Received: 29 May 2023 | Revised: 25 September 2023 | Accepted: 1 October 2023

DOI: 10.1002/met.2157

RESEARCH ARTICLE

Meteorological Applications
Science and Technology for Weather and Climate

Statistical post-processing of visibility ensemble forecasts

Sándor Baran[1] | **Mária Lakatos**[1,2]

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

[2]Doctoral School of Informatics, University of Debrecen, Debrecen, Hungary

Correspondence
Sándor Baran, Faculty of Informatics, University of Debrecen, Kassai út 26, H-4028 Debrecen, Hungary.
Email: baran.sandor@inf.unideb.hu

Funding information
Innovációs és Technológiai Minisztérium, Grant/Award Number: ÚNKP-22-3; National Research, Development and Innovation Office, Grant/Award Number: K142849

Abstract

To be able to produce accurate and reliable predictions of visibility has crucial importance in aviation meteorology, as well as in water- and road transportation. Nowadays, several meteorological services provide ensemble forecasts of visibility; however, the skill and reliability of visibility predictions are far reduced compared with other variables, such as temperature or wind speed. Hence, some form of calibration is strongly advised, which usually means estimation of the predictive distribution of the weather quantity at hand either by parametric or nonparametric approaches, including machine learning-based techniques. As visibility observations—according to the suggestion of the World Meteorological Organization—are usually reported in discrete values, the predictive distribution for this particular variable is a discrete probability law, hence calibration can be reduced to a classification problem. Based on visibility ensemble forecasts of the European Centre for Medium-Range Weather Forecasts covering two slightly overlapping domains in Central and Western Europe and two different time periods, we investigate the predictive performance of locally, semi-locally and regionally trained proportional odds logistic regression (POLR) and multilayer perceptron (MLP) neural network classifiers. We show that while climatological forecasts outperform the raw ensemble by a wide margin, post-processing results in further substantial improvement in forecast skill, and in general, POLR models are superior to their MLP counterparts.

KEYWORDS

classification, ensemble calibration, multilayer perceptron, proportional odds logistic regression, visibility

1 | INTRODUCTION

According to the definition of the World Meteorological Organization (WMO), visibility "is the greatest distance at which a black object of suitable dimensions (located on the ground) can be seen and recognized when observed against the horizon sky" (WMO, 1992). This weather quantity plays a crucial role in aviation in all phases of flight and restricted visibility also makes both ship navigation and road transportation difficult. Hence, accurate and reliable visibility forecasts result in a direct economic benefit.

In general, weather forecasts are generated using numerical weather prediction (NWP) models, which are

often run several times with varying model physics and/or initial conditions resulting in a probabilistic forecast represented by an ensemble of predictions (Bauer et al., 2015; Buizza, 2018a). With the help of a forecast ensemble, besides providing point forecasts for a given location, time point, and forecast horizon, one can also assess forecast uncertainty and estimate probability distributions of future weather variables (Gneiting & Raftery, 2005). However, in contrast to, for example, temperature, wind speed, or precipitation accumulation, most NWPs do not explicitly model visibility, so visibility forecasts must be derived from predictions of related quantities such as relative humidity or precipitation (Chmielecki & Raftery, 2011). Although several recent studies verified the efficiency of probabilistic predictions, for example, in fog forecasting (see Pahlavan et al., 2021; Parde et al., 2022), even nowadays only a few weather centers issue visibility ensemble forecasts. Examples include the multimodel Short-Range Ensemble Forecast System of the National Centers for Environmental Prediction covering the Continental US, Alaska, and Hawaii regions (Zhou et al., 2009) or the Ensemble Prediction System (EPS) of the European Centre for Medium-Range Forecasts (ECMWF; Molteni et al., 1996; ECMWF Directorate, 2012), where visibility is part of the Integrated Forecasting System (IFS) since 2015 (ECMWF, 2021).

Despite the continuous improvement in the various operational EPSs over the past decades, ensemble forecasts still might display systematic bias or suffer from lack of calibration (see, e.g., Buizza et al., 2005), which calls for some form of post-processing (Buizza, 2018b). Moreover, visibility forecasts are even more problematic, as their predictive performance, similar to total cloud cover (Haiden et al., 2021), is highly below the skill of ensemble forecast of, for example, temperature, wind speed, pressure, or precipitation accumulation (see, e.g., Zhou et al., 2012). There are several reasons that make visibility prediction challenging. For instance, the coarse vertical grid resolution of the operational NWP models might make it difficult to appropriately resolve processes affecting fog formulation (Gultepe et al., 2006) and there might also be problems in the parameterization of factors affecting fog microphysics in a polluted environment (see, e.g., Wagh et al., 2023).

Nowadays, one can select from a large collection of post-processing methods developed for a wide variety of weather variables; for an overview of the state-of-the-art techniques, see, for example, Wilks (2018) or Vannitsem et al. (2021). Parametric approaches like nonhomogeneous regression (Gneiting et al., 2005) or Bayesian model averaging (BMA; Raftery et al., 2005) provide full predictive distribution of the investigated weather quantity, whereas quantile regression-based approaches (see, e.g., Friederichs & Hense, 2007; Bremnes, 2019) result in probabilistic forecasts by estimating the quantiles of the forecast distribution. Moreover, in the last few years, machine learning techniques such as quantile regression forests (Taillardat et al., 2016), distributional regression network (Rasp & Lerch, 2018), Bernstein quantile network (Bremnes, 2020), or the classification and interpolation-based approach of Scheuerer et al. (2020) gain more and more popularity; for a recent comparison of neural network-based approaches, we refer to Schultz and Lerch (2022). However, according to the best knowledge of the authors, not many of the above techniques are applied in the context of visibility forecasting. Chmielecki and Raftery (2011) propose a BMA model for calibrating ensemble forecasts of visibility, Ryerson and Hacker (2018) consider a nonparametric post-processing approach, whereas machine learning-based methods are mainly used only for generating visibility forecasts from the NWP outputs of other weather quantities (see, e.g., Marzban et al., 2007; Dietz et al., 2019).

While visibility forecasts can be considered as continuous (e.g. ECMWF ensemble forecasts are provided in 1-m steps), most synoptic observation (SYNOP) stations report observations according to the WMO suggestions, that is "100–5 000 m in steps of 100 m, 6–30 km in steps of 1 km, and 35–70 km in steps of 5 km" (WMO, 2018, section 9.1.2), and observed visibility is rounded down to the nearest reported value. In this way, one has just 84 different values, so the corresponding predictive distribution is a discrete probability law. Hence, post-processing of visibility forecasts can be considered as an 84-group classification problem resulting in the probabilities of the different reported values. The situation is similar to the case of total cloud cover (TCC) reported in eighths of the sky covered by clouds called oktas, taking just nine different values, only in this case the number of classes is nearly 10 times larger. To calibrate TCC ensemble forecasts, Hemri et al. (2016) proposed multiclass logistic regression (MLR; Izenman, 2008) and proportional odds (or ordered) logistic regression (POLR; McCullagh, 1980), whereas Baran et al. (2021) investigated the forecast skill of several machine learning-based classification methods such as multilayer perceptron neural network (MLP; Goodfellow et al., 2016), gradient boosting machine (Friedman, 2001), and random forest algorithms (Breiman, 2001).

In the present work, we investigate the predictive performance of POLR and MLP approaches to calibration of ECMWF visibility ensemble forecasts, as these two methods exhibit the best forecast skill among the techniques investigated by Baran et al. (2021) in the context of post-processing ensemble predictions of TCC. As reference forecasts, we consider the raw visibility ensemble and climatology.

The article is organized as follows. After a brief description of the studied visibility datasets in Section 2, we review the POLR and MLP methods in Section 3 and

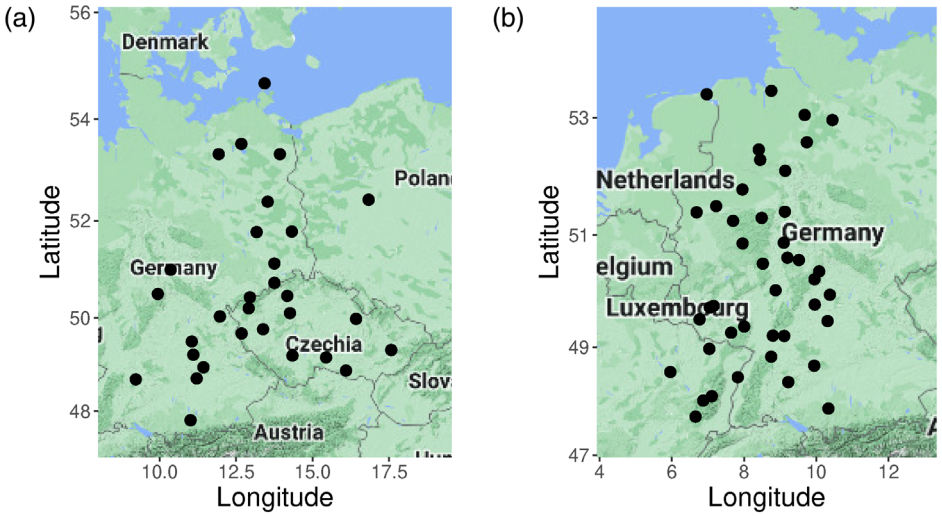

FIGURE 1 Locations of SYNOP observation stations corresponding to (a) ECMWF forecasts for 2020–2021; (b) EUPPBench benchmark dataset.

also provide the investigated approaches to training data selection and the considered verification tools. Results of our two case studies are reported in Section 4, followed by a short discussion and conclusions in Section 5.

2 | DATA

We consider two datasets consisting of ECMWF visibility ensemble forecasts and corresponding validating observations covering different time periods but having slightly overlapping ensemble domains. As mentioned, observations are reported according to the WMO suggestions in values:

$$\mathcal{Y} = \{0, 100, 200, \ldots, 4900, 5000, 6000, 7000, \ldots,$$

$$29000, 30000, 35000, 40000, \ldots, 65000, 70000\},$$

whereas the matching of forecasts (given in 1 m steps) and observations is performed by rounding down to the closest reported value.

In the case of the first dataset, as predictions, we have the operational 51-member ECMWF visibility ensemble forecasts (control forecast (CTRL) and 50 members (ENS) generated using random perturbations, which are statistically indistinguishable and should be treated as exchangeable) for calendar years 2020 and 2021 for 30 SYNOP stations in Germany, Czech Republic, and Poland (Figure 1a). All forecasts are initialized at 0000 UTC, and we consider 40 different lead times from 6 to 240 h with a time step of 6 h. This data set is fairly complete as there are no missing observations at all and one has just 2 days when some predictions are missing; namely, there are no exchangeable ensemble members with forecast horizon 132 h initialized 2 June 2021 and 114 h initialized 17 December 2021.

We also investigate visibility data from the *EUPPBench* benchmark dataset (Demaeyer et al., 2023), where besides the 51-member operational ECMWF ensemble, the high-resolution (HRES) forecast is also available. The studied dataset consists of ensemble forecasts for calendar years 2017–2018 initialized at 0000 UTC with a forecast horizon of 120 h and temporal resolution of 6 h for 42 SYNOP stations in Germany and France (Figure 1b). Here around 1.5% of the forecast cases is incomplete due to missing station observations. Note that the *EUPPBench* benchmark dataset also contains gridded data including ECMWF visibility forecasts and reforecasts. However, the use of reforecasts would not be consistent with the more recent dataset; moreover, reforecasts are available only on dates corresponding to Mondays and Thursdays.

3 | POST-PROCESSING METHODS AND VERIFICATION TOOLS

Following the notations by Baran et al. (2021), denote by $Y \in \mathcal{Y} = \{y_1, y_2, \ldots, y_{84}\}$ the observed visibility for a given location and time point with $\mathcal{Y}$ defined in Section 2, and let $\boldsymbol{f} = (f_1, f_2, \ldots, f_{52})$ be the corresponding ensemble forecast with a given forecast horizon, where $f_1 = f_{HRES}$ denotes the high-resolution forecast, $f_2 = f_{CTRL}$ is the control run, and $f_3, f_4, \ldots, f_{52}$ correspond to the 50 exchangeable members $f_{ENS,1}, f_{ENS,2}, \ldots, f_{ENS,50}$. As Y is a discrete random variable, the predictive distribution of Y is specified by conditional probabilities with respect to the ensemble forecast

$$\mathrm{P}(Y = y_k | \boldsymbol{f}), \quad k = 1, 2, \ldots, 84. \tag{1}$$

Naturally, instead of the raw forecast $\boldsymbol{f}$, in Equation (1), the conditional probability can be taken with respect

to any feature vector $\boldsymbol{x}$ derived from the ensemble and/or other covariates such as forecasts of additional weather quantities or station-specific information like location, altitude, or land use.

3.1 | Proportional odds logistic regression

As mentioned in the Introduction, in the case of discrete weather quantities, post-processing is reduced to a classification problem, where MLR is a frequently used and powerful tool. However, given an M-dimensional feature vector $\boldsymbol{x}$, an MLR model for an 84-group classification has $83(M+1)$ free parameters to be estimated from the training data. Hence, to avoid numerical issues in the estimation process, one requires an extremely large training data set consisting of past forecast–observation pairs.

A more parsimonious approach is the POLR model designed to fit ordered data, like the visibility observations at hand. In this case, the conditional cumulative distribution of Y with respect to an M-dimensional feature vector $\boldsymbol{x}$ is given as

$$\mathrm{P}(Y \le y_k|\boldsymbol{x}) = \frac{\mathrm{e}^{\mathscr{L}_k(\boldsymbol{x})}}{1+\mathrm{e}^{\mathscr{L}_k(\boldsymbol{x})}}, \quad \text{with}$$

$$\mathscr{L}_k(\boldsymbol{x}) := \alpha_k + \boldsymbol{x}^\top \boldsymbol{\beta}, \quad k = 1, 2, ..., 84,$$

where $\alpha_k \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^M$ and we assume $\alpha_1 < \alpha_2 < \cdots < \alpha_{84}$, so the POLR model has $84+M$ unknown parameters.

3.2 | Multilayer perceptron neural network

For both continuous and discrete weather quantities, the application of neural networks has increased in popularity for calibration. For the latter, instead of a parametric approach, one can use a classical feedforward multilayer perceptron (MLP) neural network for classification. A classical MLP consists of multiple layers and neurons or nodes, each of which is a transformed (by an activation function) weighted sum of the node values from the previous layer plus an additional bias term. The features are "fed" to the input layer, and the predictions of the distribution of the various classes are made on the output layer. The number of hidden layers and neurons in them are tuning parameters of the network and provide the level of abstraction. For additional information, we refer to Goodfellow et al. (2016).

3.3 | Training data selection

Both parameters of the POLR model and weights of the MLP neural network are estimated with the help of training data, and the spatial and temporal composition of this set of forecast–observation pairs is a key issue in statistical post-processing. The basic approaches to spatial selection are local and regional (global) modeling (Thorarinsdottir & Gneiting, 2010). In the local case only past data of the station of interest are used to obtain the predictive distribution resulting in distinct POLR model parameters and MLP networks for the different stations. Provided one has a long enough training period to avoid numerical issues (for optimal training period lengths for different weather quantities see, for example, Hemri et al., 2014), local models usually outperform their regional counterparts, which pool training data of the whole ensemble domain and all stations share the same set of POLR model parameters and a single MLP network. Regional modeling can be performed with the help of rather short training periods; however, this approach is not really suitable for large and heterogeneous domains. As a bridge between these two traditional spatial selection methods, Lerch and Baran (2017) propose a semi-local approach, where first clusters of stations with similar characteristics are formed using k-means clustering, and then within each cluster a regional estimation is performed. Clustering-based semilocal modeling is more adaptive to the differences, for example, in the climatology of the various observation stations than the regional approach, and preferable in situations when one does not have enough training data for local modeling (see, e.g., Baran et al., 2020; Szabó et al., 2023).

Concerning the temporal selection, a popular approach is the use of rolling training periods, where models are trained with the help of ensemble forecasts and validating observations from the preceding n calendar days. This method enables models to quickly adapt, for example, to seasonal changes; however, in many situations, the use of large fixed training data sets (see, e.g., Rasp & Lerch, 2018; Ghazvinian et al., 2021) or yearly training considering data of the last couple of years before the year of the given date (see, e.g., Hemri et al., 2016; Baran et al., 2021) might also be beneficial.

3.4 | Verification scores

As proposed by Gneiting et al. (2007), the evaluation of the predictive performance of probabilistic forecasts should be based on the idea of "maximizing the

sharpness of the predictive distributions subject to calibration". Calibration means a statistical consistency between forecasts and observations, while sharpness refers to the concentration of the predictive distribution.

A simple graphical tool for assessing calibration is the probability integral transform (PIT) histogram (Wilks, 2019, section 9.5.4). The PIT is defined as the value of the predictive cumulative distribution function (CDF) evaluated at the verifying observation, where for noncontinuous laws randomization should be applied at points of discontinuity (Gneiting & Ranjan, 2013). For a properly calibrated probabilistic forecast, the PIT follows a standard uniform distribution; moreover, the shape of the PIT histogram can provide hints to the possible reason for the lack of calibration.

The sharpness of a probabilistic forecast can be investigated by examining the widths of various prediction intervals. In the case studies of Section 4, similar to Hemri et al. (2016), we compare the competing forecasts in terms of the average width of the centered 90% prediction interval. Furthermore, one can also calculate the coverage of this interval defined as the proportion of validating observations located between the lower and upper 5% quantiles of the predictive distribution. The deviation in coverage from the confidence level can serve as a measure of calibration.

Calibration and sharpness can also be assessed simultaneously using proper scoring rules, which are loss functions assigning numerical values to forecast–observation pairs (F,x). Here we consider the continuous ranked probability score (CRPS; Wilks, 2019, section 9.5.1) and the logarithmic score (LogS; Good, 1952), which are the most popular scoring rules in atmospheric sciences. Note that in the case of visibility reported according to WMO suggestions (see Section 2), a forecast F is characterized by a probability mass function (PMF) $p_F(y)$ specifying a discrete probability distribution on $\mathcal{Y}$. Hence, the CRPS equals

$$CRPS(F,x) = \sum_{k=1}^{84} p_F(y_k)\,|y_k - x| \qquad (2)$$

$$-\sum_{k=2}^{84}\sum_{\ell=1}^{k-1} p_F(y_k)p_F(y_\ell)\,|y_k - y_\ell|,$$

whereas the LogS is defined as

$$LogS(F,x) := -\log(p_F(x)),$$

that is, as the negative logarithm of the PMF at the observation. Both CRPS and LogS are negatively oriented (the smaller the better), and expression (2) is the discrete form of the representation of the CRPS

$$CRPS(F,x) = \mathrm{E}\,|X - x| - \frac{1}{2}\mathrm{E}\,|X - X'|$$

provided by Gneiting and Raftery (2007), where X and X' are independent random variables with distribution F and finite first moment.

For each of the investigated forecast horizons, the predictive performance of various probabilistic forecasts is quantified by the mean CRPS ($\overline{\mathrm{CRPS}}$) and mean LogS ($\overline{\mathrm{LogS}}$) over all forecast cases in the verification data. Furthermore, one can obtain a deeper insight into the smaller differences in forecast skill of the compering predictions by examining continuous ranked probability skill scores (CRPSS) and logarithmic skill scores (LogSS), which measure the improvement in terms of CRPS and LogS of a forecast F with respect to a reference forecast F_{ref}, respectively (see, e.g., Murphy, 1973; Gneiting & Raftery, 2007). CRPSS and LogSS are defined as

$$\mathrm{CRPSS} := 1 - \frac{\overline{\mathrm{CRPS}}}{\overline{\mathrm{CRPS}}_{ref}} \quad \text{and}$$

$$\mathrm{LogSS} := 1 - \frac{\overline{\mathrm{LogS}}}{\overline{\mathrm{LogS}}_{ref}},$$

where $\overline{\mathrm{CRPS}}$, $\overline{\mathrm{LogS}}$ and $\overline{\mathrm{CRPS}}_{ref}$, $\overline{\mathrm{LogS}}_{ref}$ denote the mean score values corresponding to forecasts F and F_{ref}, respectively. Note that skill scores are positively oriented, so larger values mean better predictive performance.

Furthermore, as point forecasts, we consider the ensemble mean and the means of the predictive distributions of the competing forecasts, which are evaluated with the help of the corresponding root mean squared errors (RMSEs).

Finally, the uncertainty of the verification scores and the statistical significance of the score differences are addressed by providing confidence intervals for the skill scores. The standard deviations required for the confidence bounds are calculated from 2000 block bootstrap samples obtained using the stationary bootstrap scheme, where the mean block length is computed according to Politis and Romano (1994).

4 | RESULTS

In the case studies of Sections 4.2 and 4.3, the forecast skill of the POLR and MLP approaches described in Sections 3.1 and 3.2, respectively, are evaluated.

Both methods require training data, which should be large enough to allow reliable modeling. We consider 350-day rolling training periods, which in the case of regional estimation means 10,500 forecast cases for the first dataset and 14,700 forecast cases for the second one for each training step. In what follows, regional POLR and MLP models are referred to as *POLR-R* and *MLP-R*, respectively. We also investigate clustering-based semilocal modeling, where the clusters of stations are derived using k-means clustering of feature vectors depending on station climatology over the training period. In particular, each station is represented by a three-dimensional feature vector providing the frequencies of visibility intervals 0–5000, 5000–30,000, and 30,000–70,000 m. With the shift in the training data, the clusters are recalculated dynamically, and in general, four clusters are formed in the case of forecasts for 2020–2021, and eight clusters in the case of the EUPPBench data, provided each cluster contains at least four locations. Otherwise, the number of clusters is reduced and the clustering-based estimation might even fall back to regional modeling. For the semilocal POLR and MLP approaches, notations *POLR-C* and *MLP-C* are used and note that the above feature vectors and the applied number of clusters are based on a preliminary analysis of the climatology of each station. Finally, local training is also investigated, and the corresponding models are denoted as *POLR-L* and *MLP-L*.

In both case studies, the forecast skill of the competing post-processing methods is tested on data for a complete calendar year. As reference forecasts, we consider the raw ECMWF visibility ensemble and climatology, where observations of a rolling training period of a given length are considered as a forecast ensemble.

Before calculating the input features of the POLR and MLP models, a normalization of the forecasts is performed; namely, all forecast values are divided by 70,000. Then we consider the following set of predictors: the (normalized) control member of the ensemble $\widetilde{f}_{CTRL} := f_{CTRL}/70000$, the mean $\overline{f}_{ENS}$ of the 50 (normalized) exchangeable members, the variance s^2 of the 51-member (normalized) operational ensemble, in the case of the EUPPBench dataset the (normalized) high-resolution forecast $\widetilde{f}_{HRES} := f_{HRES}/70000$, and the proportions p_1, p_2, and p_3 of ensemble members predicting visibility up to 1000 m, 1000–2000 m, and more than 30,000 m, respectively. The use of the control member, the mean of the exchangeable members, the ensemble variance, and possibly the high-resolution forecast as predictors is in line with the suggestions by Hemri et al. (2016) and rather common in parametric modeling based on ECMWF ensemble forecasts (see, e.g., Gneiting, 2014), whereas covariates p_1, p_2, and p_3 (both the number of such proportions to be considered and the boundaries) were selected based on a detailed data analysis. Note that normalization is required to have all input features commensurate, which makes modeling numerically more stable. Finally, following the suggestions of, for example, Dabernig et al. (2017), seasonal variations are addressed by adding annual base functions

$$\beta_1(d) := \sin(2\pi d/365) \quad \text{and} \quad \beta_2(d) := \cos(2\pi d/365)$$

to the input features, where d denotes the day of the year.

4.1 | Implementation details

POLR models are fit using the R package MASS (Venables & Ripley, 2002), while MLP classification is based on the package RSNNS, making the Stuttgart Neural Network Simulator (Zell et al., 1994) available in R. The neural network has two hidden layers with 25–25 neurons, the maximal number of iterations to learn is restricted to 200, the learning rate is 0.2, and both hidden layers use the logistic activation function, which parameterization is a result of extended experimentation.

To handle numerical problems with LogS calculation caused by zero predicted probabilities resulting in infinite score values, we follow the procedure applied by Hemri et al. (2016) and Baran et al. (2021). Extremely low values of the PMF are replaced with a probability $p_{\min}$, which ensures that the corresponding reported visibility at the given observation hour appears at least once a year with a pre-specified probability π. This means that for an observation y_j instead of $p_F\left(y_j\right)$, we consider $\max\left\{p_{\min}, p_F\left(y_j\right)\right\}$, where $p_{\min} = 1-(1-\pi)^{1/365}$, and normalize the obtained values to get a PMF again. For $\pi = 0.01$ suggested by Hemri et al. (2016), this approach results in $p_{\min} = 2.75 \times 10^{-5}$. Note that this modification of the PMF does not result in visible changes in other verification scores.

4.2 | Calibration of 51-member visibility ensemble forecasts

We study the post-processing of visibility ensemble forecasts with the help of locally, semi-locally, and regionally trained POLR and MLP models based on the eight-dimensional feature vector $\left(\widetilde{f}_{CTRL}, \overline{f}_{ENS}, s^2, p_1, p_2, p_3, \beta_1, \beta_2\right)^\top$. Following the suggestions by Hemri et al. (2016), for the POLR models, the weights of $\widetilde{f}_{CTRL}$ and $\overline{f}_{ENS}$ are kept non-negative by excluding iteratively covariates with negative

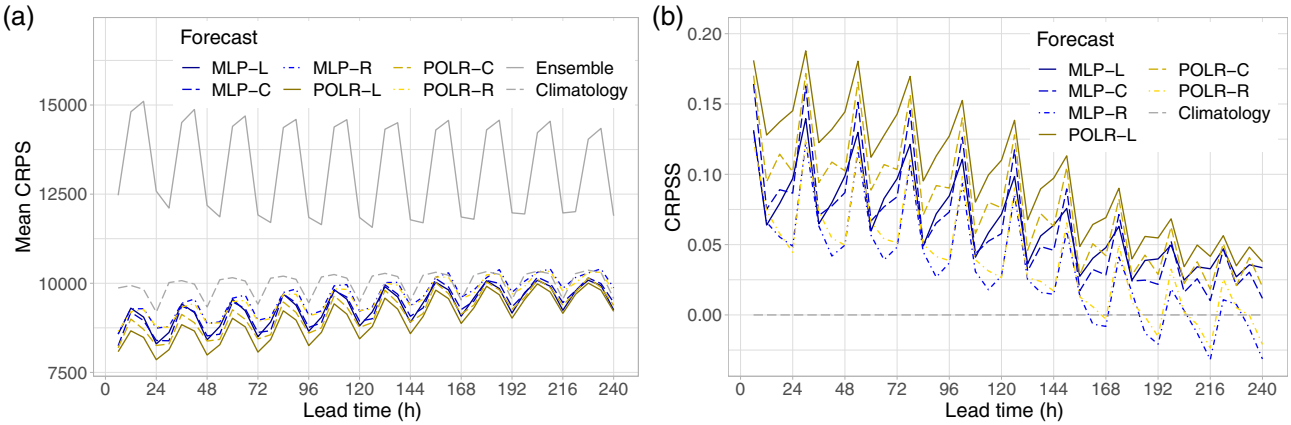

FIGURE 2 Mean CRPS of post-processed, raw, and climatological visibility forecasts for calendar year 2021 (a) and CRPSS of post-processed forecasts with respect to climatology (b) as functions of the lead time.

coefficients. The predictive performance of the competing post-processing methods is tested on data for the calendar year 2021. In the case of the reference climatological forecasts, we consider a 30-day rolling training period, which value is chosen by the comparison of the overall mean CRPS, RMSE of mean forecasts and coverage of 90% central prediction intervals for training periods of length 25, 30, ..., 50 and 350 days. The 350-day training period was tested to match the training period of the post-processing models; however, such a long time interval cannot capture seasonal effects as well as the short ones.

According to Figure 2a, in terms of the mean CRPS, all post-processing methods outperform the raw ECMWF ensemble forecast for all forecast horizons by a wide margin and climatology also performs rather well. The clear diurnal cycle in the mean CRPS corresponds to the different observation times (0000, 0600, 1200, and 1800 UTC), indicating that the skill of the predictions depends on the time of the day. Raw ensemble forecasts show the best predictive performance at 0600 UTC, whereas both for climatology and for post-processed predictions, the lowest mean CRPS is attained at 0000 UTC. Note that the superiority of climatology over the raw visibility forecasts is fully in line with the results of Chmielecki and Raftery (2011). A deeper insight into the differences between the various calibration techniques can be obtained from Figure 2b depicting the CRPSS values with respect to climatology. In general, clustering-based modeling is superior to regional, and the advantage of post-processing with respect to climatology decreases with the increase in the lead time. For POLR models, the chosen training period length is long enough even for reliable local estimation, and the POLR-L approach results in the highest skill score for all lead times. This is not the case for the MLP models, where the CRPSS of the semilocally estimated MLP-C is often above the skill score of the MLP-L, especially for lead times corresponding to 0600 UTC. POLR models outperform their MLP counterparts for most of the forecast horizons; however, up to 156 h, even the poorest MPL-R has better skill than climatology. A possible ranking of the competing methods can be obtained from Table 1, giving the overall mean score values of calibrated and climatological forecasts as proportions of the corresponding mean score of the raw ECMWF ensemble. POLR-L results in the lowest mean CRPS, closely followed by POLR-C.

As Figure 3a and Table 1 show, the difference between raw and post-processed forecasts in terms of the mean LogS is even more pronounced than in terms of the mean CRPS; however, the general picture slightly changes. On the one hand, all post-processing methods outperform climatology by a wide margin. On the other hand, in terms of the LogS, clustering-based modeling is clearly superior to local estimation. According to Figure 3b, for most of the forecast horizons, MLP-L underperforms even MLP-R, and at lead times corresponding to 0600 and 1200 UTC, POLR-R also results in positive LogSS with respect to POLR-L.

To enlighten the uncertainty of the investigated verification scores and statistical significance of the score differences, in Figure 4, the CRPSS and LogSS values of some post-processing approaches are accompanied with 95% confidence intervals. According to Figure 4a, the advantage of the best performing POLR-L model over climatology in terms of the mean CRPS is clearly significant for all lead times but 204 and 228 h. Regarding the LogS (Figure 4b), POLR-L significantly outperforms MLP-L for all lead times, whereas the difference between POLR-L and the best performing POLR-C is significant at a 5% level only at forecast horizons corresponding to 0600 UTC.

Furthermore, PIT histograms of Figure 5 also demonstrate the positive effect of post-processing and the

TABLE 1 Overall mean CRPS/LogS of post-processed and climatological visibility forecasts for calendar year 2021 as proportion of the mean CRPS/LogS of the raw ECMWF ensemble.

	MLP-L	MLP-C	MLP-R	POLR-L	POLR-C	POLR-R	Climatology
CRPS	70.65%	70.75%	72.94%	68.09%	69.60%	72.52%	75.60%
LogS	45.31%	44.10%	44.88%	44.28%	43.45%	44.44%	66.05%

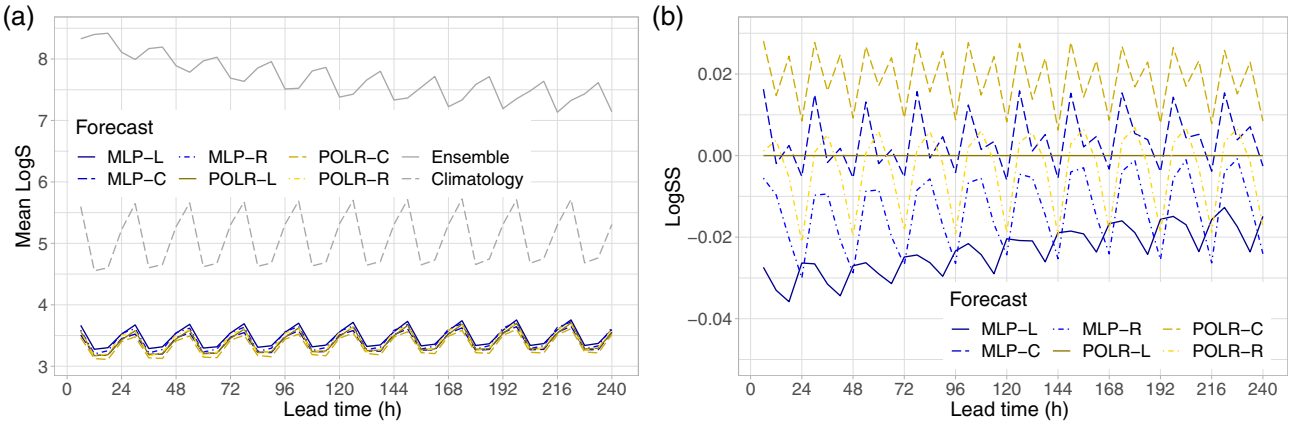

FIGURE 3 Mean LogS of post-processed, raw, and climatological visibility forecasts for calendar year 2021 (a) and LogSS of post-processed forecasts with respect to the POLR-L model (b) as functions of the lead time.

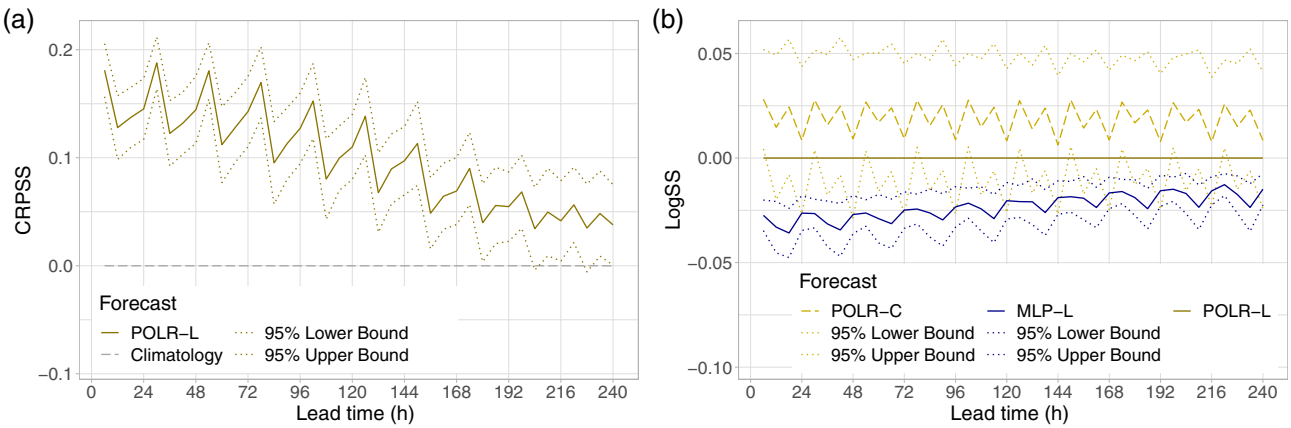

FIGURE 4 CRPSS of POLR-L post-processed visibility forecasts for calendar year 2021 with respect to climatology (a) and LogSS of POLR-C and MLP-L approaches with respect to the POLR-L model (b) together with 95% confidence intervals as functions of the lead time.

superiority of climatology over the raw visibility ensemble forecasts. The histograms of the latter are rather U-shaped with a more and more pronounced bias with the increase of the forecast horizon. All other forecasts result in PIT values that are much closer to the desired standard uniform distribution; however, climatology still exhibits a small underdispersion, whereas MLP post-processed predictions are slightly overdispersive. For all calibrated forecasts, the moment-based α^0_{1234} test (Knüppel, 2015) rejects uniformity at a 5% level of significance for all 40 investigated lead times. Nevertheless, a possible ranking of the competing post-processing approaches in terms of goodness of fit of PIT can be obtained from Table 2 reporting the mean p-values of the α^0_{1234} tests over all considered lead times (the larger the better). Note that for raw and climatological forecasts, the PIT values are concentrated around 52 and 31 bins (ensemble size +1), respectively. Hence, these PIT histograms fall back to the corresponding verification rank histograms (Wilks, 2019, section 9.7.1).

The improved calibration of post-processed and climatological forecasts can also be observed in Figure 6a,

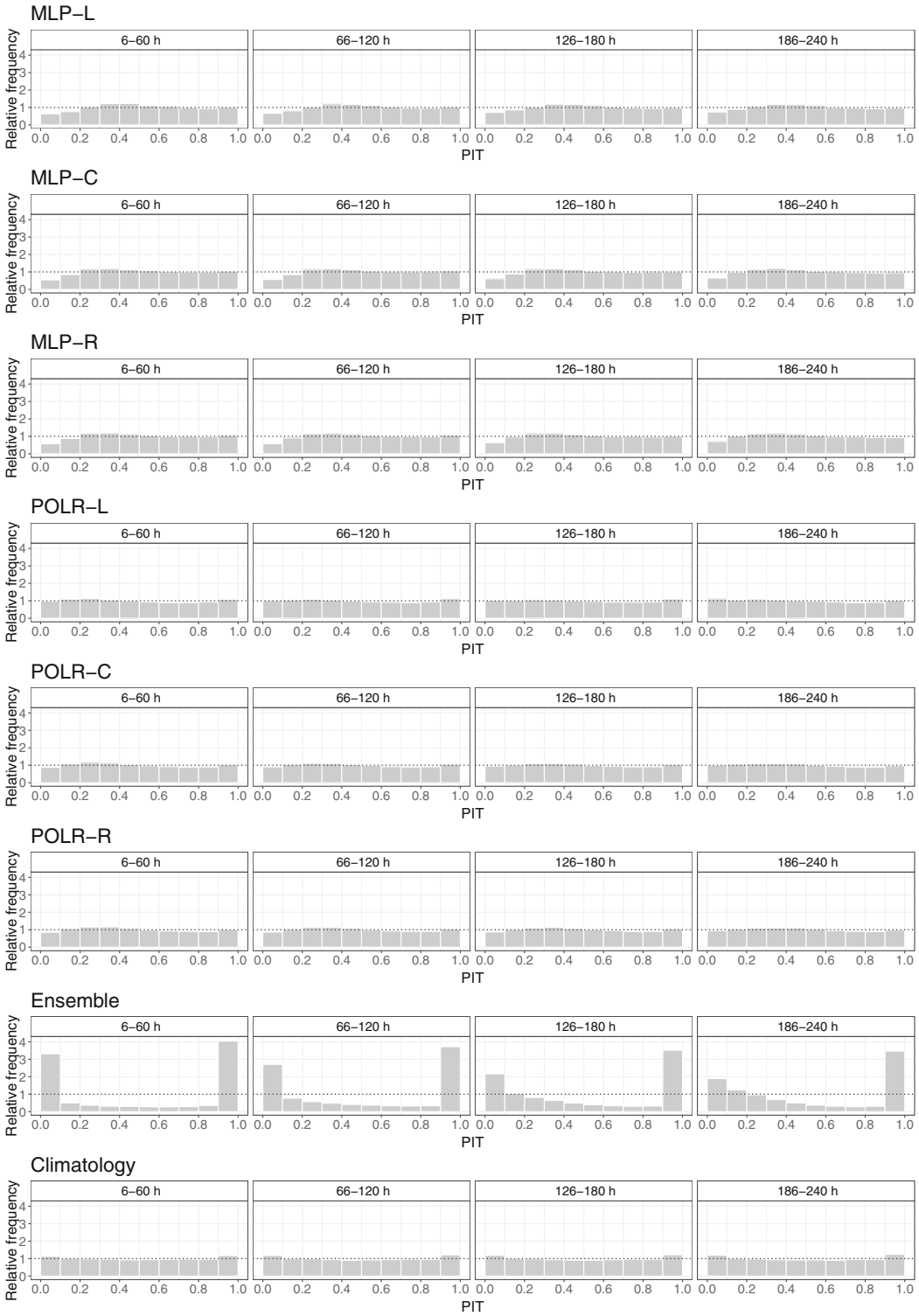

FIGURE 5 PIT histograms of post-processed, climatological, and raw visibility forecasts for calendar year 2021 for lead times 6–60, 66–120, 126–180, and 186–240 h.

TABLE 2 Overall mean p-values of the α^0_{1234} tests for uniformity of the PIT values for calendar year 2021.

MLP-L	MLP-C	MLP-R	POLR-L	POLR-C	POLR-R
1.98×10^{-15}	6.54×10^{-14}	1.15×10^{-7}	4.31×10^{-5}	1.54×10^{-4}	2.00×10^{-5}

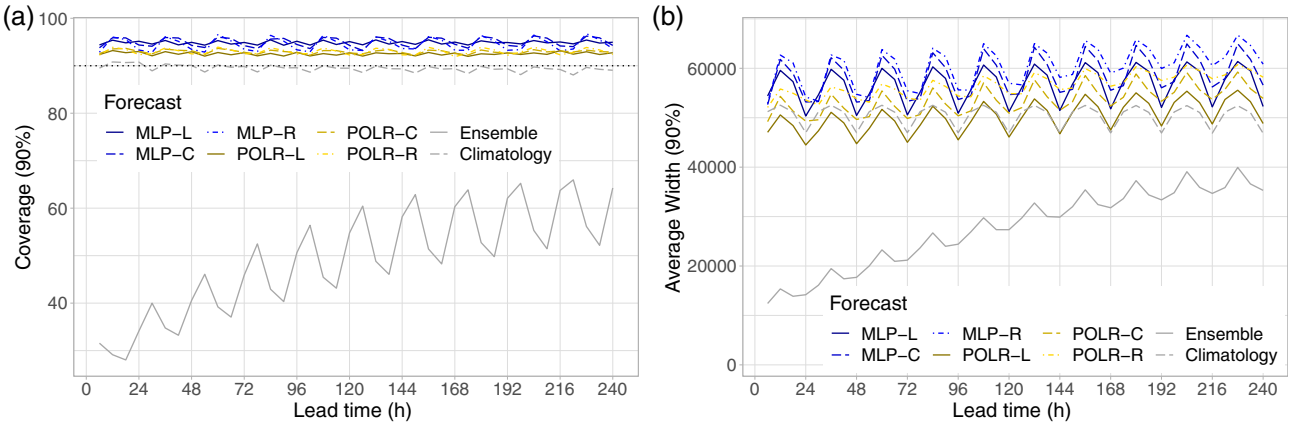

FIGURE 6 Coverage (a) and average widths (b) of 90% central prediction intervals of raw and post-processed visibility forecasts for calendar year 2021 as functions of the lead time. In panel (a) the ideal coverage is indicated by the horizontal dotted line.

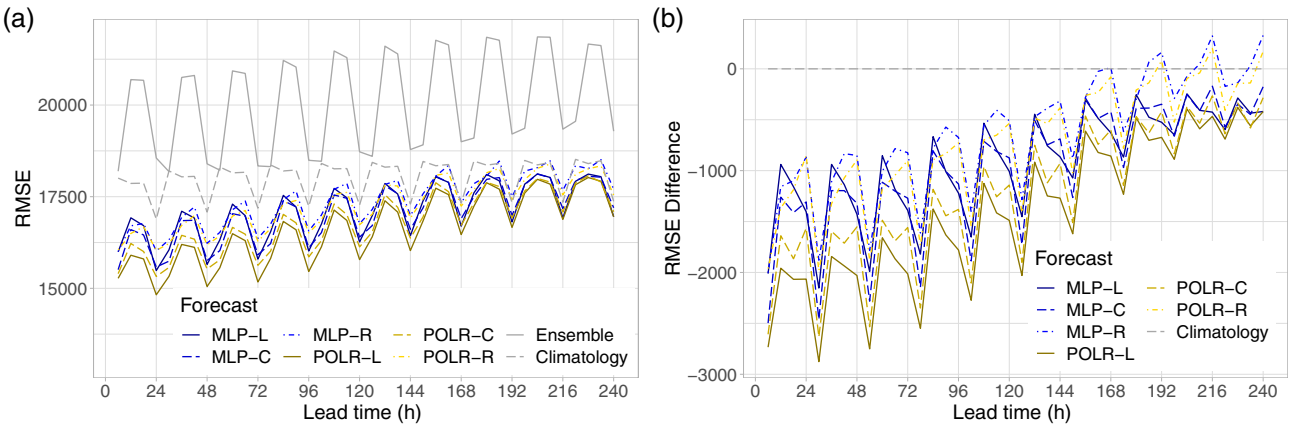

FIGURE 7 RMSE of the mean forecasts for calendar year 2021 (a) and difference in RMSE from climatology (b) as functions of the lead time.

showing the coverage of 90% central prediction intervals. For the raw forecasts, this coverage ranges from 28% to 66% and shows an increasing trend with a clear diurnal cycle. For post-processed and climatological forecasts, the diurnal cycle is far less pronounced and one cannot observe dependence on the forecast horizon. The coverage of climatology is almost perfect, whereas post-processed forecasts result in coverages slightly above 90%. For the POLR models, the mean absolute deviations over all lead times from the desired 90% are 2.53% (POLR-L), 2.96% (POLR-C), and 2.99% (POLR-R), while for the MLP approaches, one gets 4.86% (MLP-R), 4.98% (MLP-C), and 4.52% (MLP-R). Note that for climatological forecasts, this mean absolute deviation is just 0.69%. Naturally, the price for the better calibration of post-processed and climatological forecasts has to be paid in the loss of sharpness. As depicted in Figure 6b, the raw ECMWF visibility ensemble results in far the narrowest 90% central prediction intervals. This is a rather common phenomenon stemming from the highly underdispersive character of the raw forecasts, which in our case improves with the increase of the lead time (see Figure 5). Post-processing usually increases the spread of the predictions, which directly leads to deterioration in sharpness. Furthermore, one should also note that in general, the average width values of the different forecasts are fairly consistent with the corresponding coverages.

Finally, according to Figure 7a, compared with the raw ensemble, all POLR and MLP models substantially improve the accuracy of the mean forecast, whereas the advantage of climatology is far less pronounced, especially for shorter forecast horizons. Note that the ranking of the different forecasts with respect to the RMSE of the mean (Figure 7b) is completely in line with the ordering based on the mean CRPS (Figure 2b), and the POLR-L approach again outperforms the competitors for all lead times.

Besides the 30-day reference climatology, the predictive performance of climatological forecasts based on a

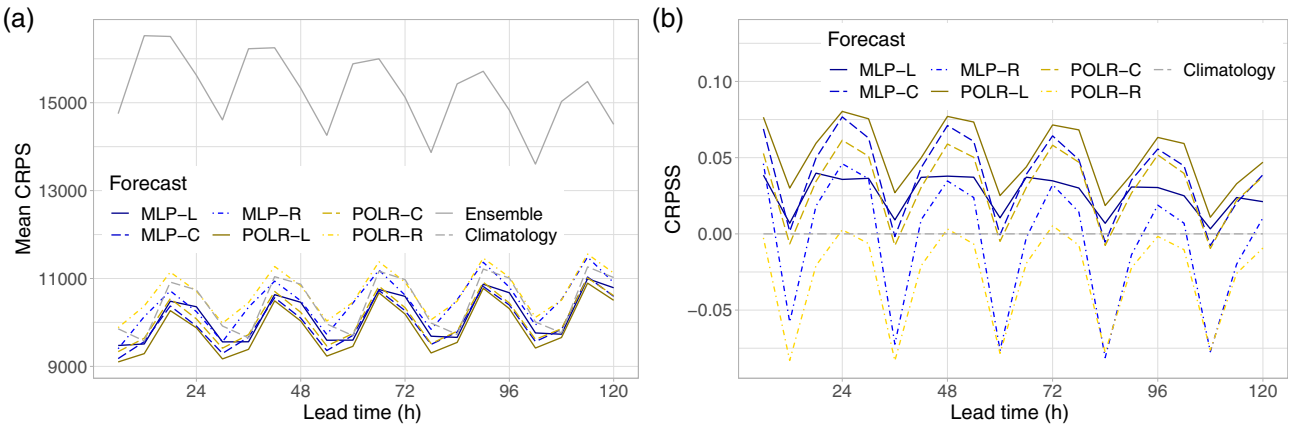

FIGURE 8 Mean CRPS of post-processed, raw, and climatological EUPPBench visibility forecasts for calendar year 2018 (a) and CRPSS of post-processed forecasts with respect to climatology (b) as functions of the lead time.

350-day rolling training period was also tested. However, such a long training period is unlikely to capture the seasonal variations in visibility. In terms of the mean CRPS and RMSE of the mean, this forecast is far behind the various MLP and POLR models, its mean absolute deviation in 90% coverage from the required value is 3.90%, combined with an average width similar to the MLP approaches. The only score where the performance of the 350-day climatology is reasonable is the LogS, at 192, 216, and 240 h it can compete with the regional versions of both studied post-processing techniques.

4.3 | Calibration of EUPPBench visibility ensemble forecasts

As mentioned, in post-processing of the 52-member EUPPBench benchmark visibility forecasts as an additional predictor, we consider the normalized high-resolution forecast and investigate the performance of POLR and MLP models with an extended feature vector

$$\left(\widetilde{f}_{HRES},\widetilde{f}_{CTRL},\overline{f}_{ENS},s^2,p_1,p_2,p_3,\beta_1,\beta_2\right)^{\top}.$$

This extension is in line with the recommendation of the ECMWF Directorate (2012), suggesting that "the HRES and ENS forecasts should, wherever possible, be used together to provide the most detailed description of future weather and the associated uncertainties". Similar to the previous case study, for the local, semilocal, and regional POLR models, the coefficients of $\widetilde{f}_{HRES}$, $\widetilde{f}_{CTRL}$, and $\overline{f}_{ENS}$ are forced to be non-negative. Furthermore, the performance of the various forecasts is compared using forecast–observation pairs for the calendar year 2018, and following exactly the same selection procedure as in Section 4.2 (based on testing training intervals of different durations), for the reference climatology now a 40-day rolling training period is chosen. This training period length is still short enough to allow climatological forecasts to adapt to seasonal variations in visibility. Finally, we tested again the skill of those climatological forecasts, which are based on the same 350-day training period as the post-processing approaches.

In line with the case study of Section 4.2, both climatological and post-processed forecasts (MLP and POLR models) result in far lower mean CRPS values than the raw EUPPBench ensemble for all lead times (see Figure 8a); however, now the ranking of the competing predictions is different. As depicted in Figure 8b, climatology performs surprisingly well, clearly outperformed for all forecast horizons only by the locally trained POLR and MLP approaches. Except for lead times corresponding to 1200 UTC, POLR-C and MLP-C also yield positive CRPSS with respect to climatology and are superior to MLP-L, whereas POLR-R and MLP-R follow a similar pattern but perform much worse than the corresponding semilocal models. The ranking of forecasts based on Figure 8b (POLR-L–MLP-C–POLR-C–MLP-L–Climatology–MLP-R–POLR-R) is also confirmed by Table 3 reporting the overall improvement in mean scores with respect to the raw EUPPBench visibility ensemble forecasts.

In terms of the mean LogS, climatology is again far behind post-processed forecasts; Figure 9a is rather similar to Figure 3a. According to Figure 9b, semilocal models clearly outperform their local counterparts and except for lead times corresponding to 1200 UTC, even the regional approaches are superior to the local ones. From the studied calibration methods, POLR-C results in the lowest mean LogS, closely followed by the MLP-C, whereas MLP-L shows the poorest performance, see also Table 3.

TABLE 3 Overall mean CRPS/LogS of post-processed and climatological EUPPBench visibility forecasts for calendar year 2018 as proportion of the mean CRPS/LogS of the raw ECMWF ensemble.

	MLP-L	MLP-C	MLP-R	POLR-L	POLR-C	POLR-R	Climatology
CRPS	66.35%	65.53%	68.47%	64.67%	66.05%	69.96%	68.19%
LogS	45.48%	43.91%	44.45%	44.86%	43.66%	44.55%	64.39%

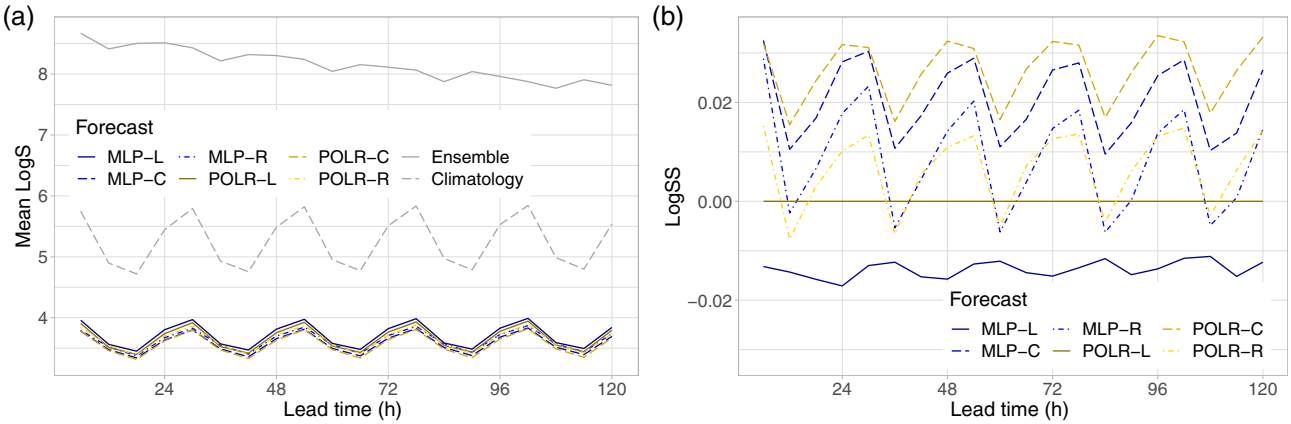

FIGURE 9 Mean LogS of post-processed, raw, and climatological EUPPBench visibility forecasts for calendar year 2018 (a) and LogSS of post-processed forecasts with respect to climatology (b) as functions of the lead time.

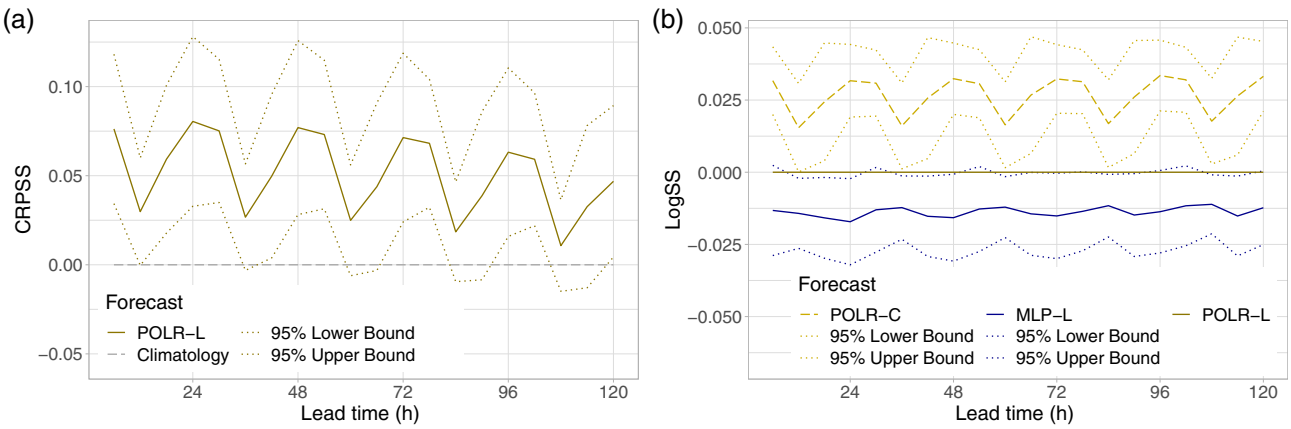

FIGURE 10 CRPSS of POLR-L post-processed EUPPBench visibility forecasts for calendar year 2018 with respect to climatology (a) and LogSS of POLR-C and MLP-L approaches with respect to the POLR-L model (b) together with 95% confidence intervals as functions of the lead time.

The good performance of climatology with regard to the mean CRPS is also confirmed by Figure 10a, as at forecast horizons 36, 60, 66, 84, 90, 108, and 114 h even the best performing POLR-L model fails to result in significantly positive CRPSS at a 5% level. Furthermore, in contrast to the case study of Section 4.2, the LogSS of POLR-C with respect to POLR-L is significantly positive for all investigated lead times, whereas the 95% upper bound for the LogSS of the least performing MLP-L almost coincides with the reference line. This means a barely significant difference between the MLP-L approach and POLR-L.

The PIT histograms of Figure 11 tell us the same story about calibration as the corresponding panels of Figure 5. Raw EUPPBench forecasts are highly underdispersive with a small bias increasing with the forecast horizon, whereas climatology and all post-processing approaches display rather flat histograms indicating improved calibration.

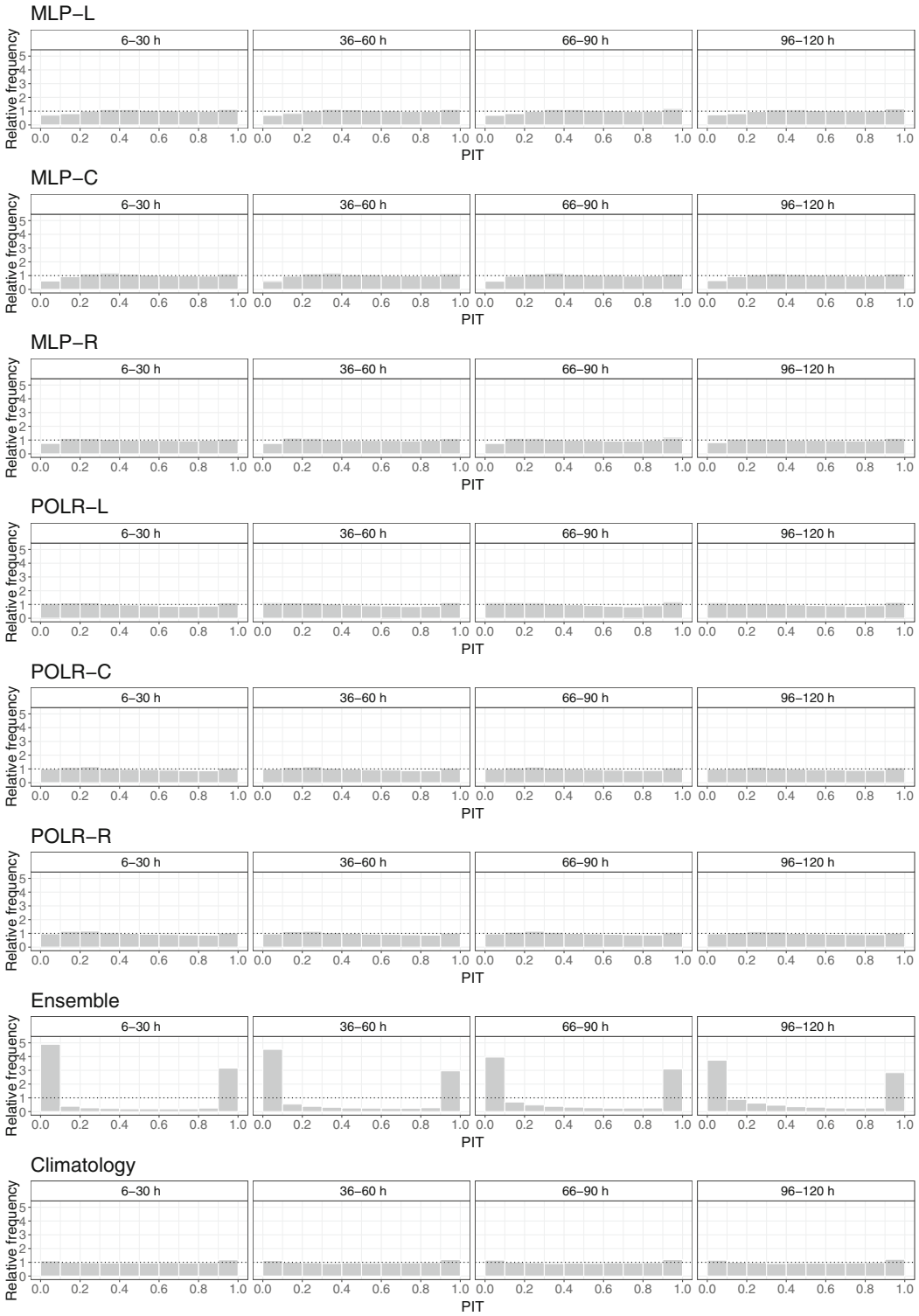

FIGURE 11 PIT histograms of post-processed, climatological, and raw visibility EUPPBench forecasts for calendar year 2018 for lead times 6–30, 36–60, 66–90, and 96–120 h.

TABLE 4 Overall mean p-values of the α^0_{1234} tests for uniformity of the PIT values for calendar year 2018.

MLP-L	MLP-C	MLP-R	POLR-L	POLR-C	POLR-R
2.19×10^{-13}	1.80×10^{-14}	2.82×10^{-4}	5.55×10^{-18}	1.15×10^{-6}	4.57×10^{-4}

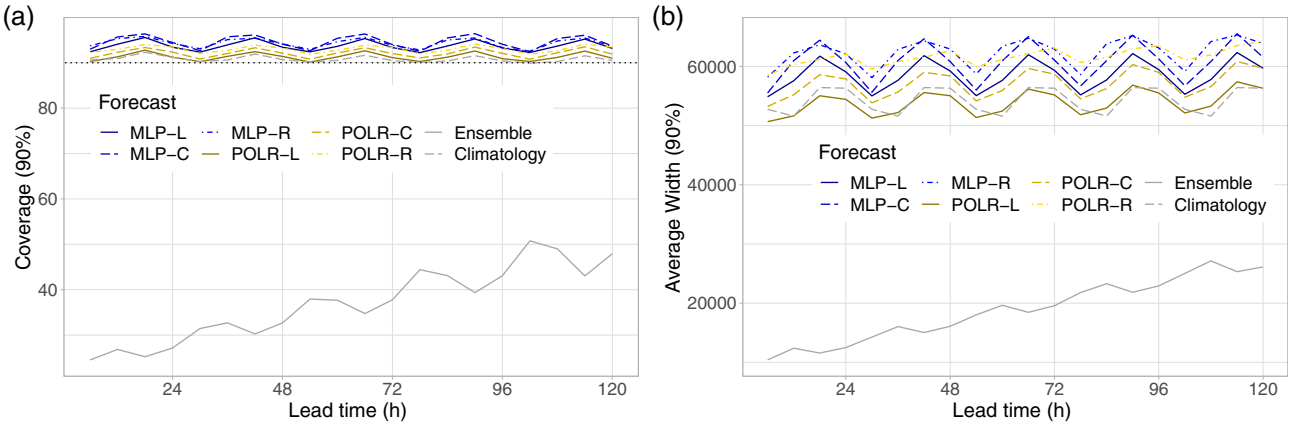

FIGURE 12 Coverage (a) and average widths (b) of 90% central prediction intervals of raw and post-processed EUPPBench visibility forecasts for calendar year 2018 as functions of the lead time. In panel (a) the ideal coverage is indicated by the horizontal dotted line.

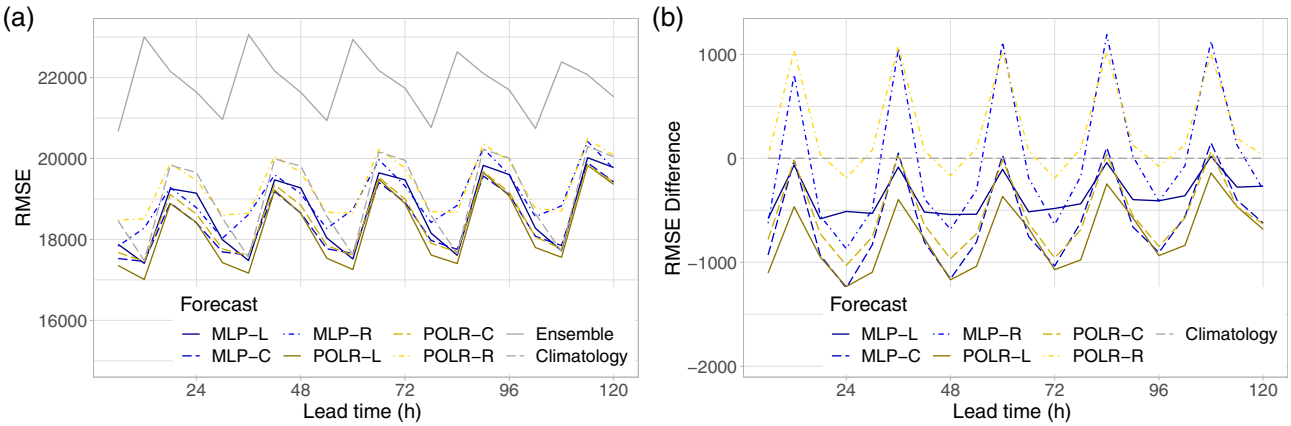

FIGURE 13 RMSE of the mean EUPPBench visibility forecasts for calendar year 2018 (a) and difference in RMSE from climatology (b) as functions of the lead time.

Nevertheless, similar to the previous case study, MLP predictions exhibit a tiny overdispersion and climatological forecasts are slightly underdispersive. The α^0_{1234} test again rejects uniformity for all calibrated forecasts for all lead times; however, now the two regional approaches result in the highest overall mean p-values, see Table 4.

According to Figure 12a, with the increase of the forecast horizon, the coverage of 90% central prediction intervals of the raw EUPPBench forecasts increases from 22% to 51%, and shows again a clear diurnal cycle. This diurnal cycle is also preserved in the coverage values of climatological and post-processed forecasts; however, these curves do not display any further dependence on the forecast horizon and are just slightly above 90%. Similar to Figure 6a, the best coverage again corresponds to climatological forecasts with a mean absolute deviation from the nominal value of 0.82%, followed by the POLR models (POLR-L: 1.32%; POLR-C: 2.12%; POLR-R: 3.03%) and MLP approaches (MLP-L: 3.69%; MLP-C: 4.59%; MLP-R: 4.35%). Again, the average widths of 90% central prediction intervals depicted in Figure 12b are nicely in line with the corresponding coverage values, with the POLR-L approach providing the sharpest post-processed forecasts.

Furthermore, in line with the case study of Section 4.2, all post-processing methods result in a considerable decrease in the RMSE of the mean forecast (see Figure 13a). Moreover, according to Figure 13b, for the EUPPBench data, climatology is fully able to catch up with the POLR and MLP models, and for forecast horizons corresponding to 1200 UTC, it definitely outperforms their local versions.

Finally, 350-day climatology performs well again only in terms of the LogS; however, in this case, even with respect to this particular score, it is still slightly behind the post-processed forecasts for all forecast horizons. The overall mean CRPS of 350-day climatological forecasts is 73.67% of the mean CRPS of the raw EUPPBench ensemble (compared with the corresponding values of Table 3), its mean absolute deviation in 90% coverage from the nominal value is 3.19%, and the sharpness of the corresponding central prediction

interval is between the sharpnesses of locally and semilocally trained MLP models.

5 | CONCLUSIONS

We investigate the predictive performance of the proportional odds logistic regression and multilayer perceptron neural network classifiers for statistical post-processing of visibility ensemble forecasts in the situation when observations are reported in discrete values following the suggestions of the WMO. In two case studies, the forecast skill of the proposed calibration methods is tested on two different datasets of ECMWF visibility ensemble forecasts covering two overlapping geographical regions and two different time domains. First, we consider the 51-member operational forecasts for 30 locations in Central Europe with a forecast horizon of 240 h and a time step of 6 h for calendar years 2020 and 2021. The second dataset is part of the EUPPBench benchmark data, where the ECMWF high-resolution forecast is also included. It contains ensemble forecasts and validating observations for 42 stations in Germany and France for calendar years 2017–2018 with a maximal lead time of 120 h and temporal resolution of 6 h. Note that both in terms of the mean CRPS and the RMSE of the ensemble mean, the EUPPBench forecasts underperform the more recent 51-member ECMWF predictions (compare Figures 8a and 13a and the corresponding parts of Figures 2a and 7a, respectively). This might be due to the different ensemble domains; however, it is more likely related to the perpetual improvement in the ECMWF IFS (6 model upgrades between 2017 and 2021). Local, semilocal, and regional training is tested for both post-processing approaches, which are based on 350-day rolling training periods. As reference, we consider raw and climatological visibility forecasts where for the latter short rolling past time intervals are chosen in order to capture seasonal variations in visibility observations.

In general, climatological and post-processed forecasts outperform the raw ensemble in terms of the mean CRPS, the mean LogS, and the RMSE of the mean forecast over the verification data by a wide margin, but the difference decreases with the increase of the forecast horizon. They result in better coverage but wider central prediction intervals than the raw ECMWF visibility forecasts and the corresponding PIT histograms are much closer to the uniform distribution than those of the raw ensemble. POLR models are superior to their MLP counterparts, which is in line with the findings by Baran et al. (2021) in the case of TCC forecasts; however, here the performance of climatology is comparable with the skill of post-processed forecasts in all investigated scores except for the mean LogS. From the competing post-processing methods, the locally trained POLR model shows the best overall performance, whereas for the machine learning-based approaches the semilocal training is often superior to the local one.

Our case studies confirm that the state-of-the-art classification methods provide efficient tools for discrete calibration of visibility forecasts. Hence, post-processing might be applied to improve the generally extremely low forecast skill of the raw visibility predictions. However, in this first study, we test merely the general performance of the POLR and MLP approaches and neither analyze the dependence of the improvement in skill on the time of the year nor compare the competing forecasts under low visibility conditions using, for example, threshold weighted scoring rules (Gneiting & Ranjan, 2011). These additional investigations would provide a deeper understanding of visibility ensemble calibration, but obviously might also reveal possible weaknesses of the proposed methodology.

Nevertheless, the results of the current work open several further avenues for future research. First, here as inputs of our post-processing models, we have used only features derived from the raw ECMWF visibility forecasts. However, as recent studies demonstrate (see, e.g., Rasp & Lerch, 2018; Baran et al., 2021; Schultz & Lerch, 2022), the incorporation of additional predictors, especially in the case of machine learning-based approaches, where the extension of the input feature set is rather straightforward, might significantly improve the forecast skill. As natural candidates, one can consider further visibility forecasts such as the ones produced by the Copernicus Atmospheric Monitoring Service, or predictions of other weather variables that might affect visibility. Furthermore, as several SYNOP stations report visibility in 10-m steps, which can be considered almost as continuous, one might try to develop techniques for estimating continuous predictive distributions for this weather quantity. The most straightforward approach is the interpolation technique of Scheuerer et al. (2020), where after discrete post-processing of visibility, for example, with the help of one of the classification methods presented here, the obtained predictive probability mass function is interpolated to a full predictive cumulative distribution function. One might also focus on the direct development of parametric models, where the predictive probability law should have a non-negative support with a point mass at the maximal reported visibility, and multimodality might also be required. The beta distribution-based BMA approach of Chmielecki and Raftery (2011) belongs to this category, which can serve as a reference method in future studies on continuous post-processing of visibility. Finally, so far, we have focused on univariate forecasts for a single location and forecast horizon. However, in the last decade, a wide range of multivariate post-processing techniques have been

developed, which are able to restore dependence structures lost during the univariate calibration, for an overview see, for example, Lerch et al. (2020) or Lakatos et al. (2023). The investigation of spatially and/or temporally consistent calibrated visibility forecasts might be another interesting direction of our future work.

AUTHOR CONTRIBUTIONS

Sándor Baran: Conceptualization (lead); data curation (supporting); formal analysis (equal); funding acquisition (lead); investigation (equal); methodology (lead); software (lead); validation (equal); visualization (equal); writing – original draft (lead); writing – review and editing (equal). **Mária Lakatos:** Conceptualization (supporting); data curation (lead); formal analysis (equal); investigation (equal); methodology (supporting); software (supporting); validation (equal); visualization (equal); writing – original draft (supporting); writing – review and editing (equal).

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of the ÚNKP-22-3 New National Excellence Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund and of the National Research, Development and Innovation Office under Grant No. K142849. They are also indebted to Zied Ben Bouallègue for providing the ECMWF visibility data. Last but not least, the authors thank the three anonymous reviewers for their valuable suggestions for improving the manuscript.

FUNDING INFORMATION

Hungarian National Research, Development and Innovation Office, Grant/Award Number: K142849; ÚNKP-22-3 New National Excellence Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund.

ORCID

Sándor Baran https://orcid.org/0000-0003-1035-004X

REFERENCES

Baran, Á., Lerch, S., El Ayari, M. & Baran, S. (2021) Machine learning for total cloud cover prediction. *Neural Computing and Applications*, 33, 2605–2620.

Baran, S., Baran, Á., Pappenberger, F. & Ben Bouallègue, Z. (2020) Statistical post-processing of heat index ensemble forecasts: is there a royal road? *Quarterly Journal of the Royal Meteorological Society*, 146, 3416–3434.

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

Breiman, L. (2001) Random forests. *Machine Learning*, 45, 5–32.

Bremnes, J.B. (2019) Constrained quantile regression splines for ensemble postprocessing. *Monthly Weather Review*, 147, 1769–1780.

Bremnes, J.B. (2020) Ensemble postprocessing using quantile function regression based on neural networks and Bernstein polynomials. *Monthly Weather Review*, 148, 403–414.

Buizza, R. (2018a) Introduction to the special issue on "25 years of ensemble forecasting". *Quarterly Journal of the Royal Meteorological Society*, 145, 1–11.

Buizza, R. (2018b) Ensemble forecasting and the need for calibration. In: Vannitsem, S., Wilks, D.S. & Messner, J.W. (Eds.) *Statistical postprocessing of ensemble forecasts*. Amsterdam: Elsevier, pp. 15–48.

Buizza, R., Houtekamer, P.L., 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.

Chmielecki, R.M. & Raftery, A.E. (2011) Probabilistic visibility forecasting using Bayesian model averaging. *Monthly Weather Review*, 139, 1626–1636.

Dabernig, M., Mayr, G.J., Messner, J.W. & Zeileis, A. (2017) Spatial ensemble post-processing with standardized anomalies. *Quarterly Journal of the Royal Meteorological Society*, 143, 909–916.

Demaeyer, J., Bhend, J., Lerch, S., Primo, C., Van Schaeybroeck, B., Atencia, A. et al. (2023) The EUPPBench postprocessing benchmark dataset v1.0. *Earth System Science Data*, 15, 2635–2653.

Dietz, S.J., Kneringer, P., Mayr, G.J. & Zeileis, A. (2019) Low-visibility forecasts for different flight planning horizons using tree-based boosting models. *Advances in Statistical Climatology, Meteorology and Oceanography*, 5, 101–114.

ECMWF. (2021) *IFS documentation CY47R3—part IV physical processes*. Reading: ECMWF. Available from: https://doi.org/10.21957/eyrpir4vj [Accessed 24th September 2023].

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

Friederichs, P. & Hense, A. (2007) Statistical downscaling of extreme precipitation events using censored quantile regression. *Monthly Weather Review*, 135, 2365–2378.

Friedman, J.H. (2001) Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29, 1189–1232.

Ghazvinian, M., Zhang, Y., Seo, D-J., He, M. and Fernando, N. (2021) A novel hybrid artificial neural network—parametric scheme for postprocessing medium-range precipitation forecasts. *Advances in Water Resources* 151, 103907. https://doi.org/10.1012/j.advwatres.2021.103907

Gneiting, T. (2014) Calibration of medium-range weather forecasts. ECMWF Technical Memorandum No. 719 Available at: http://www.ecmwf.int/sites/default/files/elibrary/2014/9607-calibration-medium-range-weather-forecasts.pdf [Accessed 24th September 2023].

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

Gneiting, T. & Raftery, A.E. (2005) Weather forecasting with ensemble methods. *Science*, 310, 248–249.

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

Gneiting, T., Raftery, A.E., Westveld, A.H. & Goldman, T. (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133, 1098–1118.

Gneiting, T. & Ranjan, R. (2011) Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29, 411–422.

Gneiting, T. & Ranjan, R. (2013) Combining predictive distributions. *Electronic Journal of Statistics*, 7, 1747–1782.

Good, I.J. (1952) Rational decisions. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, 14, 107–114.

Goodfellow, I., Bengio, Y. & Courville, A. (2016) *Deep learning*. Cambridge: MIT Press.

Gultepe, I., Müller, M.D. & Boybeyi, Z. (2006) A new visibility parameterization for warm-fog applications in numerical weather prediction models. *Journal of Applied Meteorology and Climatology*, 45, 1469–1480.

Haiden, T., Janousek, M., Vitart, F., Ben Bouallègue, Z., Ferranti, L., Prates, F. et al. (2021) Evaluation of ECMWF forecasts, including the 2021 upgrade. ECMWF Technical Memorandum No. 884 https://doi.org/10.21957/90pgicjk4 [Accessed 24th September 2023].

Hemri, S., Haiden, T. & Pappenberger, F. (2016) Discrete postprocessing of total cloud cover ensemble forecasts. *Monthly Weather Review*, 144, 2565–2577.

Hemri, S., Scheuerer, M., Pappenberger, F., Bogner, K. & Haiden, T. (2014) Trends in the predictive performance of raw ensemble weather forecasts. *Geophysical Research Letters*, 41, 9197–9205.

Izenman, A.J. (2008) *Modern multivariate statistical techniques. regression, classification and manifold learning*. New York: Springer.

Knüppel, M. (2015) Evaluating the calibration of multi-step-ahead density forecasts using raw moments. *Journal of Business & Economic Statistics*, 33, 270–281.

Lakatos, M., Lerch, S., Hemri, S. & Baran, S. (2023) Comparison of multivariate post-processing methods using global ECMWF ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 149, 856–877.

Lerch, S. & Baran, S. (2017) Similarity-based semi-local estimation of EMOS models. *Journal of the Royal Statistical Society, Series C*, 66, 29–51.

Lerch, S., Baran, S., Möller, A., Groß, J., Schefzik, R., Hemri, S. et al. (2020) Simulation-based comparison of multivariate ensemble post-processing methods. *Nonlinear Processes in Geophysics*, 27, 349–371.

Marzban, C., Leyton, S. & Colman, B. (2007) Ceiling and visibility forecasts via neural networks. *Weather Forecasting*, 22, 466–479.

McCullagh, P. (1980) Regression model for ordinal data (with discussion). *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 42, 243–268.

Molteni, F., Buizza, R., Palmer, T.N. & Petroliagis, T. (1996) The ECMWF ensemble prediction system: methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, 122, 73–119.

Murphy, A.H. (1973) Hedging and skill scores for probability forecasts. *Journal of Applied Meteorology*, 12, 215–223.

Pahlavan, R., Moradi, M., Tajbakhsh, S., Azadi, M. & Rahnama, M. (2021) Fog probabilistic forecasting using an ensemble prediction system at six airports in Iran for 10 fog events. *Meteorological Applications*, 28, e2033. https://doi.org/10.1002/met.2033

Parde, A.N., Ghude, S.D., Dhangar, N.G., Lonkar, P., Wagh, S., Govardhan, G. et al. (2022) Operational probabilistic fog prediction based on ensemble forecast system: a decision support system for fog. *Atmosphere*, 13, paper 1608.

Politis, D.N. & Romano, J.P. (1994) The stationary bootstrap. *Journal of the American Statistical Association*, 89, 1303–1313.

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

Rasp, S. & Lerch, S. (2018) Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, 146, 3885–3900.

Ryerson, W.R. & Hacker, J.P. (2018) A nonparametric ensemble postprocessing approach for short-range visibility predictions in data-sparse areas. *Weather Forecasting*, 33, 835–855.

Scheuerer, M., Switanek, M.B., Worsnop, R.P. & Hamill, T.M. (2020) Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over California. *Monthly Weather Review*, 148, 3489–3506.

Schultz, B. & Lerch, S. (2022) Machine learning methods for postprocessing ensemble forecasts of wind gusts: a systematic comparison. *Monthly Weather Review*, 150, 235–257.

Szabó, M., Gascón, E. & Baran, S. (2023) Parametric post-processing of dual-resolution precipitation forecasts. *Weather Forecasting*, 38, 1313–1322.

Taillardat, M., Mestre, O., Zamo, M. & Naveau, P. (2016) Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, 144, 2375–2393.

Thorarinsdottir, T.L. & Gneiting, T. (2010) Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. *Statistical Society, Series A (Statistics in Society)*, 173, 371–388.

Vannitsem, S., Bremnes, J.B., Demaeyer, J., Evans, G.R., Flowerdew, J., Hemri, S. et al. (2021) Statistical postprocessing for weather forecasts – review, challenges and avenues in a big data world. *Bulletin of the American Meteorological Society*, 102, E681–E699.

Venables, W.N. & Ripley, B.D. (2002) *Modern applied statistics with S*, 4th edition. New York: Springer.

Wagh, S., Kulkarni, R., Lonkar, P., Parde, A.N., Dhangar, N.G., Govardhan, G. et al. (2023) Development of visibility equation based on fog microphysical observations and its verification using the WRF model. *Modeling Earth Systems and Environment*, 9, 195–211.

Wilks, D.S. (2018) Univariate ensemble postprocessing. In: Vannitsem, S., Wilks, D.S. & Messner, J.W. (Eds.) *Statistical postprocessing of ensemble forecasts*. Amsterdam: Elsevier, pp. 49–89.

Wilks, D.S. (2019) *Statistical methods in the atmospheric sciences*, 4th edition. Amsterdam: Elsevier.

World Meteorological Organization. (1992) *International meteorological vocabulary (WMO-No.182)*. Geneva: WMO.

World Meteorological Organization. (2018) Guide to instruments and methods of observation. In: *Volume I—Measurement of meteorological variables (WMO-No.8)*. Geneva: WMO.

Zell, A., Mache, N., Hübner, R., Mamier, G., Vogt, M., Schmalzl, M. et al. (1994) SNNS (Stuttgart neural network simulator). In: Skrzypek, J. (Ed.) *Neural network simulation environments. The Kluwer international series in engineering and computer science, vol. 254*. Boston: Springer.

Zhou, B., Du, J., Gultepe, I. & Dimego, G. (2012) Forecast of low visibility and fog from NCEP: current status and efforts. *Pure and Applied Geophysics*, 169, 895–909.

Zhou, B., Du, J., McQueen, J. and Dimego, G. (2009) Ensemble forecast of ceiling, visibility, and fog with NCEP short-range ensemble forecast system (SREF). *Aviation, Range, and Aerospace Meteorology Special Symposium on Weather–Air Traffic Management Integration*, Phoenix, AZ, American Meteorological Society, extended abstract 4.5. Available at: https://ams.confex.com/ams/89annual/techprogram/paper_142255.htm [Accessed 24th September 2023].

How to cite this article: Baran, S., & Lakatos, M. (2023). Statistical post-processing of visibility ensemble forecasts. *Meteorological Applications*, *30*(5), e2157. https://doi.org/10.1002/met.2157