to the three-group model (3.2) for various training period lengths.

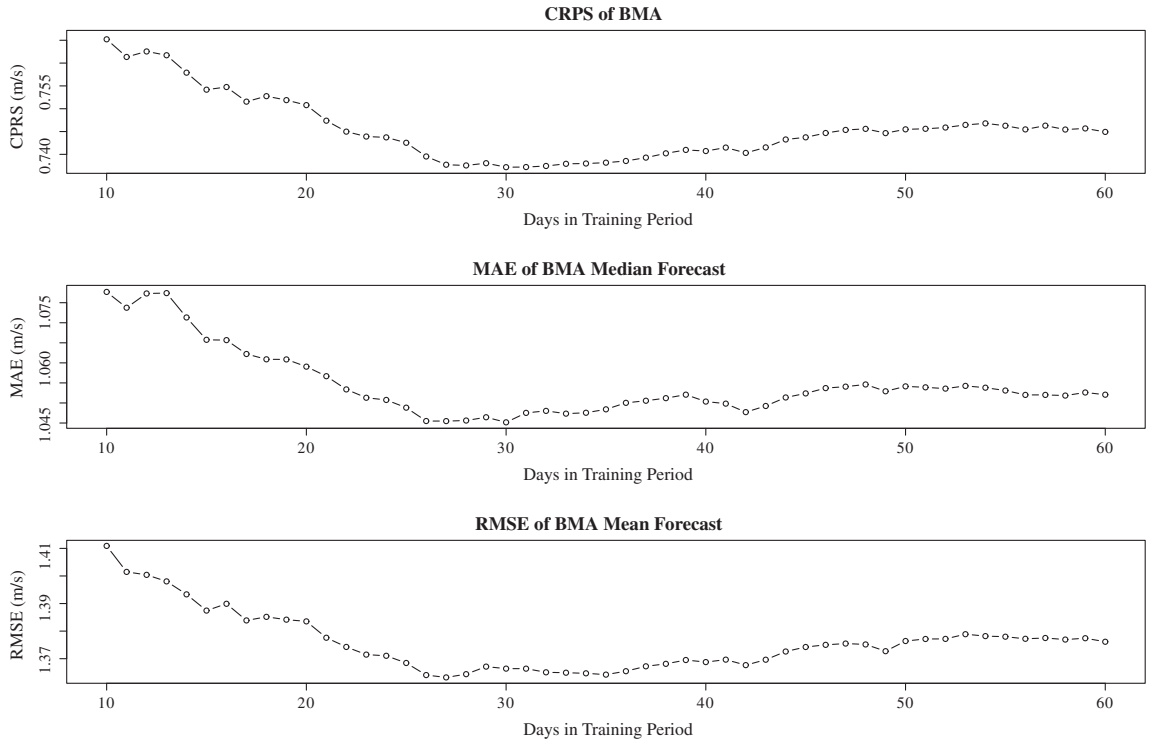

Figure 7: CRPS of the BMA predictive distribution, MAE values of the BMA median and RMSE values of the BMA mean forecasts corresponding to the three-group model (3.2) for various training period lengths.

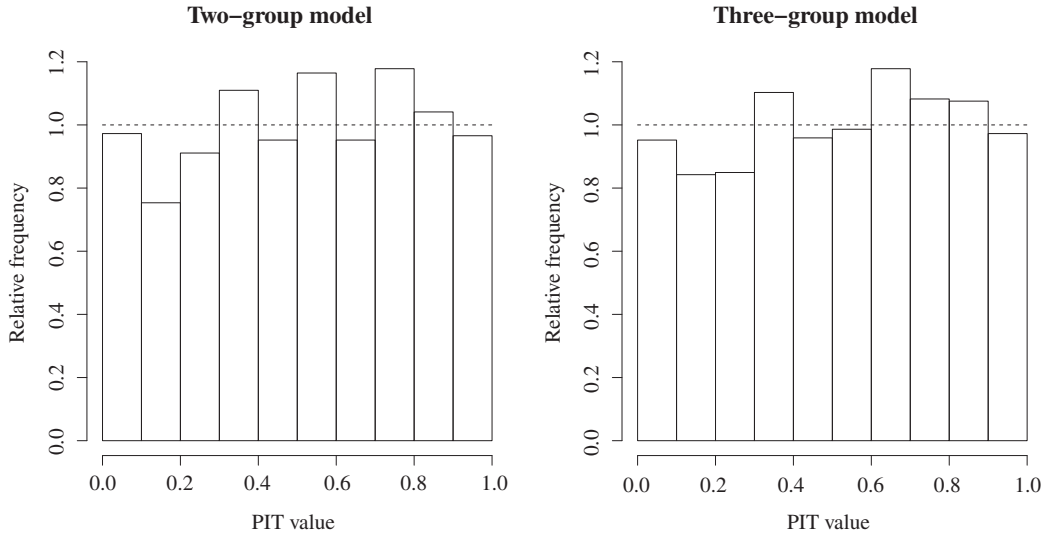

Figure 8: PIT histograms for BMA post-processed forecasts using the two-group (3.1) and three-group (3.2) models.

Table 1: Coverage and average widths of central prediction intervals.

	Coverage (%)		Average Width (m/s)	
Interval	66.7% interval	90.0% interval	66.7% interval	90.0% interval
Raw ensemble	38.70	55.14	1.4388	2.2001
BMA model (1)	68.08	90.34	2.6359	4.5297
BMA model (2)	68.36	90.21	2.6153	4.4931

and 90% central prediction intervals are slightly narrower than the corresponding intervals of model (3.1) as their coverage stabilizes above the correct 66.7% and 90% for training periods longer than about 20 and 25 days, respectively. The CRPS, MAE and RMSE plotted in Fig. 7 reach their minima of 0.7372, 1.0452 and 1.3632, respectively, 25–35 days into the training period. After 35 days, there is a slight, but rather monotonous increase in every measure. Moreover, for the 66.7% and 90% central prediction intervals, the shorter period is better in the 25–35 day range in terms of sharpness and therefore, the lower part of this interval seems appropriate for the three-group model. When comparing the diagnostic results for the two-group and three-group models, the choice between 25 and 30 days seems appropriate for the length of the training period. Finally, we have chosen the training period of length 28 days for both BMA models. It is mentioned here that the decision on the training period is subject to sampling variability and this is taken into account as much as possible in the final choice.

4.2 Predictions using BMA post-processing

According to the results of the previous subsection, to test the performance of BMA post-processing on the 11 member ALADIN-HUNEPS ensemble, we use a training period of 28 calendar days. In this way ensemble members, validating observations and BMA models are available for 146 calendar days (on 20.11.2010 all ensemble members are missing).

First we check the calibration of BMA post-processed forecasts with the help of probability integral transform (PIT) histograms. The PIT is the value of the BMA predictive cumulative distribution evaluated at the verifying observations (RAFTERY et al., 2005). The closer the histogram is to the uniform distribution, the better the calibration is. In Fig. 8, the PIT histograms corresponding to two- and three-group BMA models (3.1) and (3.2) are displayed. A comparision of the verification rank histogram of the raw ensemble (see Fig. 1) shows that post-processing improves the statistical calibration of the forecasts substantially. However, these PIT histograms are still not perfect as e.g., a Kolmogorov-Smirnov test rejects uniformity both for the two- and for the three-group model. As both corresponding p-values are 0.02, there is no difference between PITs of the two-group and of the three-group model.

Table 1 gives the coverage and average widths of 66.7% and 90.0% central prediction intervals calculated using models (3.1) and (3.2), and the corresponding measures calculated from the raw ensemble. As before, in the latter case the ensemble of forecasts corresponding to a given location and time is considered as a statistical sample from which the central prediction intervals are derived. The BMA central prediction intervals calculated

Table 2: Mean CRPS of probabilistic, MAE and RMSE of deterministic forecasts.

	Mean CRPS (m/s)	MAE (m/s)		RMSE (m/s)	
		median	mean	median	mean
Raw ensemble	0.8599	1.1215	1.1090	1.4634	1.4440
BMA model (3.1)	0.7577	1.0678	1.0763	1.4213	1.4067
BMA model (3.2)	0.7556	1.0643	1.0749	1.4153	1.4018

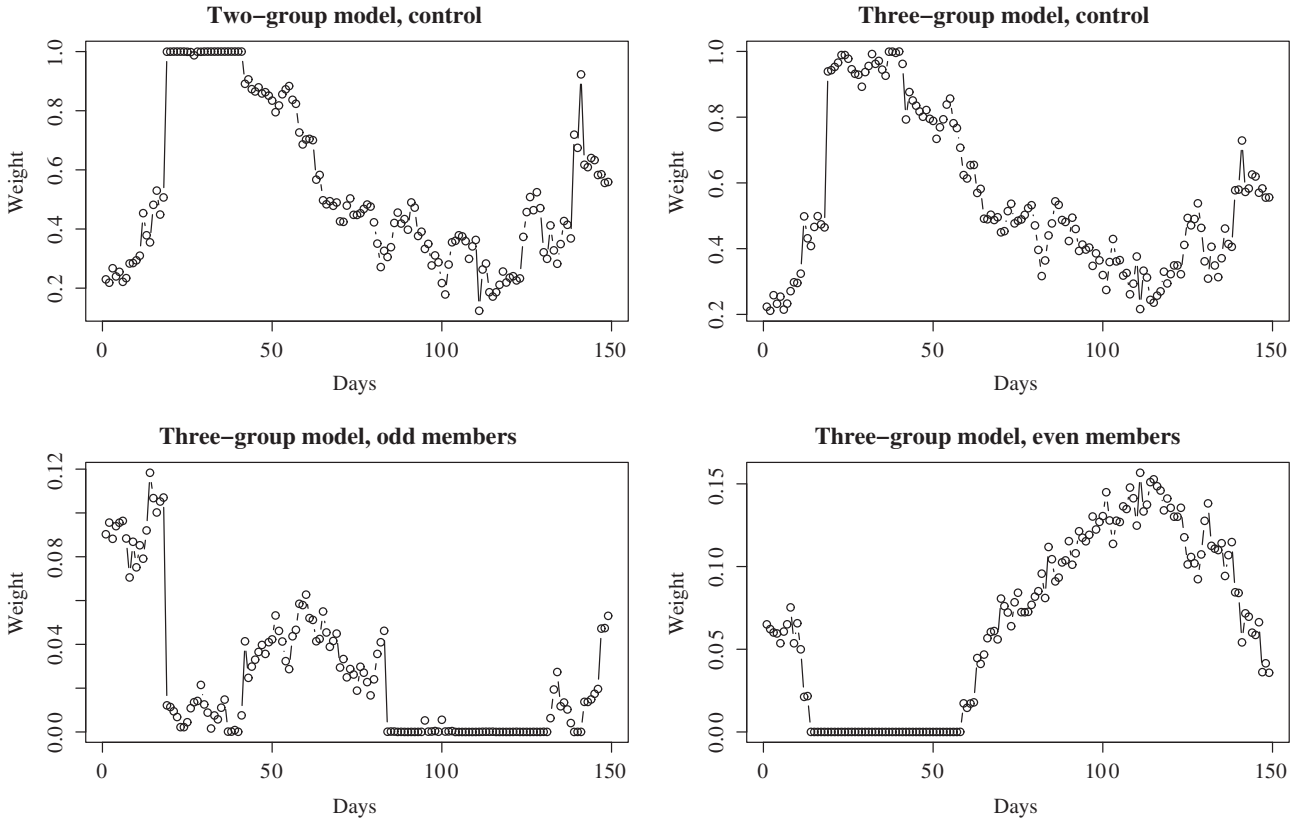

Figure 9: BMA weights of the two-group (3.1) and three-group (3.2) models.

from both models are approximately twice as wide as the corresponding intervals of the raw ensemble. This comes from the small dispersion of the raw ensemble as seen in the verification rank histogram of Fig. 1. Concerning calibration, one can observe that the coverage of both BMA central prediction intervals are rather close to the correct coverage, while the coverage of the central prediction intervals calculated from the raw ensemble are quite poor. In addition to the almost uniform PIT histogram, this shows that BMA post-processing greatly improves calibration. Further, the BMA model (3.2) yields slightly sharper predictions, but there is no great difference between the coverage of the two BMA models.

In Table 2, scores for the different probabilistic forecasts are given. Verification measures of probabilistic forecasts and point forecasts calculated using BMA models (3.1) and (3.2) are compared to the corresponding measures calculated for the raw ensemble. By examining these results, one can clearly observe the advantage of BMA post-processing which resulted in a significant decrease in all verification scores. Further, the BMA median forecasts yield slightly lower MAE values than the BMA mean forecasts for both models, while in the case of RMSE values the situation is just the opposite, which is a perfect illustration of the theoretical results of GNEITING (2011) about the optimality of these verification scores. Finally, model (3.2) distinguishing three exchangeable groups of ensemble forecasts slightly outperforms model (3.1).

Fig. 9 shows the BMA weights corresponding to models (3.1) and (3.2). Examining the behaviour of the weight, ω of the control member of the ensemble in the two-group model (3.1), one can observe that in 84.56% of the cases, there is a real mixture of gamma distributions (none of the groups has a weight which is almost 1). The values of ω which are close to 1 correspond to a time interval 17.11.2010 – 09.12.2010 when the control member of the ensemble gives much better

Table 3: MAE and RMSE of the control and exchangeable ensemble forecasts for the period 17.11.2010 – 09.12.2010.

	Control	Exchangeable members								
	f_c	$f_{\ell,1}$	$f_{\ell,2}$	$f_{\ell,3}$	$f_{\ell,4}$	$f_{\ell,5}$	$f_{\ell,6}$	$f_{\ell,7}$	$f_{\ell,8}$	$f_{\ell,9}$	$f_{\ell,10}$
MAE (m/s)	1.32	1.60	1.46	1.52	1.68	1.51	1.49	1.56	1.42	1.41	1.65
RMSE (m/s)	1.69	2.16	1.86	1.96	2.26	1.92	1.95	2.05	1.89	1.81	2.23

forecasts than the ten exchangeable ensemble members. This can clearly be seen from Table 3 where the MAE and RMSE values of the particular ensemble members are given for the above mentioned period. In all of these 23 subsequent days, $\omega > 0.995$ except for 25.11.2010 when $\omega = 0.9873$. The situation is quite different in the case of the three-group model (3.2) where the weight, ω_c of the control is close to 1 (greater than 0.98) only on 7 days. Thus in the remaining cases (95.30%), a real mixture of gamma distributions are present. Further, observe that there are 55 days (36.91%) when all BMA weights are positive, the even numbered exchangeable members have nearly zero weights (less than 0.001) in 45 cases (30.20%) at the beginning of the considered time period, while the odd numbered exchangeable members are almost zero in 53 cases (35.57%), and which occur mainly at the end of this period.

5 Conclusions

In the present study, the BMA ensemble post-processing method is applied to the 11 member ALADIN-HUNEPS ensemble of the HMS to obtain 42 hour calibrated predictions for the 10 meter wind speed. Two different BMA models are investigated. One assumes two groups of exchangeable members (control and forecasts from perturbed initial conditions), while the other considers three (control and forecasts from perturbed initial conditions with positive and negative perturbations). For both models, a 28 days training period is suggested. The comparison of the raw ensemble and of the probabilistic forecasts shows that the mean CRPS values of BMA post-processed forecasts are considerably lower than the mean CRPS of the raw ensemble. Furthermore, the MAE and RMSE values of BMA point forecasts (median and mean) are also lower than the MAEs and RMSEs of the ensemble median and of the ensemble mean. The calibration of BMA forecasts is nearly perfect as the coverage of the 66.7% and 90.0% central prediction intervals are very close to the nominal levels. The three-group BMA model slightly outperforms the two-group one and in almost all cases yields a real mixture of gamma distributions.

We therefore conclude that as the BMA post-processing of the ALADIN-HUNEPS wind speed ensemble forecasts significantly improves the calibration and accuracy of point forecasts, its operational application is worth considering.

Acknowledgments

Research has been supported by the Hungarian Scientific Research Fund under Grants No. OTKA T079128 and OTKA NK101680 and by the TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project. The project has been supported by the European Union and co-financed by the European Social Fund. The authors are indebted to Tilmann Gneiting for his useful suggestions and remarks and to Máté Mile and Mihály Szűcs from the HMS for providing the data. Last, but not least we are very grateful to Mike Rennie for revising the English of the paper.

References

BAO, L., T. GNEITING, A.E. RAFTERY, E.P. GRIMIT, P. GUTTORP, 2010: Bias correction and Bayesian model averaging for ensemble forecasts of surface wind direction. – Mon. Wea. Rev. **138**, 1811–1821.

BUIZZA, R., J. TRIBBIA, F. MOLTENI, T. PALMER, 1993: Computation of optimal unstable structures for a numerical weather prediction system. – Tellus A **45**, 388–407.

BUIZZA, R., P.L. HOUTEKAMER, Z. TOTH, G. PELLERIN, M. WEI, Y. ZHU, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. – Mon. Wea. Rev. **133**, 1076–1097.

DESCAMPS, L., C. LABADIER, A. JOLY, J. NICOLAU, 2009: Ensemble Prediction at Météo France (poster introduction by Olivier Riviere) –31st EWGLAM and 16th SRNWP meetings, 28th September – 1st October, 2009. Available at: srnwp.met.hu/Annual_Meetings/2009/download/sept29/morning/posterpearp.pdf.

FRALEY, C., A.E. RAFTERY, T. GNEITING, J.M. SLOUGHTER, 2009: EnsembleBMA: An R package for probabilistic forecasting using ensembles and Bayesian model averaging. – Technical Report 516R, Department of Statistics, University of Washington. Available at: www.stat.washington.edu/research/reports/2008/tr516.pdf.

FRALEY, C., A.E. RAFTERY, T. GNEITING, 2010: Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. – Mon. Wea. Rev. **138**, 190–202.

FRALEY, C., A.E. RAFTERY, T. GNEITING, J.M. SLOUGHTER, V.J. BERROCAL, 2011: Probabilistic weather forecasting in R. – The R Journal **3**, 55–63.

GNEITING, T., 2011: Making and evaluating point forecasts. – J. Amer. Statist. Assoc. **106**, 746–762.

GNEITING, T., A.E. RAFTERY, 2005: Weather forecasting with ensemble methods. – Science **310**, 248–249.

GNEITING, T., A.E. RAFTERY, 2007: Strictly proper scoring rules, prediction and estimation. – J. Amer. Statist. Assoc. **102**, 359–378.

GRIMIT, E.P., C.F. MASS, 2002: Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest. – Wea. Forecasting **17**, 192–205.

HÁGEL, E., 2010: The quasi-operational LAMEPS system of the Hungarian Meteorological Service. – Időjárás **114**, 121–133.

HORÁNYI, A., S. KERTÉSZ, L. KULLMANN, G. RADNÓTI, 2006: The ARPEGE/ALADIN mesoscale numerical modeling system and its application at the Hungarian Meteorological Service. – Időjárás **110**, 203–227.

HORÁNYI, A., M. MILE, M. SZŰCS, 2011: Latest developments around the ALADIN operational short-range ensemble prediction system in Hungary. – Tellus A **63**, 642–651.

HYNDMAN, R.J., Y. FAN, 1996: Sample quantiles in statistical packages. – Amer. Statist. **50**, 361–365.

LEITH, C.E., 1974: Theoretical skill of Monte-Carlo forecasts. – Mon. Wea. Rev. **102**, 409–418.

MCLACHLAN, G.J, T. KRISHNAN, 1997: The EM Algorithm and Extensions. – Wiley, New York.

PINSON, P., R. HAGEDORN, 2012: Verification of the ECMWF ensemble forecasts of wind speed against analyses and observations. – Meteorol. Appl. **19**, 484–500.

RAFTERY, A.E., T. GNEITING, F. BALABDAOUI, M. POLAKOWSKI, 2005: Using Bayesian model averaging to calibrate forecast ensembles. – Mon. Wea. Rev. **133**, 1155–1174.

SLOUGHTER, J.M., A.M. RAFTERY, T. GNEITING, C. FRALEY, 2007: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. – Mon. Wea. Rev. **135**, 3209–3220.

SLOUGHTER, J.M., T. GNEITING, A.E. RAFTERY, 2010: Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. – J. Amer. Statist. Assoc. **105**, 25–37.

TOTH, Z., E. KALNAY, 1997: Ensemble forecasting at NCEP and the breeding method. – Mon. Wea. Rev. **125**, 3297–3319.

WILKS, D.S., 1990: Maximum likelihood estimation for gamma distribution using data containing zeros. – J. Climate **3**, 1495–1501.

WILKS, D.S., 2006: Statistical Methods in the Atmospheric Sciences. 2nd ed. – Academic Press, New York.