Calibration of ALADIN/LAEF precipitation ensembles

Ján Mašek and Tomáš Vlasák, Czech HydroMeteorological Institute

1. Introduction

Aim of this study was to evaluate quality of ALADIN/LAEF precipitation ensembles and to propose and evaluate suitable calibration technique. Work was targetted to hydrological application of ensembles, this report however restricts only to its atmospheric part, which is directly linked to NWP. All work was done locally at CHMI. Technical preparation and feasibility study was done in summer/autumn 2010, work itself was finalized in autumn 2011.

2. Used data

Precipitation was evaluated in hydrological zones shown on figure 1. Study was restricted to 6 hour precipitation amounts, evaluation period was the year 2010. Zones f, s, t, u, J, N, O, P were excluded from verification due to insufficient observations (too many missing data or too big part of zone lying outside of Czech territory), which means that 29 zones out of 37 were finally used. Area of the zones varies roughly between 1300 and 4000km^2.

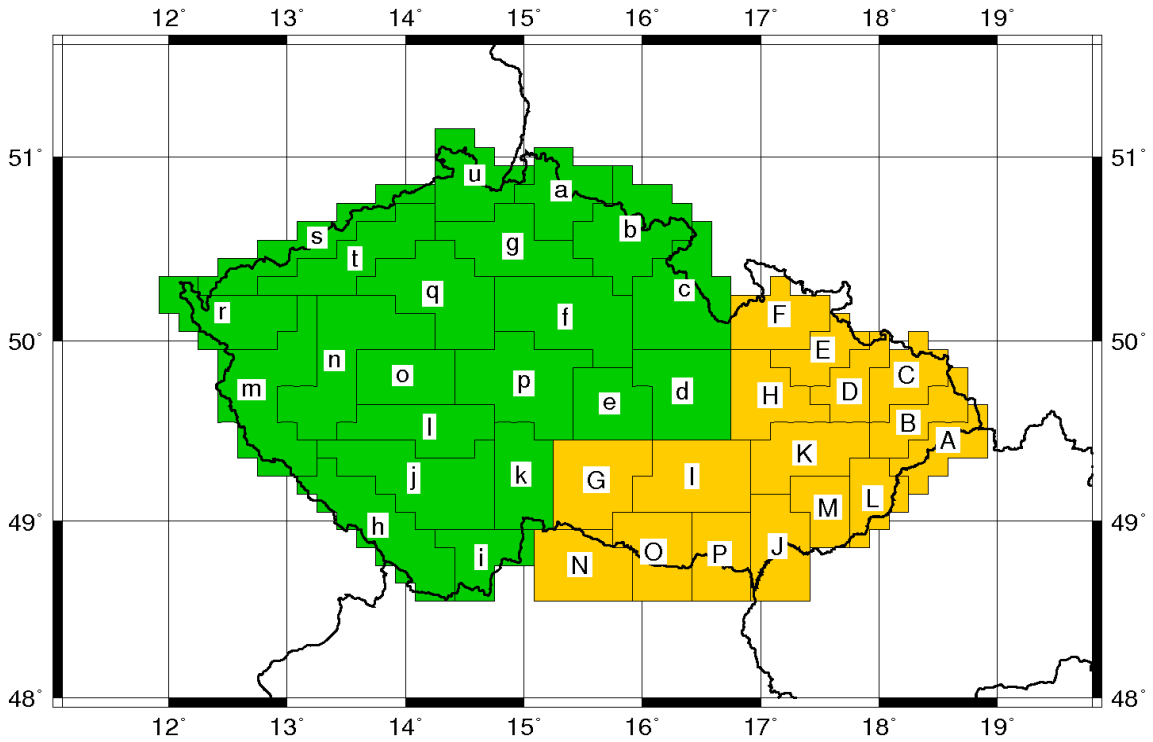

Fig. 1: *Hydrological zones used for evaluation of mean area precipitation.*

Precipitation measurements from automatic raingauges were tried originally. However, spatial density and reliability of measured data was not sufficient, that is why manual raingauges were used finally. Their advantage is higher spatial density and quality control applied on measured data. But since they provide only 24 hour cumulated precipitation, splitting to 6 hour intervals had to be done according to radar precipitation estimates. In other words, radar precipitation cleaned from RLAN interference and calibrated by manual rainguages on day by day bases was used.

Raingauge measurements were first interpolated onto regular grid, using universal kriging with linear variogram (for gridded model and radar data this step was not needed). Observed and forecasted mean area precipitation in each zone was computed from gridded data, by identifying boxes (or proportion of boxes) lying inside boundary polygon.

3. Methodology

ALADIN/LAEF ensemble consists of 17 members (control forecast plus 16 perturbed forecasts); its detailed description can be found in [5]. Study used forecasts starting at 00 UTC and going ahead for 54 hours. Comparison of forecasted and observed yearly precipitation showed that ALADIN/LAEF has overall tendency to overestimate precipitation (it strongly depends on forecast lead time, on average it is by 36%). At the same time U-shaped rank histograms indicated that the ensemble might have insufficient spread (left panel on figure 2). Basics of rank histograms and more possible mechanisms leading to their U-shape are dicussed in [1].

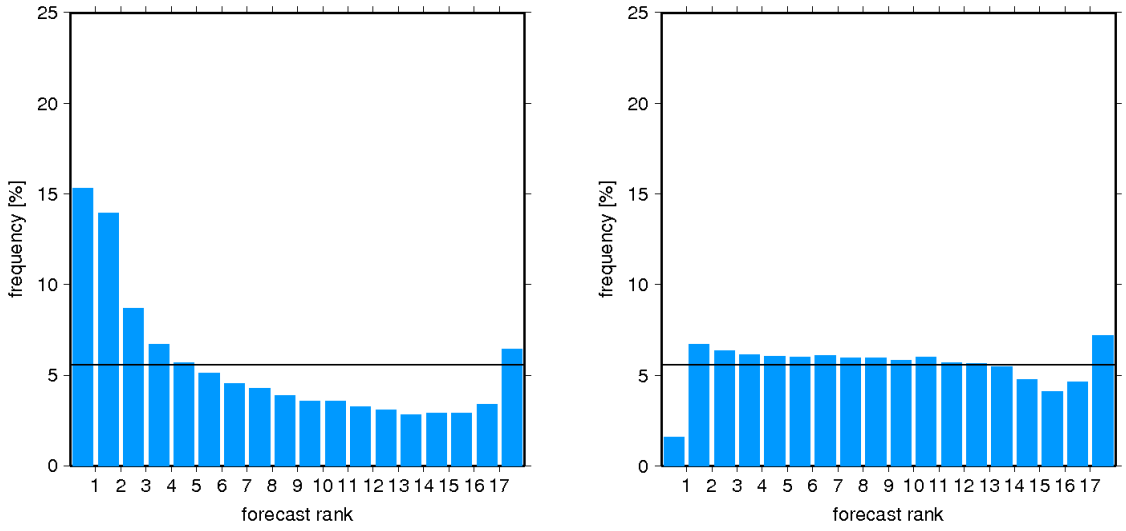

Fig. 2: *Rank histograms for 30-36 hour precipitation forecast (year 2010, all zones): left – raw ensemble; right – calibrated ensemble. On horizontal axis is forecast rank, on vertical axis frequency of observation falling between given ranks. Horizontal line denotes perfect distribution.*

Main idea behind ensemble calibration was to reduce forecasted mean yearly precipitation to observed yearly precipitaton (climatology) and at the same time to achieve as flat rank histogram as possible (flat rank histogram implies reliable forecasted probabilities). Calibration function $y = f(x)$ transforming raw forecasted precipitation x to calibrated value y must be smoothly increasing and preserving zero. Since overestimation of precipitation together with insufficient ensemble spread cannot be cured by linear rescaling, nonlinear calibration function was proposed:

$$f(x) = \begin{cases} \alpha x \left[\frac{\alpha}{\beta} + \left(1 - \frac{\alpha}{\beta}\right) \cdot \left(\frac{x}{\bar{x}}\right)^{\frac{\beta}{\alpha}} \right]; & x \le \bar{x} \\ (\alpha - \beta)\bar{x} + \beta x; & x > \bar{x} \end{cases}$$

It transforms ensemble mean $\bar{x}$ to $\alpha\bar{x}$ and for $x > \bar{x}$ has constant slope β, where $\alpha, \beta > 0$ are calibration parameters (see figure 3 for details). Number of calibration parameters was kept at minimum (i.e. two per hydrological zone), in order to reduce risk of overfitting due to only one year long data set. In all cases calibration procedure delivered $\beta > \alpha$, leading to convex function as on figure 3. Measure of histogram flatness was the sum of squared departures from ideal (i.e. uniform) frequencies.

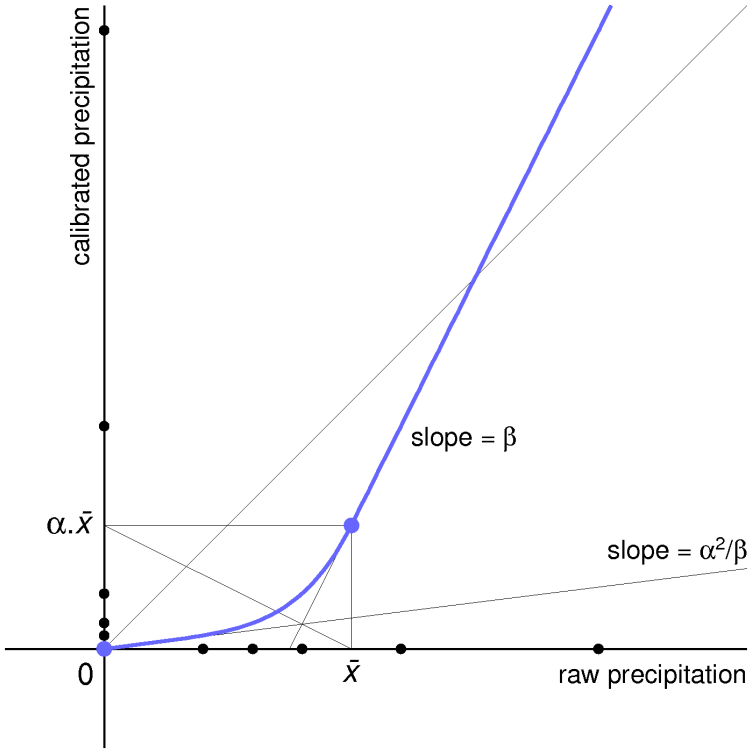

Fig. 3: *Shape of calibration function with parameters* $\alpha = 0.5$ *and* $\beta = 2$. *Dots on horizontal axis represent raw ensemble with mean value* $\bar{x}$, *dots on vertical axis are corresponding calibrated values.*

Quality of ALADIN/LAEF precipitation ensembles was evaluated using common verification techniques developed for probabilistic forecasting. Apart from rank histograms those were reliability diagrams and Brier skill score complemented by forecast relative value (in [4] there is explanation why Brier skill score alone does not have unique link to forecast value based on simple cost-loss economical model). Some deterministic scores like frequency bias or Peirce skill score were generalized to probabilistic case as well. Most important definitions are summarized below.

Brier skill score (BSS) is derived from Brier score (BS) defined as:

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2 = \underbrace{\frac{1}{N}\sum_{k=1}^{K} n_k(p_k - \bar{o}_k)^2}_{\text{(un)reliability}} - \underbrace{\frac{1}{N}\sum_{k=1}^{K} n_k(\bar{o}_k - \bar{o})^2}_{\text{resolution}} + \underbrace{\bar{o}(1-\bar{o})}_{\text{uncertainity}}$$

Here index i runs through verification period containing N cases, p_i is forecasted probability of the event (e.g. precipitation exceeding certain treshold), o_i is 1 when the event occurred and 0 when it did not occur. Brier score is negatively oriented and ranges from 0 (perfect forecast only) to 1 (perfectly wrong forecast only).

Ensemble with M members can forecast certain event only with $M + 1$ distinct probability values (0, 1/M, 2/M, …, 1), given by number of ensemble members for which the event occurred. By stratifying cases into $K = M + 1$ bins according to value of forecasted probability p_k (bin $k = 1, 2, \ldots, K$ containing n_k cases), Brier score can be decomposed to reliability, resolution and uncertainity (see [2] for details). Forecast is perfectly reliable if for each k observed frequency $\bar{o}_k$ of the event in bin k equals to forecasted probability p_k. Reliability component of Brier score measures departure from perfectly reliable forecast (strictly speaking, it should be called unrealiability since it is negatively oriented – bigger value means less reliable forecast). Resolution component describes ability of the forecast to distinguish among situations with different observed frequencies of the event. Uncertainity component depends only on climatological frequency of the event $\bar{o}$ and it equals to Brier score of climatological forecast (i.e. forecast always predicting probability of the event to be $\bar{o}$). This is because climatological forecast is completely reliable, but has zero resolution.

Disadvantage of Brier score is its dependency on climatological frequency of the event. It can be eliminated by conversion to skill score, where climatology is often taken as reference forecast with zero skill:

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{clim}}}$$

Brier skill score is positively oriented and ranges from -∞ (imperfect forecast of never or always occuring event) to 1 (perfect forecast only). It is negative for forecasts worse than climatology. By inserting Brier score decomposition and using relation $\mathrm{BS}_{\mathrm{clim}}$ = uncertainity, it can be seen immediately that BSS = [resolution – (un)reliability]/uncertainity.

Ensemble calibration can be viewed as adjusting of probabilities p_k. It means that it can reduce forecast unreliability, but it does not affect resolution and uncertainity. Well calibrated ensemble would thus have Brier skill score close to forecast resolution divided by uncertainity.

Another important measure complementing Brier skill score is relative value (RV) of the forecast based on simple cost-loss economical model (see [3] for its definition). It evaluates forecast value from the point of view of decision maker, who wants to minimize expenses (E) given as sum of costs for protection against adverse weather and losses in cases when protection was not done and the damaging event occurred. Key parameter of the model is so called cost-loss ratio, dividing costs of single protecting action by losses caused by single unprotected event. From the long term point of view expenses are minimized when protective action is taken for forecasted probability of the event exceeding cost-loss ratio and no action is taken otherwise. Relative value of the forecast is then evaluated by comparing its expenses to those of climatology (relative value 0) and perfect forecast (relative value 1):

$$\mathrm{RV} = \frac{\mathrm{E}_{\mathrm{clim}} - \mathrm{E}}{\mathrm{E}_{\mathrm{clim}} - \mathrm{E}_{\mathrm{perf}}}$$

Forecast relative value depends on cost-loss ratio via expenses. Cost-loss ratio is restricted to values less than one, since it has no meaning to use protection equally or more expensive than possible damage.

Frequency bias (FB) is defined as ratio of forecasted to observed number of events. For probabilistic forecast it reads:

$$\mathrm{FB} = \frac{\sum_{i=1}^{N} p_i}{\sum_{i=1}^{N} o_i}$$

Frequency bias ranges from 0 to +∞. If the forecast has correct climatology of the event, frequency bias is 1. This is the case not only for perfect forecast.

Peirce skill score (PSS) is defined for deterministic forecast as probability of detection (POD) minus probability of false detection (POFD). It ranges from -1 (perfectly wrong forecast only) to 1 (perfect forecast only) and is popular because of the property called equitability – it assigns the same (zero) value to all random guessing “forecasts”. Peirce skill score can be generalized to probabilistic forecast as well:

$$\mathrm{PSS} = \underbrace{\frac{\sum_{i=1}^{N} p_i o_i}{\sum_{i=1}^{N} o_i}}_{\mathrm{POD}} - \underbrace{\frac{\sum_{i=1}^{N} p_i (1-o_i)}{\sum_{i=1}^{N} (1-o_i)}}_{\mathrm{POFD}} = \frac{1}{\bar{o}(1-\bar{o})} \cdot \frac{1}{N} \sum_{k=1}^{K} n_k p_k (\bar{o}_k - \bar{o})$$

Probabilistic version of Peirce skill score can be reexpressed in a way similar to Brier score decomposition. It is useful since it clearly shows that Peirce skill score can be maximized by setting $p_k = 0$ for $\bar{o}_k < \bar{o}$ and $p_k = 1$ otherwise. It means that it favours deterministic forecasts even if they are not reliable. Because of this behaviour it is not very well suited for probabilistic forecasts, where reliability concept is important. But at least it has some informative value.

Scores were evaluated individually for each hydrological zone and collectively for all zones together. In latter case sums occuring in score definitions were computed as weighted averages of individual sums, with weights being proportional to area and number of cases in each zone (case means situation with both forecasted and observed precipitation available).

4. Results

Categoric scores were evaluated for precipitation tresholds 0.1, 1.0, 2.5 and 5.0mm/6h, where the event was defined as precipitation exceeding given treshold. First treshold is the smallest possible forecasted precipitation due to rounding, last two tresholds correspond to hydrologically significant precipitation amounts 10 and 20mm/day.

Impact of calibration on rank histogram for 30-36 hour forecast is shown on figure 2 (when not stated explicitly, situation for other forecast lead times is qualitatively similar). It can be seen that calibration improves histogram flatness, centers its distribution and reduces number of outliers. Still the distribution is not perfect, there are too few observed cases below lowest rank and too many observed cases above highest rank. Detailed analysis proved that these deficiencies are caused by hydrologically unimportant cases. Number of cases on the left edge is strongly sensitive to rounding of calibrated values, which has biggest relative impact on small precipitation amounts (rounding to 0.1mm was used in order to be consistent with precision of input data; no rounding would give too many cases). On the other hand, too many outliers on the right edge are due to cases when all ensemble members forecasted zero precipitation (thus calibration could not change it), but in reality there was small observed precipitation.

Figure 4 shows reliability diagrams for raw and calibrated 30-36 hour forecast. Raw forecast is overconfident for all tresholds (it exaggerates frequency of the event), while calibration substantially reduces this defficiency (values are closer to diagonal; deviation of grey curve on the right edge of diagram might be due to small number of cases with high forecasted probabilities). Improved reliability is direct consequence of proposed calibration procedure which flattens rank histogram.

Figure 5 shows frequency bias for raw and calibrated ensembles. Frequency bias of raw ensemble has strong variation with forecast lead time. Except from 0-6 hour forecast it is greater than one (i.e. model forecasts precipitation events more frequently than climatology) and reaches local maximum for 6-12 and 30-36 hour forecast where it reaches values between 1.7 and 2.1. After calibration, frequency bias improves greatly – it does not depend on forecast lead time any more and its value remains between 0.8 and 1.2. This is again consequence of improved reliability.

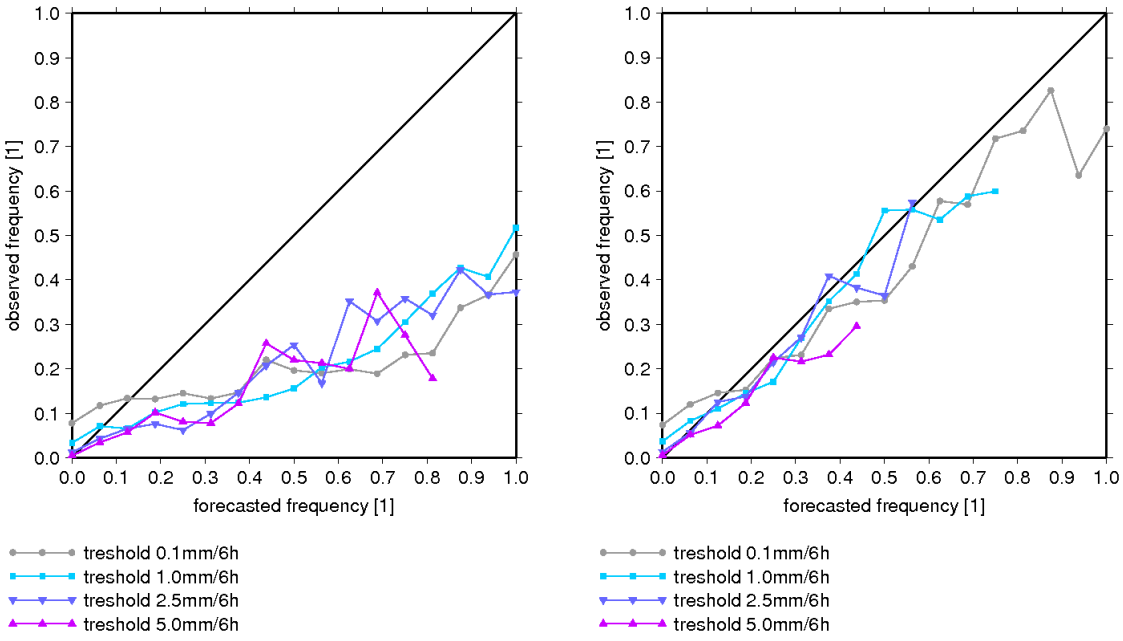

Fig. 4: *Reliability diagrams for 30-36 hour forecast and four different precipitation tresholds (year 2010, all zones): left – raw ensemble; right – calibrated ensemble. On horizontal axis is forecasted frequency, on vertical axis is associated observed frequency. Diagonal line denotes perfect reliability, only points with at least 30 forecasted cases are shown.*

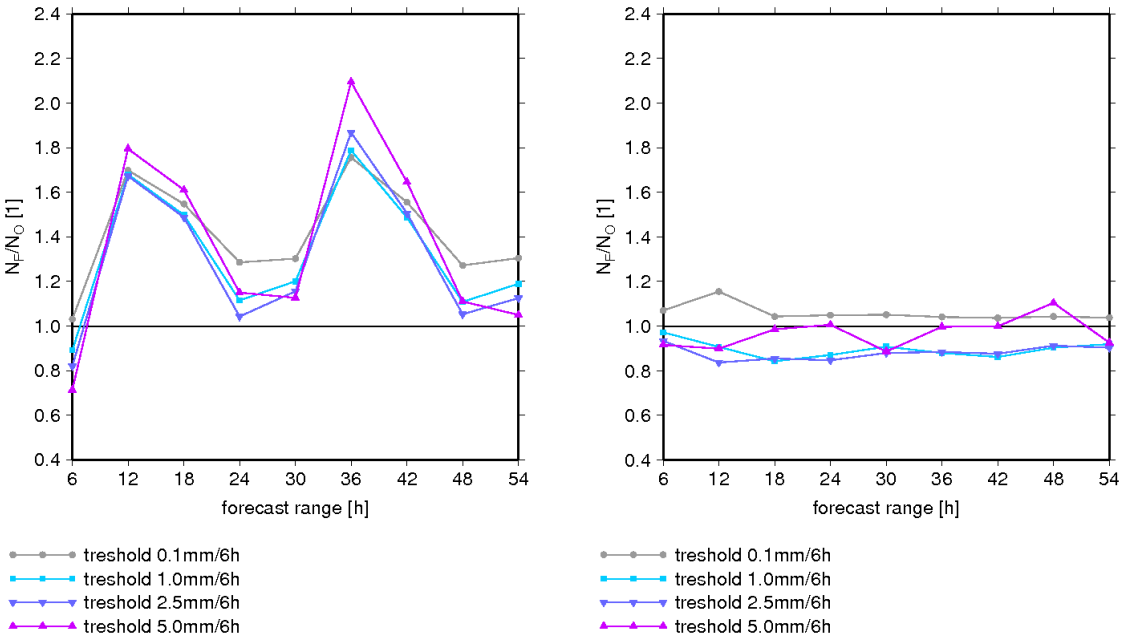

Fig. 5: *Frequency bias for four different precipitation tresholds (year 2010, all zones): left – raw ensemble; right – calibrated ensemble. On horizontal axis is forecast lead time, on vertical axis is frequency bias. Horizontal line denotes unbiased forecast.*

For probabilistic forecasts, Brier skill score is often taken as a standard quality measure, even if it is penalizing forecast unreliability too strongly. As a consequence, it can mark system with some predictive skill due to resolution to be worse than climatology which is static and has no resolution. Figure 6 shows Brier skill score for raw and calibrated ensembles. For raw ensemble it has strong dependency on forecast lead time and is dominated by negative values. After calibration all values are positive and the dependency on forecast lead time is much weaker.

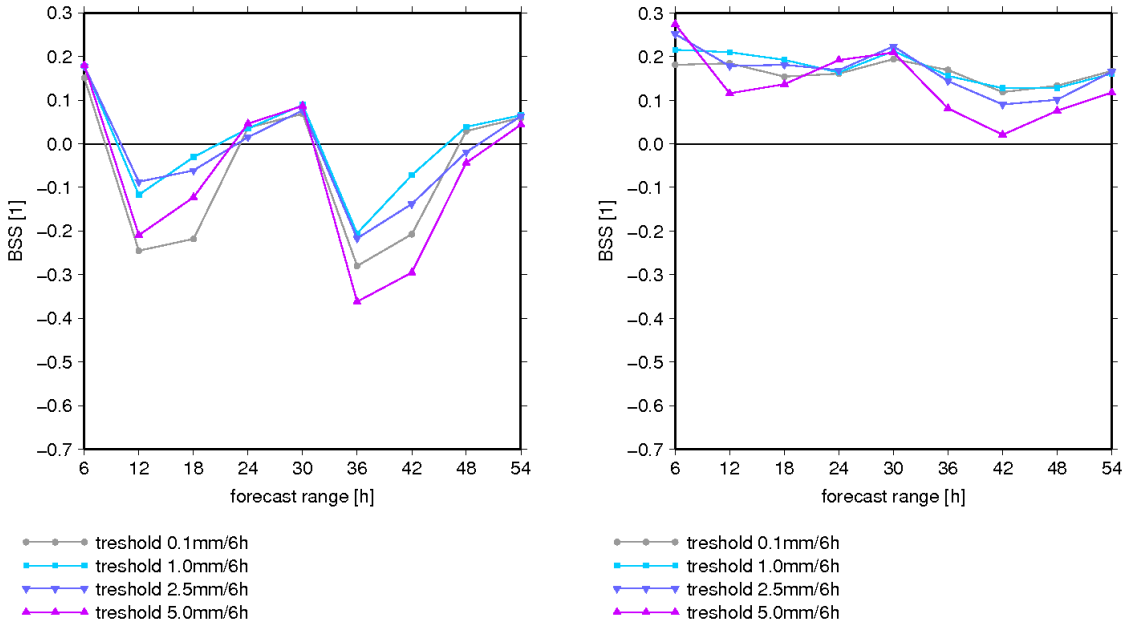

Fig. 6: *Brier skill score for four different precipitation tresholds (year 2010, all zones): left – raw ensemble; right – calibrated ensemble. On horizontal axis is forecast lead time, on vertical axis is Brier skill score. Horizontal line denotes climate forecast taken as reference with zero skill.*

Brier skill score can be complemented by forecast relative value based on simple cost-loss economical model. On figure 7 it is shown for raw and calibrated 30-36 hour forecast, with cost-loss ratio ranging from 0.001 to 1. Decision criterion for taking protective action was forecasted probability exceeding cost-loss ratio. It can be seen that forecast saves expenses with respect to climatology only for limited range of cost-loss ratios, reaching maximum savings (i.e. maximum relative value) for cost-loss ratio close to climatological frequency of the event. For cost-loss ratios much smaller than climatological frequency it is more advantageous to always take protection, since few forecast misses of the event can make losses higher than cost of permanent protection. On the other hand, for cost-loss ratios much higher than climatological frequency best strategy is to never take protection, since savings from protected events can be less than unnecessary costs due to false alarms.

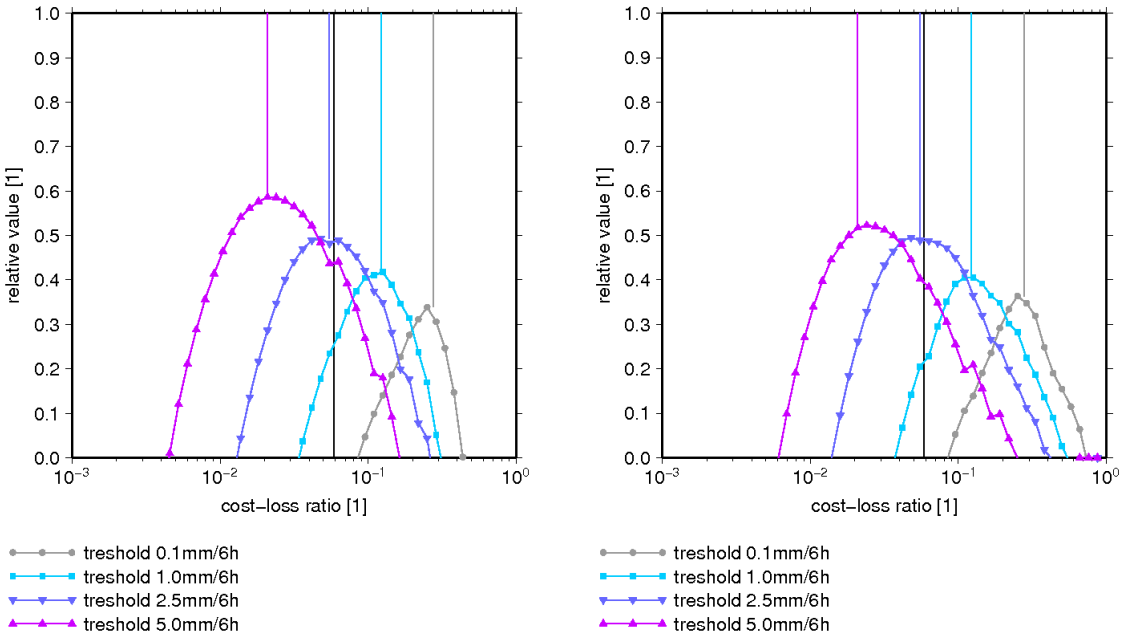

Fig. 7: *Relative value for 30-36 hour forecast and four different precipitation tresholds (year 2010, all zones): left – raw ensemble; right – calibrated ensemble. On horizontal axis is cost-loss ratio, on vertical axis forecast relative value (only positive values are plotted). Forecast value is scaled with respect to climatology (0) and perfect forecast (1). Coloured vertical lines denote climatological frequency of the event, black vertical line denotes minimum positive forecasted probability 1/M, where M is number of ensemble members.*

Figure 8 shows maximum forecast relative value for raw and calibrated ensembles. It depends on forecast lead time only slightly and the impact of calibration is roughly neutral. Maximum forecast relative value increases with precipitation treshold, i.e. it is bigger for less frequent high precipitation events.

Figure 9 shows Peirce skill score for raw and calibrated ensembles. As was explained in previous section, this score favours deterministic forecasts at the expense of reliability. It is no surprise then that it is better for raw ensemble with insufficient spread than for more reliable calibrated ensemble. Nevertheless, decrease of Peirce skill score due to calibration by roughly 25% is not dramatic. It is important that the score remains positive.

Finally, it is instructive to demostrate added value of ensemble forecast with respect to deterministic one. For this purpose, scores evaluated for ALADIN/LAEF raw control forecast are given on figure 10. In terms of Peirce skill score (top left panel) control forecast is better than calibrated ensemble and only slightly better than raw ensemble (figure 9). For other scores, however, situation is different. Brier skill score (top right panel) and forecast relative value (bottom left panel) are worse than for ensemble forecast (figures 6 and 7), in case of Brier skill score much worse even than raw ensemble. Maximum forecast relative value (bottom right panel) is worse than for ensemble forecast (figure 8), strongest deterioration is visible for highest precipitation treshold.

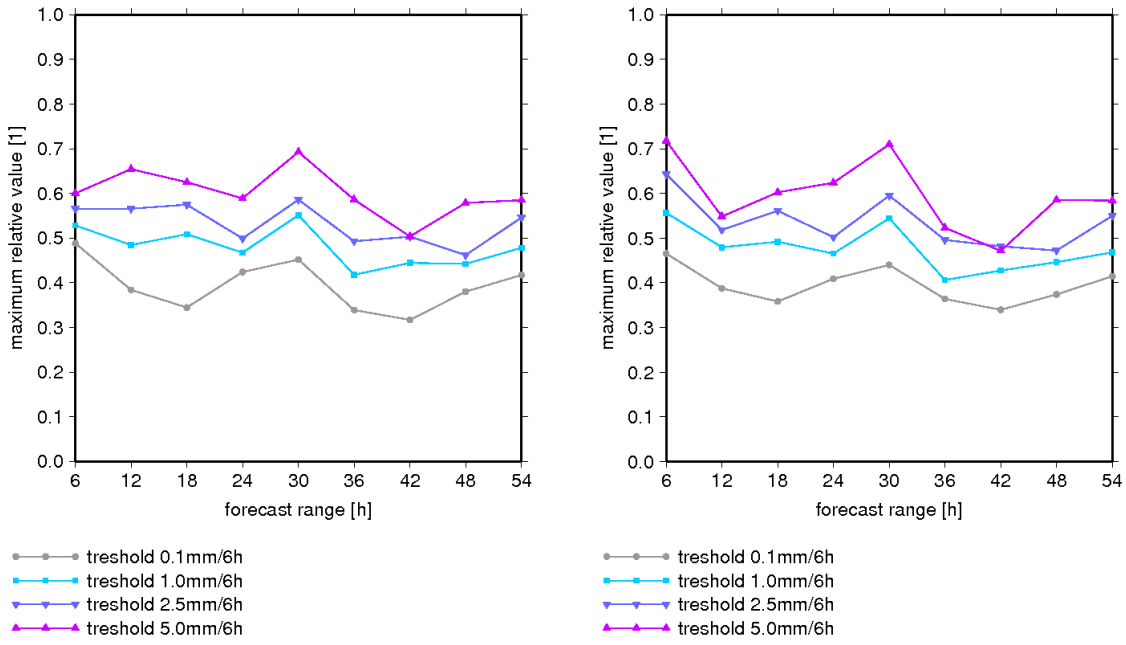

Fig. 8: *Maximum forecast relative value for four different precipitation tresholds (year 2010, all zones): left – raw ensemble; right – calibrated ensemble. On horizontal axis is forecast lead time, on vertical axis is maximum forecast relative value.*

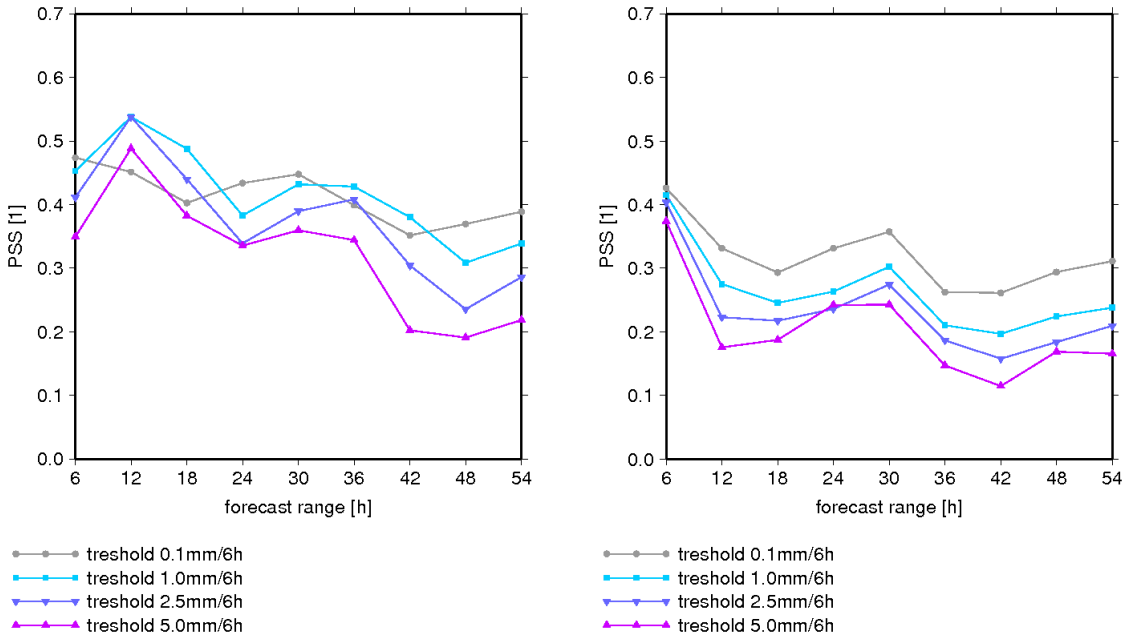

Fig. 9: *Peirce skill score for four different precipitation tresholds (year 2010, all zones): left – raw ensemble; right – calibrated ensemble. On horizontal axis is forecast lead time, on vertical axis is Peirce skill score. Climate forecast has zero skill.*

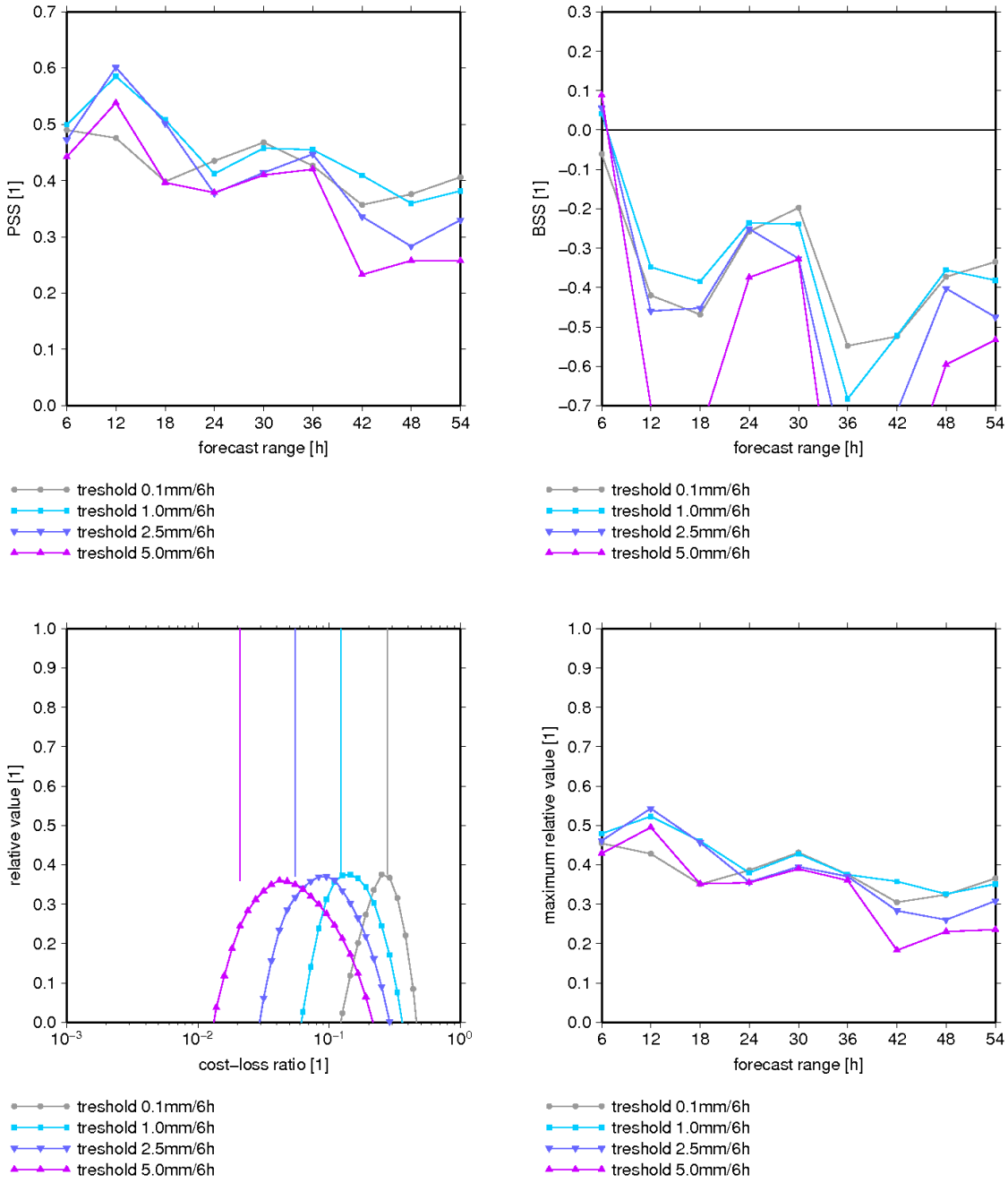

Fig. 10: *Performance of ALADIN/LAEF raw control forecast (year 2010, all zones): top left – Peirce skill score; top right – Brier skill score; bottom left – relative value of 30-36 hour precipitation forecast; bottom right – maximum forecast relative value. To be compared with figures 9, 6, 7 and 8.*

5. Summary and Conclusions

Method for calibration of ALADIN/LAEF precipitation ensembles was proposed and evaluated in 29 hydrological zones. For each zone, forecasted precipitation scaled by ensemble mean is transformed by 2-parametric nonlinear function, which enables to reduce overestimated mean precipitation and at the same time increase insufficient ensemble spread. Calibration parameters are determined from requirement that calibrated ensemble has the same yearly precipitation (climatology) as observed and its rank histogram is as flat as possible. Flattening of rank histogram improves forecast reliability, while correct climatology removes bias (more precisely yearly bias, not the seasonal one which can vary across the year).

Ensemble calibration reduces strong diurnal course visible for some scores, making model accuracy within forecast window quite uniform. Comparison of categoric scores for calibrated versus raw ensembles showed that for all used precipitation tresholds calibration significantly improves Brier skill score and almost completely removes frequency bias. Surprisingly, impact on forecast relative value is neutral (if this was the only relevant measure, calibration would not be necessary). Deterioration of Peirce skill score is not assumed critical, since this score favours overconfident forecasts which are necessarily unreliable. On the other hand, Brier skill score penalizes lack of reliability too strongly, often prefering climatology (which is perfectly reliable but has zero resolution) to model forecast apparently having some added value. This is why Brier skill score was complemented by forecast relative value based on simple cost-loss economical model, which is more relevant for decision makers. Forecast relative value is positive only for limited range of cost-loss ratios, reaching maximum for cost-loss ratio close to climatological frequency of the event. Maximum relative value increases with precipitation treshold, which is advantageous for hydrological application where high precipitation events are of primary interest.

Comparison of scores against deterministic forecast represented by control run clearly demonstrates added value of precipitation ensembles both in terms of Brier skill score and forecast relative value. Ensemble calibration further shifts the situation in favour of probabilistic forecast. This is the key result justifying investments into ensemble forecasting.

Both calibration and verification were done for year 2010. Verification on independent data set would be desirable, but for this at least two years of ALADIN/LAEF forecasts will be needed (one year for calibration, another for verification). They should be available at the end of 2011. Independent data set will provide somewhat worse scores for calibrated ensemble, but deterioration of forecast relative value should be small anyhow. Ideally, calibration period of several years would be preferable in order to increase number of high precipitation events and thus get more stable statistics. Here the main problem is inhomogenity of forecast data set, caused by changes of both regional ALADIN/LAEF system and global IFS system providing perturbed boundary condtions.

Presented results were obtained for geographical region of Czech Republic, but they should remain at least qualitatively valid in much wider area. Proposed calibration procedure is applicable to arbitrary precipitation ensemble with sufficiently long data set. Verification was restricted to mean area precipitation in zones few thousand square kilometers in size. Point verification was not performed due to big representativness error of observed precipitation.

6. Acknowledgements

Work was financially supported by grant SP/1c4/16/07 of Czech ministry of environment. ALADIN/LAEF forecasts were provided by RC LACE consortium, who developed the system under the leadership of Austria and operates it at ECMWF. Precipitation measurements covering the territory of Czech Republic (both raingauge network and radars) were provided by CHMI.

7. References

[1] Hamill, T. M., 2001: Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Mon. Wea. Rev.,* **129**, 550-560.

[2] Mason, S. J., 2004: On Using “Climatology” as a Reference Strategy in the Brier and Ranked Probability Skill Scores. *Mon. Wea. Rev.,* **132**, 1891-1895.

[3] Murphy, A. H., 1977: The Value of Climatological, Categorical and Probabilistic Forecasts. *Mon. Wea. Rev.,* **105**, 803-816.

[4] Murphy, A. H., and M. Ehrendorfer, 1987: On the Relationship between the Accuracy and Value of Forecasts in the Cost-Loss Ratio Situation. *Wea. Forecasting,* **2**, 243-251.

[5] Wang, Y. et al., 2009: The Central European limited area ensemble forecasting system: ALADIN-LAEF. *RC LACE report, available at* http://www.rclace.eu/File/Predictability/2009/laef4lace.pdf.