Stochastic Environmental Research and Risk Assessment (2021) 35:215–230
https://doi.org/10.1007/s00477-020-01928-4

ORIGINAL PAPER

Area-covering postprocessing of ensemble precipitation forecasts using topographical and seasonal conditions

Lea Friedli[1] · **David Ginsbourger**[2] · **Jonas Bhend**[3]

Accepted: 29 October 2020 / Published online: 16 November 2020

Abstract

Probabilistic weather forecasts from ensemble systems require statistical postprocessing to yield calibrated and sharp predictive distributions. This paper presents an area-covering postprocessing method for ensemble precipitation predictions. We rely on the ensemble model output statistics (EMOS) approach, which generates probabilistic forecasts with a parametric distribution whose parameters depend on (statistics of) the ensemble prediction. A case study with daily precipitation predictions across Switzerland highlights that postprocessing at observation locations indeed improves high-resolution ensemble forecasts, with 4.5% CRPS reduction on average in the case of a lead time of 1 day. Our main aim is to achieve such an improvement without binding the model to stations, by leveraging topographical covariates. Specifically, regression coefficients are estimated by weighting the training data in relation to the topographical similarity between their station of origin and the prediction location. In our case study, this approach is found to reproduce the performance of the local model without using local historical data for calibration. We further identify that one key difficulty is that postprocessing often degrades the performance of the ensemble forecast during summer and early autumn. To mitigate, we additionally estimate on the training set whether postprocessing at a specific location is expected to improve the prediction. If not, the direct model output is used. This extension reduces the CRPS of the topographical model by up to another 1.7% on average at the price of a slight degradation in calibration. In this case, the highest improvement is achieved for a lead time of 4 days.

Keywords Ensemble postprocessing · Ensemble model output statistics · Precipitation accumulation · Censored logistic regression · Weighted scoring rule estimator · Continuous ranked probability score

1 Introduction

Today, medium-range weather forecasts are generated by Numerical Weather Prediction (NWP) systems which use mathematical (or physics-based, numerical) models of the atmosphere to predict the weather. NWP forecasts are affected by considerable systematic errors due to the imperfect representation of physical processes, limited spatio-temporal resolution, and uncertainties in the initial state of the climate system. This initial condition uncertainty and the fact that the atmosphere is a chaotic system, where small initial errors can grow into large prediction errors, make weather forecasting challenging (Wilks and Vannitsem 2018). Therefore, attention has turned to probabilistic weather forecasting to quantify weather-dependent predictability from day to day.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00477-020-01928-4) contains supplementary material, which is available to authorized users.

✉ Lea Friedli
lea.friedli@unil.ch

David Ginsbourger
ginsbourger@stat.unibe.ch

Jonas Bhend
jonas.bhend@meteoswiss.ch

[1] Institute of Earth Sciences, University of Lausanne, Lausanne, Switzerland

[2] Institute of Mathematical Statistics and Actuarial Science and Oeschger Center for Climate Change Research, University of Bern, Bern, Switzerland

[3] Federal Office of Meteorology and Climatology MeteoSwiss, Zurich, Switzerland

Probabilistic forecasts are generated using different forecasting scenarios (referred to as ensemble members) based on slightly perturbed initial conditions and perturbed physical parameterizations in the NWP system. Unfortunately, such ensemble forecasts are not able to capture the full forecasting uncertainty as it is difficult to represent all sources of error reliably and accurately (Buizza 2018). Hence ensemble forecasts are often biased and over-confident (Wilks 2018). Statistical postprocessing can be used to calibrate ensemble forecasts. A proper postprocessing method providing accurate weather forecasts is fundamental for risk quantification and decision making in industry, agriculture and finance. One example is flood forecasting, where reliable precipitation forecasts are a necessary prerequisite for predicting future streamflow (e.g. Aminyavari and Saghafian 2019).

The objective of statistical postprocessing is to find structure in past forecast-observation pairs to correct systematic errors in future predictions. Various approaches to postprocess ensemble predictions have been developed over the last years, a selection of them is listed for example in Wilks (2018). His overview covers parametric approaches that assume a predictive distribution belonging to a class of probability distributions and nonparametric approaches that avoid such distributional assumptions. For the class of parametric methods, the two main approaches he lists are *Bayesian model averaging* (BMA; Raftery et al. 2005) and *Ensemble model output statistics* (EMOS; Gneiting et al. 2005). The BMA approach generates a predictive probability density function (PDF) using a weighted average of PDFs centred on the single ensemble member forecasts. There are numerous applications of this method, for example in the studies about ensemble precipitation postprocessing of Sloughter et al. (2007) and Schmeits and Kok (2010). The EMOS approach provides a predictive PDF using a parametric distribution whose parameters depend on the ensemble forecast. One of the most frequently used EMOS models is the Nonhomogeneous Gaussian regression approach (NGR; Gneiting et al. 2005). While in a *homogeneous* regression model the variance of the predictive distribution is assumed to be constant, in the *inhomogeneous* approach it is expressed as a function of the ensemble variance. The NGR model, that assumes a Gaussian predictive distribution, has been extensively applied to postprocess temperature forecasts, see for instance Baran (2013) or Hemri et al. (2014). For precipitation, as a non-negative quantity, EMOS with a left-censoring of the forecast distribution at zero is usually applied. A range of parametric distributions have been explored for precipitation postprocessing including the censored Generalized Extreme Value distribution (Scheuerer 2014) , the censored shifted Gamma distribution (Baran and Nemoda 2016), and the censored Logistic distribution (Messner et al. 2014).

We seek to postprocess precipitation forecasts for all of Switzerland. With its complex topography as shown in Fig. 1, Switzerland provides a challenging case for precipitation forecasting. From a climatic perspective, the study area can be classified into different regions for which precipitation characteristics differ quite considerably. First and foremost, the Alps separate the country into a northern and southern part. The Alpine ridge often creates strong contrasts with intense precipitation on the windward slopes and dry conditions downwind. The intensity of such orography-induced precipitation also differs with much more intense precipitation frequently occuring in the south due to the advection of warm, humid air masses from the Mediterranean. The large inner-alpine valleys on the other hand are often shielded from advection of precipitation and thus tend to exhibit much drier climates than the surrounding areas. In addition to pronounced spatial variability, precipitation in Switzerland also exhibits a strong seasonal cycle. While passing fronts and large-scale precipitation dominate in the cold part of the year, in summer and autumn, convection and thunderstorms frequently occur. Convection is usually initiated away from the highest peaks on the northern and southern slope of the Alps and in the Jura mountains in the northwest. During a typical summer day, isolated showers and storms therefore start to appear there and subsequently spread according to the prevailing upper-level winds. Due to its chaotic nature and spatial heterogeneity, predicting convective precipitation is one of the key challenges in weather forecasting.

Starting from an EMOS model, we aim to provide a postprocessing method that enables spatially comprehensive yet locally specific forecasts. To discuss alternatives to achieve this, we distinguish between *global* models that use all available forecast observation pairs to estimate model coefficients, *local* models that use data from a specific location only, and *semi-local* models that use weighting to

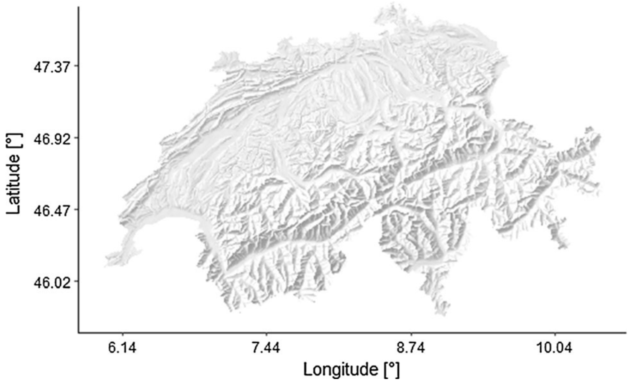


Fig. 1 The relief of the study area with respect to global coordinates (WGS84)

pool information in a suitably defined neighbourhood. The second represents the most locally specific approach and local models therefore generally outperform global models (Thorarinsdottir and Gneiting 2010). It is important to note, however, that by using local models alone, calibration at unobserved sites is not possible. Here we use a semi-local approach to enable local effects without binding the model to the stations.

An ensemble postprocessing algorithm allowing for training data to be weighted individually has first been introduced by Hamill et al. (2008). In their study, they calibrate ensemble precipitation forecasts using logistic regression (for a given threshold) whereby in the fitting procedure, the training data pairs are weighted with respect to the relationship of their ensemble mean and the threshold in question. Another ensemble postprocessing study where the training data is assigned with individual weights has been presented by Lerch and Baran (2018). Using an EMOS approach to postprocess ensemble wind speed predictions, they weight the training data pairs depending on the similarity of their location of origin and the prediction location. Thereby, they measure the similarity of two locations with distances based on the geographical location, the station climatology and the station ensemble characteristics. As an alternative to the distance based weighting approach, Lerch and Baran (2018) suggest to cluster the observational sites based on the same similarity measures and to perform the postprocessing for each cluster individually. The motivation behind the two semi-local approaches of Lerch and Baran (2018) is to solve numerical stability issues of the local model, which requires long training periods since only the data of one station is used for training, they do not aim for an area-covering postprocessing method as this study does. But our study not only has another underlying motivation, we are also using new similarity measures and focus on a rich set of topographical features that are relevant for postprocessing in an area with complex topography such as Switzerland (Fig. 1).

Over the last years, other approaches enabling postprocessing at unobserved sites have been developed. For the interpolation technique, the postprocessing parameters of a local model are spatially interpolated using geostatistical methods. The introduction of geostatistical interpolation for a BMA postprocessing framework has been provided by Kleiber et al. (2011) as *Geostatistical Model Averaging*. Their methodology developed for normally distributed temperature forecasts has been modified by Kleiber et al. (2011) to allow its application to precipitation forecasts. For EMOS models, a geostatistical interpolation procedure has been presented in Scheuerer and Büermann (2013) and extended in Scheuerer and König (2014). Both studies base on locally adaptive postprocessing methods avoiding location-specific parameters in the predictive distribution. Instead, they use local forecast and observation anomalies (with respect to the climatological means) as response and covariates for the regression to get a site-specific postprocessing method. Consequently, they do not have to interpolate the parameters but the anomalies. This method has been modified by Dabernig et al. (2017) using a standardized version of the anomalies and accounting additionally for season-specific characteristics. In contrast, in the study of Khedhaouiria et al. (2019), the regression coefficients are fitted locally and then interpolated geostatistically. In their study, this two-step procedure of interpolating the coefficients is compared with an integrated area-covering postprocessing method relying on Vector Generalized Additive Models (VGAM) with spatial covariates. A comprehensive intercomparison of all the proposed approaches for area-covering postprocessing is beyond the scope of this study, instead we discuss avenues for future research in Sect. 5.

In addition to the area-covering applicability, the seasonal characteristics of the considered weather quantity present a challenge for the EMOS models. In this context, the temporal selection of the training data plays an important role. The already mentioned studies of Scheuerer and Büermann (2013) and Scheuerer and König (2014) have been using a rolling training period of several tenths of days. This means that the model has to be refitted every day and that only part of the training data can be used for the fitting. In the study of Dabernig et al. (2017) which is also based on anomalies, a full seasonal climatology is fitted and subtracted such that the daily fitting can be avoided and the whole training data set can be used during the regression. In the work of Khedhaouiria et al. (2019), the post-processing parameters are also fitted for every day individually. They account for seasonality by adding sine and cosine functions of seasonal covariates. We have tested similar approaches for our case study and used different choices of training periods, additional regression covariables and a weighting of the training data to account for the seasonality. Our case study highlights that a key difficulty of postprocessing ensemble precipitation forecasts lies in summer and early autumn, when in many places postprocessing leads to a degradation of the forecast quality, be it using a local or global approach. The presented seasonal approaches account for seasonal variations in the postprocessing but do not enable its renouncement. For this reason, we introduce a novel approach referred to as *Pretest*. The later first evaluates whether a postprocessing at a given station in a given month is expected to improve forecast performance. A comparison of the performances shows that for our case study the Pretest approach performs best (see supplementary material for details).

In summary, the aim of this paper is to provide calibrated and sharp precipitation predictions for the entire area of Switzerland by postprocessing ensemble forecasts. The postprocessing model should account for seasonal specificities and while it is developed at observation locations, it should also be applicable at unobserved locations and thereby allow to produce area-covering forecasts. The remainder of this paper is organized as follows: Sect. 2 introduces the data, the notation and the verification techniques. The elaboration of the postprocessing model is presented in Sect. 3. In Sect. 4 we show the results of the external model validation; a discussion of the presented methodology follows in Sect. 5. Finally, in Sect. 6 the presented approach and application results are summarised in a conclusion.

2 Data, notation and verification

2.1 Data

This study focusses on postprocessing of ensemble precipitation predictions from the NWP system *COSMO-E* (COnsortium of Small-scale MOdelling). At the time of writing, this is the operational probabilistic high-resolution NWP system of MeteoSwiss, the Swiss national weather service (Meteo Schweiz 2018). *COSMO-E* is run at 2x2km resolution for an area centered on the Alps extending from northwest of Paris to the middle of the Adriatic. An ensemble of 21 members is integrated twice daily at 00 : 00 and 12 : 00 UTC for five days (120 hours) into the future. Boundary conditions for the *COSMO-E* forecasts are derived from the operational global ensemble prediction system of the European Centre for Medium-range Weather forecasting (ECMWF).

We focus on daily precipitation amounts in Switzerland. As suggested by Messner (2018), the ensemble predictions and the observations of the daily precipitation amount are square-root transformed before the postprocessing. We use observed daily precipitation at MeteoSwiss weather stations to calibrate the ensemble forecasts. This paper relies on two observation datasets, one for the elaboration of the methodology and one for the subsequent assessment of the models of choice. The datasets are presented in Table 1. The first dataset (subsequently referred to as Dataset 1) consists of ensemble forecasts and verifying observations for the daily precipitation amount between January 2016 and July 2018 whereby a day starts and ends at 00 : 00 UTC. The data is available for 140 automatic weather stations in Switzerland recording sub-daily precipitation. The second dataset provides the observed daily precipitation amounts of 327 additional stations (on top of the 140 ones of Dataset 1) that record only daily sums. For historical reasons, these daily sums are aggregated from 06 : 00 UTC to 06 : 00 UTC of the following day. For the purpose of uniformity, these daily limits are adopted for all stations in Dataset 2. The larger number of stations from Dataset 2 is only available from June 2016 to July 2019. All stations from Dataset 1 are also part of Dataset 2. The stations of both datasets are depicted in Fig. 2.

Since *COSMO-E* makes forecasts for five days into the future, the different intervals between the forecast initialization time and the time for which the forecast is valid have to be taken into account. This is referred to as forecast *lead time*. The possible lead times for a prediction initialized at 00 : 00 UTC are 1, 2, 3, 4, 5 days and 1.5, 2.5, 3.5 and 4.5 days for one initialized at 12 : 00 UTC respectively. Figure 3 illustrates the possible lead times for both datasets. For Dataset 2 the forecast lead times increase from 24 hours to 30 hours for the first day due to the different time bounds of aggregation. Also, only four complete forecasts can be derived from each *COSMO-E* forecast with Dataset 2. We use Dataset 1 reduced to forecast-observation pairs with lead time equals 3 for the elaboration of the methodology. This selection procedure depends strongly on the dataset bearing the danger of overfitting. To assess this risk, the models of choice will be evaluated with Dataset 2. This is done for all lead times between 1 and 4.

In addition to the forecast-observation pairs and the station location (latitude, longitude, and altitude), topographical indices derived from a digital elevation model with 25m horizontal resolution are used. The topographical indices are available on 7 grids with decreasing horizontal resolution from 25m to 31km describing the topography from the immediate neighbourhood (at 25m resolution) to large-scale conditions (at 31km). The topographical indices include the height above sea level (DEM), a variable denoting if the site is rather in a sink or on a hill (TPI), variables describing aspect and slope of the site and variables describing the derivative of the site in different directions.

2.2 Notation

In this paper, the observed precipitation amount (at any specified location and time) is denoted as y. y is seen as a realization of a non-negative valued random variable Y. The K ensemble members are summarized as $\boldsymbol{x} = (x_1, \ldots, x_K)$. Predictive distributions for Y are denoted by F and stand either for cumulative distribution functions (CDFs) or probability density functions (PDFs). In literature, a capital F is used to refer indistinguishably to the probability measure or its associated CDF, a loose convention that we follow for simplicity within this paper. A

Table 1 The properties of Dataset 1 and Dataset 2

	Dataset 1		Dataset 2
Purpose	Elaboration methodology (Sect. 3)		Assessment (Sect. 4)
Available months	Jan 2016–Jul 2018		Jun 2016–Jul 2019
Aggregation period	00 : 00 UTC–00 : 00 UTC		06 : 00 UTC–06 : 00 UTC
Ensemble forecasts:			
Spatial resolution	2 × 2 km		2 × 2 km
Temporal resolution	Hourly accumulation		Hourly accumulation
Observations:			
Available stations	140		467 (140 stations of Dataset 1, 327 additional stations)
Temporal resolution	Hourly accumulation		Hourly accumulation for the 140 stations of Dataset 1 Daily accumulation for the 327 additional stations
Model training:	Sect. 3.1	Sect. 3.2–3.4	
Stations	Global models: All stations Dataset 1 (140) Local model: Station of interest (1 station each)	Global models: Cross-validation, remove prediction station (139 stations each) Local model: Station of interest (1 station each)	Global models: Stations of Datatset 1 (140) Local model: Station of interest (1 station each)
Months	Cross-Validation, remove prediction month (30 months each, using Jan 2016–Jul 2018)	12 months prior to prediction month (using Jan 2016–Jun 2018)	12 months prior to prediction month (using Jun 2016–Apr 2019)
Model validation:	Sect. 3.1	Sect. 3.2–3.4	
Stations	All stations Dataset 1 (140)	All stations Dataset 1 (140)	Additional stations of Dataset 2 (327)
Months	Jan 2016–Jul 2018	Jan 2017–Jul 2018	Jun 2017–May 2019

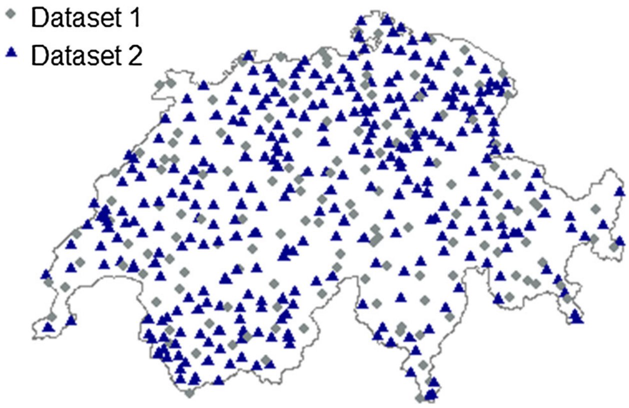
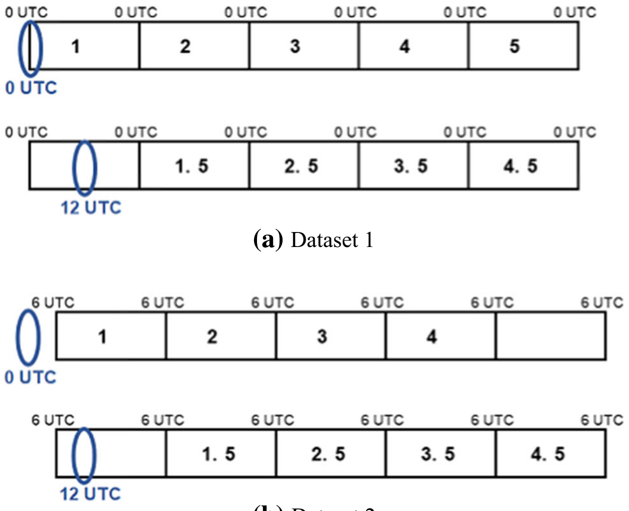


Fig. 2 The 140 stations of Dataset 1 (points) and the additional 327 stations of Dataset 2 (triangles)



(a) Dataset 1

(b) Dataset 2

Fig. 3 The lead times of Dataset 1 and Dataset 2, which classify the forecasts with respect to the time interval between the prediction time (circle) and the predicted day

forecast-observation pair written as $(\boldsymbol{x}_i, y_i)$ refers to a raw ensemble forecast and the corresponding observation. A pair written as (F_i, y_i) generally indicates here, on the other hand, that the forecast is a postprocessed ensemble prediction.

2.3 Verification

To assess a postprocessing model, the conformity of its forecasts and the associated observations is rated. In the case of probabilistic forecasts, a predictive distribution has to be compared with single observation points. We follow Gneiting et al. (2007) by aiming for a predictive distribution maximizing the *sharpness* subject to *calibration*.

Calibration refers to the statistical consistency between forecasts and observations (Gneiting et al. 2007; Thorarinsdottir and Schuhen 2018) and while several notions of calibration do exist (see Gneiting et al. 2007 for a detailed discussion with examples highlighting their differences), the notion of *probabilistic calibration* can probably be considered as the most common one. As recalled in Gneiting et al. (2007), "probabilistic calibration is essentially equivalent to the uniformity of the PIT values" (*Probability Integral Transform*; Dawid 1984). In practice, the n available forecast-observation pairs (F_i, y_i) with $i = 1, 2, \ldots, n$ out of the test dataset are examined by having a look at the histogram of the PIT values

$$F_1(y_1), \ldots, F_n(y_n). \tag{1}$$

While a flat histogram with equally populated bins is necessary for a forecasting method to be ideal, "checks for the uniformity of the PIT values have been supplemented by tests for independence" (Gneiting et al. 2007, referring to Frühwirth-Schnatter 1996 and Diebold et al. 1998). To investigate the calibration of the raw ensemble, the discrete equivalent of the PIT histogram called *verification rank histogram* is used (Hamill and Colucci 1997). It is generated by ranking the values

$$\{x_1, x_2, \ldots, x_K, y\} \tag{2}$$

of every ensemble-observation pair. A histogram of the ranks of the observations shows how they are distributed within the ensemble. Again, a flat histogram indicates a calibrated forecasting method.

Hamill (2001) pointed out that the flatness of the PIT histogram is necessary but not sufficient for a forecast to be ideal. Gneiting et al. (2007) took these results as a motivation to aim for a predictor which maximizes the sharpness while being calibrated. Sharpness relates to the concentration of the predictive distribution; a more concentrated predictive distribution means a sharper forecast. Being a characteristic of the forecast only and not comparing it with the actual observations, sharpness is typically considered in conjunction with calibration rather than individually (Wilks 2011).

Accuracy The *accuracy* of a forecast is assessed with summary measures addressing both calibration and sharpness simultaneously. These functions called *Scoring Rules* map each forecast-observation pair (F, y) to a numerical penalty, where a smaller penalty indicates a better forecast and vice versa (Thorarinsdottir and Schuhen 2018). Let

$$S : \mathcal{F} \times \mathcal{Y} \to \mathbb{R} \cup \{\infty\} \tag{3}$$

be a Scoring Rule, where $\mathcal{Y}$ is a set with possible values of the quantity to be predicted and $\mathcal{F}$ a convex class of probability distributions on $\mathcal{Y}$.The Scoring Rule is said to be *proper* relative to the class $\mathcal{F}$ if

$$\mathbb{E}_{Y\sim G} S(G, Y) \leq \mathbb{E}_{Y\sim G} S(F, Y), \tag{4}$$

where $F, G \in \mathcal{F}$ are probability distributions and G in particular is the true distribution of the random observation Y. The subscript $Y \sim G$ at the expected value denotes that the expected value is computed under the assumption that Y has distribution G (Gneiting et al. 2007).

We will focus on two popular Scoring Rules: The *Continuous Ranked Probability Score* (CRPS; Matheson and Winkler 1976) and the *Brier Score* (Brier 1950). The CRPS is a generalization of the absolute error for probabilistic forecasts. It can be applied to predictive distributions with finite mean and is defined as follows (Thorarinsdottir and Schuhen 2018):

$$CRPS(F, y) = \int_{-\infty}^{\infty} \left(F(x) - \mathbb{1}_{[y,\infty)}(x)\right)^2 dx, \tag{5}$$

where $\mathbb{1}_A(x)$ denotes the indicator function for a set $A \subseteq \mathbb{R}$ which takes the value 1 if $x \in A$ and 0 otherwise. To compare the performance of the postprocessing models with the one of the raw ensemble, we need a version of the CRPS for a predictive distribution F_{ens} given by a finite ensemble $x_1, \ldots, x_K$. We use the following definition by Grimit et al. (2006):

$$\begin{aligned} CRPS(F_{ens}, y) = & \frac{1}{K}\sum_{k=1}^{K} |x_k - y| \\ & - \frac{1}{2K^2}\sum_{k=1}^{K}\sum_{l=1}^{K} |x_k - x_l|. \end{aligned} \tag{6}$$

In practice, competing forecasting models are compared by calculating the mean CRPS values of their predictions over a test dataset. The preferred method is the one with the smallest mean score. We use a *Skill Score* as in Wilks (2011) to measure the improvement (or deterioration) in accuracy achieved through the postprocessing of the raw ensemble:

$$Skill(F, F_{ens}, y) = 1 - \frac{CRPS(F, y)}{CRPS(F_{ens}, y)}, \tag{7}$$

where the $Skill(F, F_{ens}, y)$ characterizes the improvement in forecast quality by postprocessing ($CRPS(F, y)$) relative to the forecast quality of the raw ensemble ($CRPS(F_{ens}, y)$).

The definition of the CRPS given in Equation (5) corresponds to the integral over another Scoring Rule: The Brier Score assesses the ability of the forecaster to predict the probability that a given threshold u is exceeded. The following definition of the Brier Score is taken from Gneiting et al. (2007):

$$BS(F, y|u) = (F(u) - \mathbb{1}_{[y,\infty)}(u))^2 \tag{8}$$

If the predictive distribution F_{ens} is provided by a finite ensemble $x_1, ..., x_K$, we use the following expression for the Brier Score:

$$BS(F_{ens}, y|u) = \left\{ \left(\frac{1}{K} \sum_{k=1}^{K} \mathbb{1}_{[x_k,\infty)}(u) \right) - \mathbb{1}_{[y,\infty)}(u) \right\}^2. \tag{9}$$

3 Postprocessing model

The aim of this chapter is to find a suitable postprocessing model by comparing the forecast quality of different approaches on the basis of Dataset 1.

3.1 Censored logistic regression

In this section, we present a conventional ensemble postprocessing approach as a starting point for later extensions. We compared several EMOS models and Bayesian Model Averaging in a case study with Dataset 1 whereby a censored inhomogeneous Logistic regression (cNLR; Messner et al. 2014) model turned out to be the most promising approach. The performance of the models has been assessed by cross-validation over the months of Dataset 1. Of the 31 months available, the months are removed one at a time from the training data set and the model is trained with the remaining 30 months. Predictive performance is then assessed based on comparison between the observations at each left-out month from the models trained on the remaining months. The basic models are tested in two versions: For the global version, the model is trained with the data of all stations allowing the later application of the model to all stations simultaneously. The local version requires that models are trained individually for each station with the past data pairs of this station. More details on the alternative approaches to the cNLR model and the results from the comparison of approaches can be found in the supplementary material.

The cNLR approach is a distributional regression model: We assume that the random variable Y describing the observed precipitation amount follows a probability distribution whose moments depend on the ensemble forecast. To choose a suitable distribution for Y, we take into account that the amount of precipitation is a non-negative quantity that takes any positive real value (if it rains) or the value zero (if it does not rain). These properties are accounted for by appealing to a zero censored distribution. We assume that there is a latent random variable Y^* satisfying the following condition (Messner et al. 2016):

$$Y = \begin{cases} Y^* & \text{for } Y^* > 0, \\ 0 & \text{for } Y^* \leq 0. \end{cases} \tag{10}$$

In this way, the probability of the unobservable random variable Y^* being smaller or equal than zero is equal to the probability of Y being exactly zero.

For the choice of the distribution of Y^*, we have compared different parametric distributions: a Logistic, Gaussian, Student, Generalized Extreme Value and a Shifted Gamma distribution. For the Logistic distribution, which has achieved the best results, $Y^* \sim \mathcal{L}(m, s)$ with location m and scale s has probability density function

$$f(y; m, s) = \frac{\exp\left(-\frac{y-m}{s}\right)}{s\left(1 + \exp\left(-\frac{y-m}{s}\right)\right)^2}. \tag{11}$$

The expected value and the variance of Y^* are given by:

$$\mathbb{E}(Y^*) = m, \quad Var(Y^*) = \frac{s^2\pi^2}{3}. \tag{12}$$

The location m and the scale s of the distribution have to be estimated with the ensemble members. We note that in the *COSMO-E* ensemble, the first member x_1 is obtained with the best estimate of the initial conditions whereas the other members are initialized with randomly perturbed initial conditions. The first member is thus not exchangeable with the other members and it is therefore reasonable to consider this member separately. Members $x_2, x_3, ..., x_{21}$ are exchangeable, meaning that they have no distinct statistical characteristics. Within the location estimation, they can therefore be summarized by the ensemble mean without losing information (Wilks 2018). Taking this into account, we model the location m and the scale s of the censored Logistic distribution as follows:

$$m = \beta_0 + \beta_1 x_1 + \beta_2 \bar{x}, \tag{13}$$

$$log(s) = \gamma_0 + \gamma_1 SD(\boldsymbol{x}), \tag{14}$$

where $\bar{x}$ is the ensemble mean and $SD(\boldsymbol{x})$ is the ensemble standard deviation. The five regression coefficients are summarized as $\psi = (\beta_0, \beta_1, \beta_2, \gamma_0, \gamma_1)$.

Optimization and implementation: Let (F_i, y_i) be n forecast-observation pairs from our training dataset $(i = 1, 2, \ldots, n)$. The predictive distributions F_i are censored Logistic with location and scale depending on the ensemble forecasts $\boldsymbol{x_i}$'s and the coefficient vector ψ. We use *Scoring Rule estimation* (Gneiting et al. 2005) for the fitting of ψ. Therefore we select a suitable Scoring Rule and express the mean score of the training data pairs as a function of ψ. Then, ψ is chosen such that the training score is minimal. In this study, the Scoring Rule of choice is the CRPS. For the implementation of the cNLR model we use the R-package *crch* by Messner et al. (2016).

3.2 Topographical extension

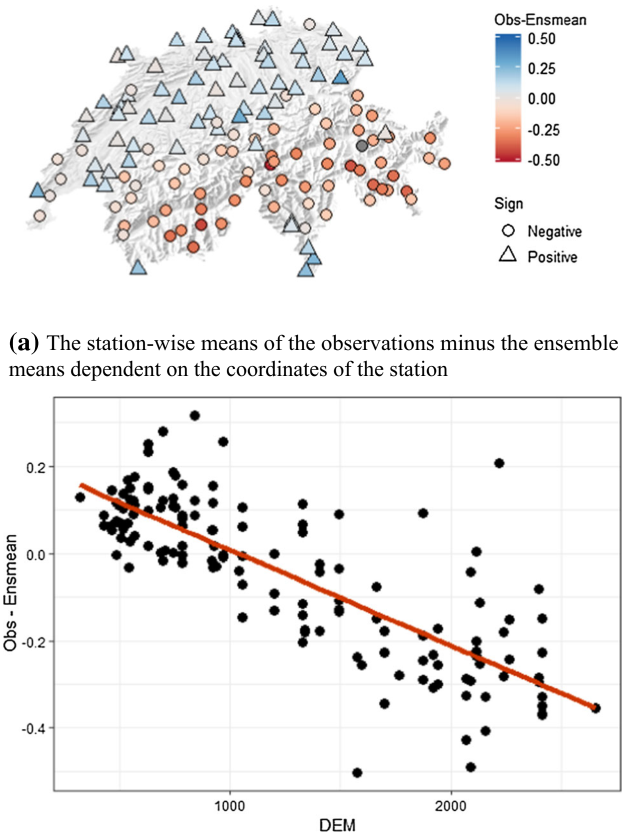
A case study with Dataset 1 showed that the performance of the local cNLR model is better than the one of the global cNLR model (see supplementary material). Similar results have been presented for example by Thorarinsdottir and Gneiting (2010) in a study about wind speed predictions. The local model, however, cannot be used for area-covering postprocessing. To improve the performance of the global model, it is enhanced with topographical covariates. The idea is to fit the regression coefficients only with training data from weather stations which are topographically similar to the prediction site.

We assume that the training dataset consists of n forecast-observation pairs $(\boldsymbol{F_i}, y_i)$ with $i = 1, 2, \ldots, n$. The global cNLR model estimates ψ by minimizing the mean CRPS value of all these training data pairs. To select only or favour the training pairs from similar locations, we use a weighted version of the mean CRPS value[1] as the cost function c, which is minimized for the fitting of the coefficient vector ψ:

$$c(\psi; s) = \sum_{i=1}^{n} w_i(s) CRPS(F_i^{\psi}, y_i). \tag{16}$$

$CRPS(F_i^{\psi}, y_i)$ refers to the CRPS value of data pair (F_i^{ψ}, y_i) where the predictive distribution F_i^{ψ} depends on the coefficient vector ψ. We use (F_i, y_i) as shorthand for (F_i^{ψ}, y_i). The weight $w_i(s)$ of training data pair (F_i, y_i) depends on the similarity between the location it originated and the location s where we want to predict. We set $w_i(s) = 1$ if training data pair i originated in one of the L closest (be it with respect to the euclidean distance or to some other dissimilarity measure) stations to the prediction site s. For the training pairs from the remaining stations, we set $w_i(s) = 0$. This ensures that the training dataset is restricted to the data from the L stations which are most similar to the prediction site s. Consequently, the coefficient vector $\psi^*(s)$ which minimizes $c(\psi; s)$ depends on the location s.

Following Lerch and Baran (2018), the similarity between locations is quantified with a distance function, which, in our case, is intended to reflect the topography of the respective locations. From the topographical dataset we have about 30 variables in 8 resolutions at our disposal. To get an insight regarding which ones to use, we examine the topographical structure of the raw ensemble's prediction error. We compare the observed daily precipitation amounts with the ensemble means and take the station-wise averages of these differences. These preliminary analyses were made with the first year (2016) of Dataset 1. The mean prediction errors per station are depicted in Fig. 4. The topographical structure of the ensemble prediction



(a) The station-wise means of the observations minus the ensemble means dependent on the coordinates of the station

(b) The same means as in (a) dependent on the height above the sea level (DEM) in resolution 31km, the solid line depicts the best linear function through the points

Fig. 4 The station-wise means of the observations minus the ensemble means for the data from 2016 of Dataset 1

[1] Literature knows another kind of weighted CRPS value: Threshold and quantile weighted versions of the CRPS are used when wishing to emphasize certain parts of the range of the predicted variable. The threshold weighted version of the CRPS is given by

$$CRPS_u(F, y) = \int_{-\infty}^{\infty} (F(x) - \mathbb{1}_{[y,\infty]}(x))^2 u(x) dx, \tag{15}$$

where u is a non-negative weight function on the real line (Gneiting and Ranjan 2011, consult their paper for the analogous definition of the quantile weighted version of the CRPS).

error seems to be linked to the mountainous relief of Switzerland.

In a first approach, we define the similarity of two locations via the similarity in their distances to the Alps. It turns out that such an approach depends on numerous parameters (we refer the reader to the supplementary material for more details). The proposed alternative is to focus on the variable DEM describing the height above sea level and use the values provided by a grid with low resolution (here 31 km horizontal grid spacing). This ensures, that the small-scale fluctuations in the height are ignored such that the large-scale relief is represented. Figure 4 shows the same station-wise means as Fig. 4, but this time the values of each station are plotted versus their height above the sea level (DEM) in resolution 31*km*. The best fit with a polynomial is achieved by modelling the ensemble mean bias as a linear function of the DEM variable; the solid line is depicting this linear function.

The linear dependency appearing in Fig. 4 motivates the following choice of a distance function to measure the similarity between two locations. Let us define a function *DEM* which maps a location s to its height above the sea level in the resolution 31km:

$$DEM : \mathcal{D} \to \mathbb{R}, \quad s \mapsto DEM(s), \tag{17}$$

where $\mathcal{D} \subseteq \mathbb{R}^2$ is a set with all the coordinate pairs lying in Switzerland. The similarity of locations s_1 and s_2 is then measured by the following distance:

$$d_{DEM}(s_1, s_2) = |DEM(s_1) - DEM(s_2)|. \tag{18}$$

Based on this distance, we determine the L stations of the training data set which are most similar to the prediction location, i.e. the stations which have the smallest distances d_{DEM}. Let

$$D^1_{d_{DEM}}(s) \leq D^2_{d_{DEM}}(s) \leq \ldots \leq D^m_{d_{DEM}}(s) \tag{19}$$

be the ordered distances $d_{DEM}(s, s_j)$ between the m stations from the training data and the prediction location s. Let s_i be the location of the station where forecast-observation pair (F_i, y_i) originated. Then, the weights with respect to the distance d_{DEM} are defined as:

$$w^L_i(s) = \begin{cases} 1 & \text{for } d_{DEM}(s, s_i) \leq D^L_{d_{DEM}}(s), \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

These weights ensure that the data pairs, which originated at one of the L most similar stations, get weight 1 and the remaining get weight 0.

Besides this approach, several other topographical extensions of the cNLR model have been tested (with Dataset 1): For their spatial modelling, Khedhaouiria et al. (2019) propose to vary the postprocessing parameters by expressing them as a function of spatial covariates. We have applied a similar approach and integrated the topographical covariates in the location estimation of the cNLR model. To reduce the number of predictor variables, the topographical variables have been summarized by *Principal Components*. Additionally, we used the *glinternet* algorithm of Lim and Hasti (2013) to uncover important additive factors and interactions in the set of topographical variables. A more basic weighted approach has been based on Euclidean distances in the ambient (two and three dimensional) space. All extensions of the cNLR model have been compared with the local and the global fit of this very model.

As the target of this work is to develop an area-covering postprocessing method, the extended and the global models are trained with data where the predicted month and the predicted station are left out. This simulates the situation where postprocessing must be done outside a weather station, i.e. without past local measurements. The training period is set to the last year (12 months) before the prediction month. Consequently, forecasting performance can only be assessed with (test) data from 2017 and 2018. The case study with Dataset 1 showed that all other topographical approaches are less successful than the DEM approach, more details and the results can be found in the supplementary material (in the section about extension approaches).

3.3 Seasonal extension

In addition to our efforts to quantify similarities between locations, we also aim to investigate ways of further improving postprocessing outside of measurement networks by accounting for seasonal specificities. To examine the seasonal behavior of the local and the global cNLR model, we focus on their monthly mean CRPS values and

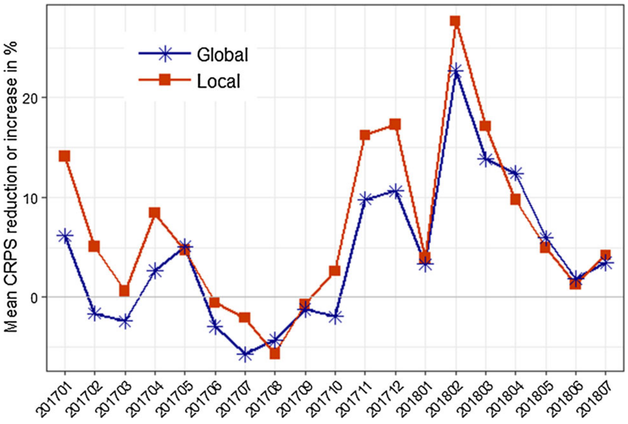


Fig. 5 The monthly skill of the local and the global cNLR model which compares the monthly mean CRPS value of the model with the one of the raw ensemble, the values describe the reduction or increase of the mean CRPS value in percent

compare them with the ones of the raw ensemble. Figure 5 shows the monthly skill of the global and the local cNLR model. We use the mean over all the stations from Dataset 1 and depict the months between January 2017 and July 2018 such that we have 12 months of training data in each case. A positive skill of 10% means for example that the mean CRPS value of the raw ensemble is reduced by 10% through postprocessing, a negative skill indicates analogously that the mean CRPS value is higher after the postprocessing. The global model has negative skill in February and March 2017 and between June and October 2017. The values are especially low in July and August 2017. The local model has a positive skill in most months, the postprocessing with this model decreases the forecasting performance only between June and September 2017. We use these results as a first indication that the postprocessing of ensemble precipitation forecasts is particularly challenging in summer and early autumn.

Next, we are interested in whether there are regional differences in the model performance within a month. The global cLNR model is used as we will extend this model afterwards. We plot the maps with the station-wise means of the skill exemplary for the month with the best skill (February 2018) and the one with the worst (July 2017). The maps depicted in Fig. 6 show that the skill of the global cNLR model varies between different weather stations. Again, the structure seems to be related to the mountainous relief of Switzerland. We note that for both months the skill in the Alpine region is distinctly higher than in the flat regions.

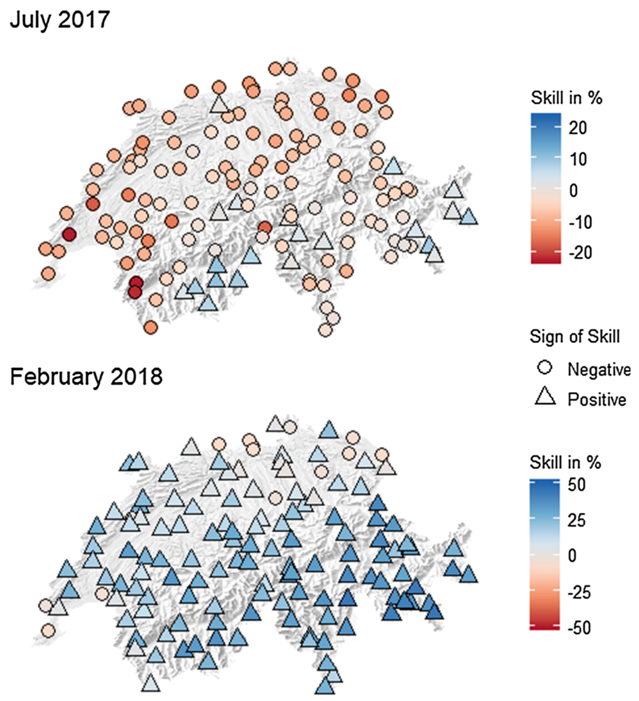


Fig. 6 The station-wise skill of the global cNLR model for July 2017 and February 2018 which compares the mean CRPS value of the model with the one of the raw ensemble, the values describe the change in percent

We use this knowledge to develop an approach which tries firstly to clarify whether the postprocessing in a given month at a given prediction location is worthwhile. The idea is to “pretest” the model with data of similar stations and from similar months by comparing its performance with that of the raw ensemble. For this purpose, the year of training data is first reduced to the data pairs from topographically similar stations, whereby the similarity is measured with the distance d_{DEM} defined in Eq. (18). Afterwards, this training dataset is split into two parts: *Traintrain* and *Traintest*. The model is adapted a first time with the dataset Traintrain. Afterwards, the performance of this model is assessed with the second part (Traintest) by comparing the mean CRPS of the model with the mean CRPS of the raw ensemble.

The months of the Traintest dataset are selected such that they are seasonally similar to the prediction month. To split the training dataset, three approaches are compared:

- Pretest 1: Pretest with the same month as the prediction month from the year before (Example: January 2017 for January 2018)
- Pretest 2: Pretest with the month before the prediction month (Example: December 2017 for January 2018)
- Pretest 3: Pretest with both of these months (Example: January 2017 and December 2017 for January 2018).

Let us define the set of indices of training data pairs out of Traintest:

$$h(i) = \{i \in \{1, 2, ..., n\} : (\boldsymbol{x_i}, y_i) \text{ is in Traintest}\}, \qquad (21)$$

with cardinality H. Let further $(\boldsymbol{x_i}, y_i)$ be a forecast-observation pair where the forecast is the raw ensemble. (F_i, y_i) is a pair with a postprocessed forecast. If

$$\frac{1}{H}\sum_{i\in h(i)} CRPS(\boldsymbol{x_i}, y_i) \leq \frac{1}{H}\sum_{i\in h(i)} CRPS(F_i, y_i), \qquad (22)$$

then the pretesting algorithm decides that the raw ensemble is not postprocessed in the given month at the given location. On the contrary if

$$\frac{1}{H}\sum_{i\in h(i)} CRPS(\boldsymbol{x_i}, y_i) > \frac{1}{H}\sum_{i\in h(i)} CRPS(F_i, y_i), \qquad (23)$$

then the pretesting algorithm decides that the raw ensemble is postprocessed in the given month at the given location, the fit is done a second time with the whole year of training data.

The Pretest approach has been compared with several other seasonal approaches: In a basic approach, we reduce the training period to months from the same season as the prediction month. Another approach uses the sine-transformed prediction month as an additional predictor variable to model the yearly periodicity, an approach comparable to the one of Khedhaouiria et al. (2019). The third approach reduces the training data to pairs which have a similar prediction situation (quantified with the ensemble mean and the ensemble standard deviation). The methodology for the comparison has been the same as for the topographical extensions introduced in Sect. 3.2. The Pretest approach turns out to be the most promising method for our case study with Dataset 1, more details and the comparison results can be found in the supplementary material.

3.4 Model adjustment

For the subsequent evaluation of postprocessing models with Dataset 2, we select a few postprocessing approaches to document the impact of increasing complexity on forecast quality. We will use the raw ensemble and the local such as the global version of the cNLR model as baselines. Further on, we will evaluate the cNLR model extended by the DEM similarity (cNLR DEM). Finally, we will test this same model extended a second time with the pretest approach (cNLR DEM+PT). For the last two models, we have to fix the amount L of similar stations we use for the topographical extension. For the last model we need to fix additionally the pretesting split. To determine the amount of similar stations in use, the numbers which are multiples of ten between 10 and 80 have been tested (compare Fig. 7). We use the data from August 2017 to July 2018 (seasonally balanced) from Dataset 1 and choose the number resulting in the lowest mean CRPS. For the cNLR DEM model (no Pretest) we determine $L = 40$. For the cNLR DEM+PT model, we combine the different pretesting splits with the same numbers for L. The cNLR DEM+PT model with the lowest mean CRPS value uses Pretest 3 and $L = 40$.

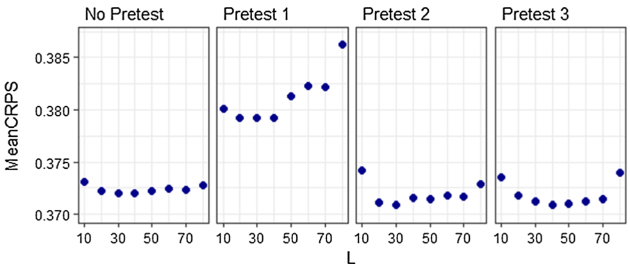


Fig. 7 The mean CRPS values for the cNLR DEM (+ Pretest) models comparing different numbers for L and the different pretesting splits

4 External validation

This chapter presents the evaluation of the different postprocessing models. As already announced, the independent Dataset 2 is used to take into account the risk of overfitting during the elaboration of the methodology.

4.1 Methodology

We are interested in the area-covering performance of the models. Therefore, we are particularly interested in the performance at locations which cannot be used in model training (as no past data is available). This is the reason why we assess the models only with the 327 additional stations of the second dataset. None of these stations have been used during the model elaboration in Sect. 3. When determining L in chapter 3.4 we used a training dataset with 139 stations (139 instead of 140 as we trained without the past data from the prediction station). For this reason we carry on with using only the 140 stations of the first dataset to train the models. This rather conservative approach could be opposed by a Cross Validation over all 467 stations, for which, however, another choice of L would probably be ideal.

The local version of the cNLR model is not able to perform area-covering postprocessing and needs the additional stations in the training from Dataset 2. Despite this, it is fitted and assessed as a benchmark here. We train the models for each of the 327 stations and each month between June 2017 and May 2019. This ensures that we have one year of training data for each month and that we have seasonally balanced results. An individual fitting per station is necessary as the selection of the most similar stations used in the DEM approaches depends on the station topography. The model must also be adapted monthly, as the pretesting procedure (and the training period) depend on the prediction month.

During the postprocessing, we used consistently the square root of the precipitation amount. The CRPS value, which is in the same unit as the observation, refers to this transformation as well. To get an idea of the actual order of magnitude, the values are converted into the original size, in which the precipitation amount is measured in mm. As a first step, 21 forecasts are drawn from the fitted censored Logistic model. Afterwards, these values and the corresponding observations are squared and the mean CRPS is calculated as for the raw ensemble. The Brier Score, which assesses the ability of the forecaster to predict if a given precipitation accumulation amount is exceeded, is also evaluated for the squared sample of the predictive distribution. The thresholds used within the Brier Score focus on

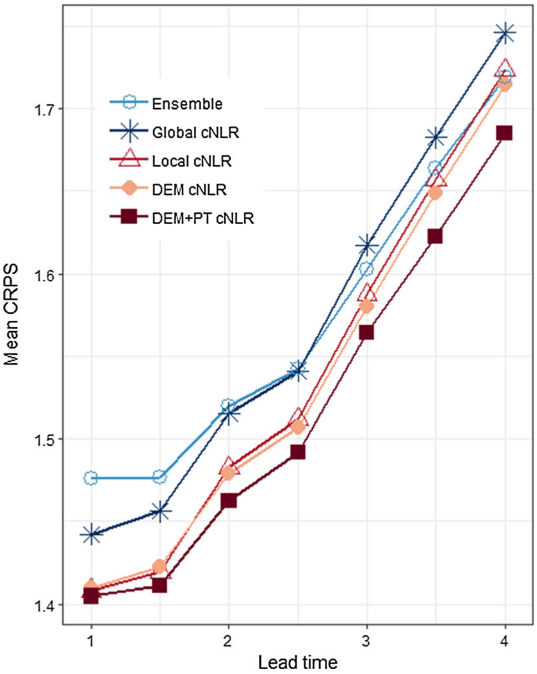


Fig. 8 Mean CRPS values for the raw ensemble and the different postprocessing models dependent on the lead time, the assessment is based on the data from June 2017 to May 2019 of Dataset 2

three precipitation accumulations: No rain (< 0.1 mm/d) , moderate rain (> 5 mm/d) and heavy rain (> 20 mm/d).

4.2 Results

First of all, let us give an overview of the different models. Figure 8 depicts the mean CRPS values for the different postprocessing approaches and lead times. We refer to Chapter 3.4 for a recap of the model adjustments concerning the DEM and DEM + Pretest model. For lead time 1, the global cNLR model is able to reduce the mean CRPS value by 2.3%. A further improvement is achieved by the local and the DEM model, which show equivalent performances and reduce the mean CRPS value by 4.5% compared to the raw ensemble. Even slightly better results are delivered by the DEM + Pretest model which reduces the mean CRPS value by 4.8%. The skill of the global model decreases with increasing lead time. While the skill is still positive for lead times 1, 1.5 and 2, the model performs roughly equally as the raw ensemble for lead time 2.5. From lead time 3, the mean CRPS value of the global model is even higher than the one of the raw ensemble. The local and the DEM model perform about the same for lead times between 1 and 2.5, for lead times above 3 the DEM model performs slightly better. The DEM+Pretest model performs best for all lead times. It reduces the mean CRPS value between 4.8% for lead time 1 and 2.0% for lead time 4. It is noticeable that the DEM + Pretest model achieves a near constant improvement in the mean CRPS of approx. 0.07 for all lead times. Relatively—i.e. as a Skill Score—this corresponds to less and less of the total forecast error. We note additionally, that the improvement which is achieved through the extension of the DEM model with the Pretest depends on the lead time. While the Pretest reduces the mean CRPS of the DEM model only by 0.4% for lead time 1, the obtained reduction corresponds to a proportion of 1.7% for lead time 4.

Figure 8 summarizes the average performance of the models over all months. To assess the seasonal performance of the different approaches, the monthly means of the Skill Score are plotted in Fig. 9. We use the raw ensemble forecast as reference and depict the results for lead time 1 and lead time 4. For lead time 1, we note that the DEM + Pretest model is the only one with non-negative skill in almst all months, implying that this model only rarely degrades the quality of the ensemble prediction. While the model delivers in summer and early autumn equivalent results as the raw ensemble, the monthly mean CRPS value can be reduced by up to 20% in winter months. The same improvement is achieved with the models without Pretest, but they have a slightly worse overall performance as they degrade forecast performance during summer and autumn. For longer lead times (illustrated exemplarily for lead time 4, right panel of Fig. 9), postprocessing is less successful in improving forecast quality with forecasts in summer often deteriorating for all but the DEM + Pretest method. With pretesting the seasonal cycle in quality improvements is much less apparent for lead time 4 than for lead time 1. This is likely due to the combination of calibration + pretesting which is performed at individual stations, which guarantees (in expectation) that the quality of postprocessed forecasts is at least as good as that of the direct model output. If, on the other hand, there is considerable miscalibration of forecasts even if only at a few stations, this can be exploited. We also detect noticeable differences in the improvements which are achieved by extending the DEM model with the Pretest between lead times 1 and 4: The improvement for lead time 4 is higher in most months, especially for June to August 2017 and August to October 2018.

We have also examined the spatial skill of the models. Therefore, we have compared the station-wise mean CRPS of the models with the one of the raw ensemble. The skill, which is very similar at neighbouring stations, increases in the Alps and is marginal or non-existent in the Swiss plateau. The resulting spatial distribution of the skill looks similar as the mean bias depicted on Fig. 4, the maps are therefore not shown here.

As proposed by Thorarinsdottir and Schuhen (2018), we use more than one Scoring Rule for the assessment of our postprocessing methods and apply the Brier Score to

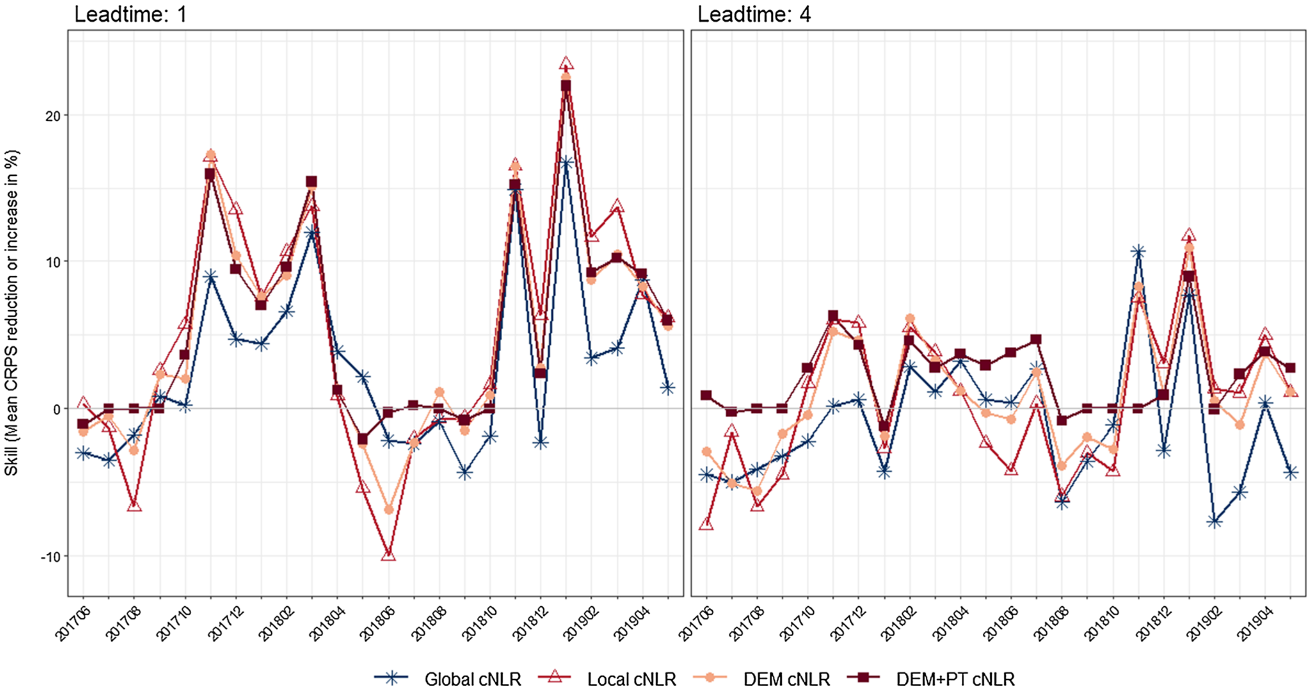


Fig. 9 Reduction and increase (in %) of the monthly mean CRPS value of the raw ensemble by the different postprocessing approaches, the assessment is done with the data from June 2017 to May 2019 of Dataset 2

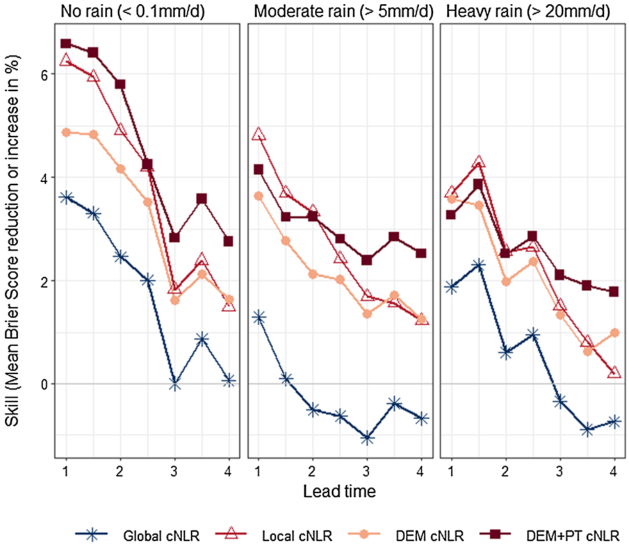


Fig. 10 Skill of the different postprocessing approaches and lead times, depicted through the reduction and increase (in %) of the mean Brier Score of the raw ensemble for the thresholds of 0.1mm/d, 5mm/d and 20mm/d. The data from June 2017 to May 2019 in use is taken from Dataset 2

evaluate forecast quality of specific events. For precipitation forecasts, the ability to predict whether it will rain or not is of particular interest, this is captured by the Brier score for daily rainfall < 0.1 mm (precision of observation measurements). We extend the assessment by also considering forecasts of moderate and heavy precipitation characterized by daily rainfall above 5 mm/d and 20 mm/d. Figure 10 illustrates the skill of the different postprocessing models by comparing the mean Brier Scores of the different models, thresholds and lead times with the ones of the raw ensemble forecasts. The assessment with the Brier Score confirms that the improvement achieved with the DEM + Pretest model is higher than the one with the other models. Only for lead times 1 and 1.5 and moderate or heavy rainfall, the local model outperforms the DEM + Pretest approach. Overall, the skill decreases with increasing threshold and increasing lead time. This is to be expected given that the postprocessing focuses on improving forecasts on average and exceedances of high thresholds are relatively rare (4% of the observed daily rainfall falls in the heavy rainfall category). Also, we use square-root transformed precipitation in the optimization which further reduces the importance of heavy precipitation events in the training set. The plot confirms further that the global model performs worst, for moderate or heavy rainfall and a lead time above 1.5 respectively 2.5 even worse than the raw ensemble. As in measures of the CRPS, the local and the DEM model score comparable for all thresholds and lead times between 2.5 and 4. For smaller lead times, the local model performs better for all thresholds.

Raw ensemble forecasts are often underdispersed and have a wet bias (Gneiting et al. 2005). This holds for the ensemble precipitation forecasts used in this study as well.

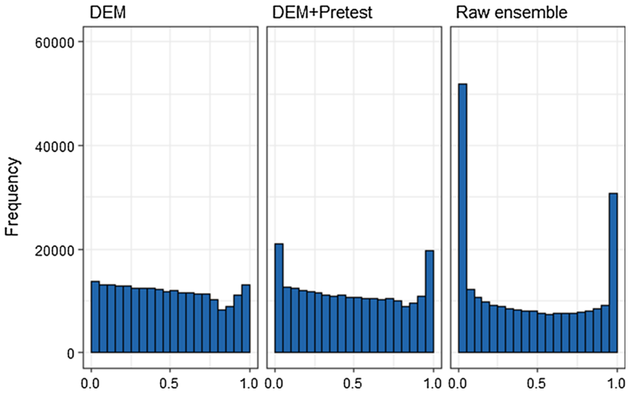


Fig. 11 The PIT histograms for the DEM and DEM + Pretest models and the verification rank histogram for the raw ensemble forecasts (the lead time in use is 1)

Figure 11 (right) shows the verification rank histograms for the raw ensemble. Again, we use the data from June 2017 to May 2019 of Dataset 2 for this assessment and depict the results for lead time 1. We note that the histogram for the raw ensemble has higher bins at the left and right marginal ranks and higher bins for the ranks which lie in the first half of 1, 2, ...,21. Therefore, it indicates that the ensemble forecasts are underdispersed and tend to have a wet bias. This raises the question whether the results of the DEM + Pretest model are still calibrated.

To be able to evaluate the calibration of the full predictive distribution of the postprocessing models, we do not use the reverse transformation of the square root for this assessment. Additionally, we have to use a randomized version of the PIT as our predictive distribution has a discrete component (Thorarinsdottir and Schuhen 2018):

$$\lim_{y \uparrow Y} F(y) + V\Big(F(Y) - \lim_{y \uparrow Y} F(y)\Big), \tag{24}$$

where $V \sim \mathcal{U}([0,1])$ and $y \uparrow Y$ means that y approaches Y from below.

Figure 11 shows the PIT histograms for the DEM and the DEM+Pretest model. As expected, the PIT histogram of the Pretest model lies between the one of the raw ensemble and the one of the DEM model without Pretest. The first and the last bins, which are higher than the other bins, indicate that the Pretest model is underdispersed. However, it seems much less gravely than for the raw ensemble. Since the remaining tested seasonal approaches produce worse results, this slight miscalibration of the DEM + Pretest model is a disadvantage we have to accept for the moment.

Finally, we want to get an idea of the acceptance behaviour of the DEM + Pretest model. Figure 12 shows when and where the DEM + Pretest model does (light point) and does not (dark point) postprocess the raw ensemble. Again, we focus on the results of lead time 1. The plots show that the months where the model uses the raw ensemble lie mostly in summer and autumn. The postprocessing during these months is accepted at stations which lie in bands of different widths parallel to the Alps. The model postprocesses the raw ensemble at all stations during almost all months in between December and May, the only exception are March and April 2019. Therefore, it appears that the Pretest approach can address the seasonal difficulties of postprocessing ensemble precipitation forecasts.

5 Discussion

To enable area-covering postprocessing, we use a model that weights the training data depending on the similarity between its location of origin and the prediction location. This basic principle could be applied to any postprocessing problem where the prediction and the observation locations do not match. However, some of the choices made in this case study are quite specific and data dependent, in particular the presented procedure used to determine the number of most similar stations with which the models are trained. The models use a similarity based weighted CRPS estimator to fit the regression coefficients. The clarification of the asymptotic behaviour of such an estimator could help determining the ideal number of stations to train with and making the elaboration of the methodology less sensitive to the data in use.

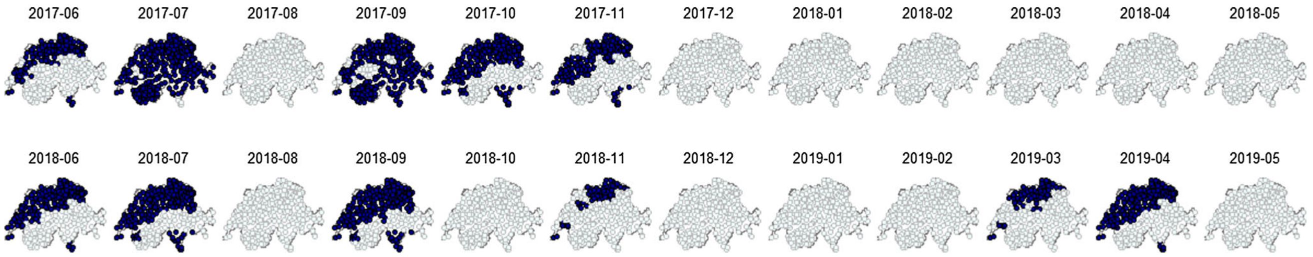


Fig. 12 Maps depicting the acceptance behaviour of the DEM+−Pretest model. The model evaluates for each of the 327 stations and 24 months of the test dataset if a postprocessing the raw ensemble seems worthwhile, the maps show when and where the model does (light point) and does not (dark point) postprocess the raw ensemble (the lead time in use is 1)

The Pretest, which decides whether a postprocessing is worthwhile in a given setting has the disadvantage that the calibration of the resulting forecast is not guaranteed. Yet, although numerous alternative seasonal approaches have been tested, the CRPS of the Pretest model could not be levelled. In addition, making a Pretest means that the model must be adjusted twice, which is computationally expensive. But the strength of this approach is that it is fairly universally applicable also to problems outside meteorology—given that one is willing to accept some loss in model calibration for the obtained gain in accuracy.

There are various directions in which the model could be further expanded: More meteorological information could be added such as covariates describing the large-scale flow. Further meteorological knowledge could also be incorporated by supplementing the DEM-distance with a further distinction between north and south of the Alps, for instance. The scale estimation of the final model is based only on the standard deviation of the ensemble. This estimation could be further extended with additional predictors to ensure that the ensemble dispersion is adjusted with respect to the prediction setting as well (as done for the location estimation in the alternative extension approaches).

The evaluation with the Brier Score displays that for the case of no rain, all postprocessing models perform better than the raw ensemble. For the case of moderate rain, the global model is superseded by far by the models integrating a local aspect (local, DEM and DEM + Pretest model). The differences between the local and global models are more moderate for the case of heavy rain, but while all local approaches exceed the performance of the raw ensemble, this is not the case with the global model. Investigating further the behaviour of different approaches on the range / threshold of interest and eventually developing a postprocessing method with a focus on rare events would open exciting avenues of research. The work of Friederichs et al. (2018) offers an introduction to ensemble postprocessing of extreme weather events, an exemplary application for extreme rainfall intensity can be found in Shin et al. (2019). To avoid local fluctuations and reflect the spatial dependencies between neighbouring locations, Shin et al. (2019) use a spatial extreme model, namely a max-stable process. The indicated potential to link postprocessing of extreme weather events to area-covering approaches is left for future research.

6 Conclusion

The aim of this case study was to produce improved probabilistic precipitation forecasts at any place in Switzerland by postprocessing *COSMO-E* ensemble forecasts enhanced with topographical and seasonal information. During the elaboration of the methodology, a censored nonhomogeneous Logistic regression model has been extended step by step; the final model combines two approaches.

A semi-local approach is used for which only data within a neighbourhood around the prediction location are used to establish the postprocessing model. The training data used to fit the regression coefficients is weighted with respect to the similarity between its location of origin and the prediction location. This similarity is determined based on the smoothed elevation, i.e. the topographical variable DEM in a resolution of 31km. Using this approach, the weighting of the training data can be adapted for any prediction location and the model can be applied to the entire area of Switzerland thus fullfilling the first requirement of this study.

In addition, a seasonal Pretest ensures that the model only postprocesses the raw ensemble forecast when a gain is expected—as assessed in the training sample. This extension addresses the second objective of this study and ensures that the postprocessing model accounts for seasonal specificities such as enhanced frequency of convective precipitation in the summer months. As such, the Pretest represents a flexible approach to successively integrate data-driven methods when a benchmark—here direct output from NWP—is available. This situation is expected to frequently arise in applications where training data is limited.

The resulting final model is able to outperform a local version of the cNLR model and reduces the mean CRPS of the raw ensemble (depending on the lead time) by up to 4.8%. Forecast quality might be further improved by adding meteorological and additional topographic predictors to more specifically address spatio-temporal variability of precipitation formation.

Acknowledgements A substantial part of the presented research was conducted when Lea Friedli was a master student at the University of Bern and David Ginsbourger was mainly affiliated with the Idiap Research Institute at Martigny. The calculations were performed on UBELIX, the High Performance Computing (HPC) cluster of the University of Bern.

Funding Open access funding provided by University of Lausanne.

References

Aminyavari S, Saghafian B (2019) Probabilistic streamflow forecast based on spatial post-processing of TIGGE precipitation forecasts. Stoch Environ Res Risk Assess 33:1939–1950

Baran S, Horányi A, Nemoda D (2013) Comparison of BMA and EMOS statistical calibration methods for temperature and wind speed ensemble weather prediction. arXiv preprint arXiv:1312.3763

Baran S, Nemoda D (2016) Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting. Environmetrics 27(5):280–292

Brier GW (1950) Verification of forecasts expressed in terms of probability. Mon Weather Rev 78:1–3

Buizza R (2018) Ensemble forecasting and the need for calibration. In: Vannitsem S, Wilks DS, Messner JW (eds) Statistical Postprocessing of Ensemble Forecasts, 1st edn. Elsevier, pp 15–48

Dabernig M, Mayr GJ, Messner JW, Zeileis A (2017) Spatial ensemble post-processing with standardized anomalies. Q J R Meteorol Soc 143(703):909–916

Dawid AP (1984) Present position and potential developments: some personal views statistical theory the prequential approach. J R Stat Soc Ser A Gen 147(2):278–290

Diebold FX, Gunther TA, Tay AS (1998) Evaluating density forecasts with applications to financial risk management. Int Econ Rev 39:863–883

Friederichs P, Wahl S, Buschow S (2018) Postprocessing for Extreme Events. Statistical Postprocessing of Ensemble Forecasts. 1st edn. Elsevier, pp 127–154

Frühwirth-Schnatter S (1996) Recursive residuals and model diagnostics for normal and non-normal state space models. Environ Ecol Stat 3(4):291–309

Gneiting T, Balabdaoui F, Raftery AE (2007) Probabilistic forecasts, calibration and sharpness. J R Stat Soc Ser B Stat Methodol 69(2):243–268

Gneiting T, Raftery AE, Westveld AH, Goldman T (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Mon Weather Rev 133:1098–1118

Gneiting T, Ranjan R (2011) Comparing density forecasts using threshold- and quantile-weighted scoring rules. J Bus Econ Stat 29:411–422

Grimit EP, Gneiting T, Berrocal VJ, Johnson NA (2006) The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. Q J R Meteorol Soc 132:1942–2925

Hamill TM (2001) Interpretation of rank histograms for verifying ensemble forecasts. Mon Weather Rev 129(3):550–560

Hamill TM, Colucci S (1997) Verification of Eta-RSM short-range ensemble forecasts. Mon Weather Rev 125(6):1312–1327

Hamill TM, Hagedorn R, Whitaker JS (2008) Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: precipitation. Mon Weather Rev 136:2620–2632

Hemri S, Scheuerer M, Pappenberger F, Bogner K, Haiden T (2014) Trends in the predictive performance of raw ensemble weather forecasts. Geophys Res Lett 41(24):9197–9205

Khedhaouiria D, Mailhot A, Favre AC (2019) Regional modeling of daily precipitation fields across the Great Lakes region (Canada) using the CFSR reanalysis. Stoch Environ Res Risk Assess. https://doi.org/10.1007/s00477-019-01722-x

Kleiber W, Raftery AE, Baars J, Gneiting T, Mass CF, Grimit E (2011) Locally calibrated probabilistic temperature forecasting using geostatistical model averaging and local Bayesian model averaging. Mon Weather Rev 139(8):2630–2649

Kleiber W, Raftery AE, Gneiting T (2011) Geostatistical model averaging for locally calibrated probabilistic quantitative precipitation forecasting. J Am Stat Assoc 106(496):1291–1303

Lerch S, Baran S (2018) Similarity-based semi-local estimation of EMOS models. arXiv preprint arXiv:1509.03521

Lim M, Hasti T (2013) Learning interactions through hierarchical group-lasso regularization. arXiv preprint arXiv:1308.2719

Matheson JE, Winkler RL (1976) Scoring rules for continuous probability distributions. Manag Sci 22:1087–1096

Messner JW (2018) Ensemble Postprocessing with R. In: Vannitsem S, Wilks DS, Messner JW (eds) Statistical Postprocessing of Ensemble Forecasts, 1st edn. Elsevier, pp 291–321

Messner JW, Mayr GJ, Wilks DS, Zeileis A (2014) Extending extended logistic regression: extended versus separate versus ordered versus censored. Mon Weather Rev 142:3003–3014

Messner JW, Mayr GJ, Zeileis A (2016) Heteroscedastic censored and truncated regression with crch. R J 8(1):173–181

Meteo Schweiz (2018) COSMO-Prognosesystem. https://www.meteoschweiz.admin.ch/home/messund-prognosesysteme/warn-und-prognosesysteme/cosmo-prognosesysteme.html

Raftery AE, Gneiting T, Balabdaoui F, Polakowski M (2005) Using Bayesian model averaging to calibrate forecast ensembles. Mon Weather Rev 133:1155–1174

Scheuerer M (2014) Probabilistic quantitative precipitation forecasting using ensemble model output statistics. Q J R Meteorol Soc 140(680):1086–1096

Scheuerer M, Büermann L (2013) Spatially adaptive post-processing of ensemble forecasts for temperature. J R Stat Soc Ser C Appl Stat 63(3):405–422

Scheuerer M, König G (2014) Gridded, locally calibrated, probabilistic temperature forecasts based on ensemble model output statistics. Q J R Meteorol Soc 140:2582–2590

Schmeits MJ, Kok KJ (2010) A comparison between raw ensemble output, (modified) Bayesian model averaging, and extended logistic regression using ECMWF ensemble precipitation reforecasts. Mon Weather Rev 138(11):4199–4211

Shin Y, Lee Y, Choi J, Park JS (2019) Integration of max-stable processes and Bayesian model averaging to predict extreme climatic events in multi-model ensembles. Stoch Environ Res Risk Assess 33:47–57

Sloughter JML, Raftery AE, Gneiting T, Fraley C (2007) Probabilistic quantitative precipitation forecasting using Bayesian model averaging. Mon Weather Rev 135(9):3209–3220

Thorarinsdottir TL, Gneiting T (2010) Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. J R Stat Soc Ser A Stat Soc 173(2):371–388

Thorarinsdottir TL, Schuhen N (2018) Verification: assessment of calibration and accuracy. In: Vannitsem S, Wilks DS, Messner JW (eds) Statistical postprocessing of ensemble forecasts, 1st edn. Elsevier, pp 155–186

Wilks DS (2011) Forecast verification. Int Geophys 100:301–394

Wilks DS (2018) Univariate ensemble postprocessing. In: Vannitsem S, Wilks DS, Messner JW (eds) Statistical Postprocessing of Ensemble Forecasts, 1st edn. Elsevier, pp 49–89

Wilks DS, Vannitsem S (2018) Uncertain forecasts from deterministic dynamics. In: Vannitsem S, Wilks DS, Messner JW (eds) Statistical postprocessing of ensemble forecasts, 1st edn. Elsevier, pp 1–13