On the Implementation of Postprocessing of Runoff Forecast Ensembles

JON OLAV SKØIEN,^a KONRAD BOGNER,^b PETER SALAMON,^a AND FREDRIK WETTERHALL^c

^a *European Commission, Joint Research Centre, Ispra, Italy*

^b *Swiss Federal Research Institute WSL, Birmensdorf, Switzerland*

^c *European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom*

(Manuscript received 15 January 2021, in final form 3 August 2021)

ABSTRACT: Different postprocessing techniques are frequently employed to improve the outcome of ensemble forecasting models. The main reason is to compensate for biases caused by errors in model structure or initial conditions, and as a correction for under- or overdispersed ensembles. Here we use the ensemble model output statistics method to postprocess the ensemble output from a continental-scale hydrological model, LISFLOOD, as used in the European Flood Awareness System (EFAS). We develop a method for local calibration and interpolation of the postprocessing parameters and compare it with a more traditional global calibration approach for 678 stations in Europe based on long-term observations of runoff and meteorological variables. For the global calibration we also test a reduced model with only a variance inflation factor. Whereas the postprocessing improved the results for the first 1–2 days lead time, the improvement was less for increasing lead times of the verification period. This was the case both for the local and global calibration methods. As the postprocessing is based on assumptions about the distribution of forecast errors, we also present an analysis of the ensemble output that provides some indications of what to expect from the postprocessing.

SIGNIFICANCE STATEMENT: We analyzed different methods for postprocessing of runoff ensemble forecasts, created by passing different meteorological ensemble forecasts through the same hydrological model, for a large number of stations over all Europe. We checked to which degree transformations of the forecasts will affect the postprocessing, and if a method with local calibration and interpolation of the postprocessing parameters would work, and if it could improve the results. Transformation of the data did not improve the results for our dataset. The interpolation of postprocessing parameters in itself works well. However, whereas local calibration was better for some lead times, it can be questioned whether it is worth the extra complications in an operational forecasting system. Postprocessing was less efficient for longer lead times, and we analyzed factors that might affect this.

KEYWORDS: Interpolation schemes; Statistical techniques; Ensembles; Forecasting; Operational forecasting; Probability forecasts/models/distribution

1. Introduction

Ensemble forecasting has a long tradition in meteorological forecasting but has only been widely applied in hydrological forecasting since a few years. Previous studies such as Cloke and Pappenberger (2009) and Wetterhall et al. (2013) showed how the use of ensemble streamflow forecasts has increased during the last years. Alfieri et al. (2014) presented an evaluation of ensemble streamflow predictions in Europe, with a focus on the European Flood Awareness System (EFAS).

The idea of ensemble modeling is that the spread (or variability) within the ensemble should reflect the uncertainty of the forecast. Whereas the variance can be seen as an indication of the reliability, it is usually assumed that the ensembles need postprocessing (Gneiting et al. 2005) and that this is essential also for hydrological variables (Hemri 2018).

Typically, for meteorological and hydrological forecasting, the ensembles will often exhibit both bias and dispersion errors.

Biases occur if the mean of the ensembles is systematically lower or higher than the observation, and dispersion errors occur when the ensemble variance indicates too low or too high confidence in the forecast, compared with the observations. Ideally, the ensemble spread should match the forecast error growth with lead time. It is therefore usually necessary to postprocess the forecasts to improve their usefulness in decision-making. The target of postprocessing is that the future observation should be statistically indistinguishable from any of the postprocessed ensemble members (Wilks 2017).

Two of the most common methods for this are Bayesian model averaging (BMA; Raftery et al. 2005) and ensemble model output statistics (EMOS; Gneiting et al. 2005). The first method creates a weighted average of a probability distribution associated to each ensemble forecast, whereas the second method models the mean as a weighted average of the individual forecasts and the variance as a function of the ensemble variance. Whereas BMA includes the possibility of multimodal postprocessed distribution functions, EMOS gives a single-modal output. However, BMA will often include more parameters and be computationally more demanding. The last aspect is the reason for choosing the EMOS-method in this study, as we are testing different postprocessing methods for a large number of stations. This would similarly be the case for

Denotes content that is immediately available upon publication as open access.

Corresponding author: Jon Olav Skøien, jon.skoien@gmail.com

many more complex methods that have been developed during the last years, summarized by [Vannitsem et al. \(2020\)](#).

Most postprocessing methods, including EMOS, originally evolved in meteorology, where it was common to find a single set of postprocessing weights to be used for an entire region, with a global calibration method. From the postprocessed distributions for each location, samples were drawn to generate a postprocessed ensemble. These were spatially independent, but [Berrocal et al. \(2007\)](#) used the spatial structure of the errors to generate a spatially structured covariance matrix which can be used to generate spatially consistent samples, still using the same set of weights for all locations. [Hamill et al. \(2008\)](#) introduced a method for pooling similar forecasting locations. Later, [Lerch and Baran \(2017\)](#) tested different methods for semi-local calibration, and showed that these outperformed the completely local or regional approaches.

[Hemri et al. \(2013\)](#) developed a BMA-based method for postprocessing runoff forecasts for individual stations, using the methods of [Berrocal et al. \(2007\)](#) for incorporating the correlation between lead times, which is higher for runoff than for meteorological variables. Their method was particularly useful when combining forecasts with different maximum lead times. [Engeland and Steinsland \(2014\)](#), on the other hand, presented a method fitting different weights to each observation location and lead time, but with the same number of forecasts for each lead time.

Our study of ensemble forecasts is related to the EFAS, an operational, pan European flood forecasting system, which is part of the Copernicus Emergency Management Service. EFAS already presents ensemble forecasts of runoff ([Alfieri et al. 2014](#); [Smith et al. 2016](#); [Thielen et al. 2009](#); [Bartholmes et al. 2009](#)). Meteorological ensemble forecasts and deterministic forecasts from different numerical weather prediction models are used as forcing data for a hydrological model, creating a set of runoff ensembles for each grid point. These are currently presented to the user as a set of boxplots, one plot for each ensemble forecast, together with the deterministic forecasts. Flood notifications are sent to partners based on a weighted average from the different forecasts, with fixed weights based on a study by [Bartholmes et al. \(2009\)](#). The method is quite simple, and we want to analyze if the EFAS ensemble forecasts would benefit from postprocessing. Forecasts might have regional differences and we will therefore present a method that allows for spatially variable weights in addition to the method with the fixed global weights. The dataset and modeling region are considerably larger than in the studies described above, rendering the method of [Engeland and Steinsland \(2014\)](#) as computationally not feasible, for example.

Previous applications in hydrology only considered postprocessing of forecasts at the calibration locations (see, e.g., [Bogner and Pappenberger 2011](#)), similar to meteorological applications ([Berrocal et al. 2007](#)).

EFAS is based on a distributed hydrological model, which gives discharge forecasts for all river pixels in the region of interest. However, calibration of the postprocessing parameters

for all river pixels is computationally infeasible. Our first objective is therefore to see how locally calibrated postprocessing parameters can be interpolated with the geostatistical method top-kriging ([Skøien et al. 2006](#)).

EMOS and the regionalization methods assume that the data are normally distributed. Our second objective is therefore to consider the effect of different data transformation methods on forecast skill in general, and for floods in particular. This includes Box-Cox transformations, similar to what [Hemri et al. \(2015\)](#) used when analyzing multivariate postprocessing techniques for probabilistic hydrological forecasting, but also other transformations such as normal score and transformations based on the general extreme value (GEV) distributions.

Section 2 of this paper describes the dataset used for calibration and validation of the methodology. Section 3 describes possible transformations of the data, the EMOS framework, top-kriging, calibration, and how the methods are evaluated. Section 4 shows the results of the calibration and verification. These results are further discussed in section 5, and some additional conclusions given in section 6.

2. Data

The hydrological ensemble forecasts analyzed in this paper come from a distributed hydrological model that has been fed with a combination of meteorological deterministic and ensemble forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF), the German Weather Service (DWD) and the Limited Area Ensemble Prediction System from the Consortium for Small-Scale Modeling (COSMO-LEPS). We use forecasts from a period of 3 years (1 January 2015–31 December 2017). For each day, we use forecasts with 1–10 days lead time (if available), where the forcing meteorological data have been downscaled to the model resolution of $5\text{ km} \times 5\text{ km}$ using a mass-conservative interpolation. An overview of the forecasts is given in Table 1.

The forecasts from COSMO-LEPS do not cover the regions north of 60°N . Catchments north and south of this latitude are therefore treated separately in the interpolation. To avoid border issues, we have not used forecasts based on COSMO-LEPS for catchments that are completely or partly north of 59.5°N .

Each individual forecast member is used as input to the hydrological model LISFLOOD ([De Roo et al. 2000](#); [Van Der Knijff et al. 2010](#)), giving an ensemble of forecasted discharge hydrographs for each forecast day. LISFLOOD is a gridded model that predicts the discharge for each individual pixel on a $5\text{ km} \times 5\text{ km}$ grid, and the meteorological forecast variables are downscaled to the same grid resolution. The model has a daily temporal resolution, although some model components use intermediate internal time steps to avoid numerical issues. The extent of the forecasts and the hydrological model in current operational EFAS includes all of Europe with the eastern border at the Caspian Sea, and with Asian and African catchments that drain to the Mediterranean Sea.

TABLE 1. Overview of the meteorological forecasts used in this study.

	Forecasts	Resolution	Time step	Lead time	Reference
ECMWF deterministic	1	9 km	6-hourly	10 days	Holm et al. (2016)
ECMWF ensembles	51	18 km	Daily ^a	15 days (using day 1–10)	Holm et al. (2016)
DWD deterministic	1	6.5 km	6-hourly	5 days	DWD (2017)
		13 km		Day 6–7	
COSMO-LEPS ensembles	16	7 km	6-hourly	5.5 days (using day 1–5)	Marsigli et al. (2005)

^a Changed to 6-hourly in EFAS v4.0, which was launched in October 2020.

We use simulated runoff at 678 stations with the meteorological forcing from interpolated observed precipitation and temperature as a proxy for the observations, using the same model setup of LISFLOOD as for the forecasts. The postprocessing methodology is then calibrated and verified at the same stations as used for the hydrological model calibration.

Figure 1 shows the location of the stations that have been used for calibration, with the color indicating the upstream catchment area for each location. Some areas are sparsely covered by station locations due to unavailability of streamflow records for calibration.

We use runoff simulated from observed meteorological data because EFAS is focusing on return periods of floods, rather than the observed runoff itself, to neglect biases

resulting from the hydrological model error misspecifications. Thus, the bias will come mostly from the deviation between the calibration based on meteorological observations and the forecast based on meteorological forecasts. For a better comparability, all discharge data are given in millimeters per day (mm day^{-1}).

3. Method

The postprocessing of hydrological ensemble forecasts with EMOS contains several steps, depending on which method to use. In this method section we first describe data transformations, and then parameter calibration and incorporation of spatial dependencies, before analyses and validation of the results.

FIG. 1. Overview of locations used for calibration of EMOS parameters. The color scale reflects upstream catchment area (km^2). Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL.

TABLE 2. The different data transformations applied in this paper.

Number	Name	Description and comments
1	Untransformed	$X_t = X$
2	Normal score transform	$X_t = \text{NST}(X)$ Uses rank matching to achieve a dataset closely resembling the normal distribution. A limitation is that the back transformation is not able to create values outside the original range.
3	Box–Cox	$X_t = (x^\lambda - 1)/\lambda$ The standard Box–Cox transformation, with λ fitted for each time series
4	Box–Cox with $\lambda = -0.3$	Box–Cox transformation with the average λ from all time series
5	Lognormal transformation	$X_t = \ln(X)$ Lognormal transformation, can be seen as a special case of the Box–Cox transformation with $\lambda = 0$
6	Box–Cox with $\lambda \geq 0$	Box–Cox transformation with λ fitted for each time series, but not allowing for negative values, setting them to 0
7	General extreme value distribution	$X_t = Q_N[F_{\text{GEV}}(X)]$, where Q_N is the normal quantile function [transforming from probabilities in the range $(0, 1)$ to an $N(0, 1)$ distribution] and $F_{\text{GEV}}(X) = \exp(-\{1 + \zeta[(x - u)/\sigma]\}^{-1/\zeta})$ is the cumulative distribution function for the extreme value distribution, where ζ is a shape parameter, σ is a scale parameter, and u is a location parameter. Scale and shape were fitted, whereas the location parameter was set to 0 as it was challenging to fit this automatically.
8	General extreme value distribution with shape $\zeta = -0.25$	GEV with fixed shape. No fitting necessary for the scale parameter σ , as it can be found from the data: $\sigma = \text{mean}(\text{obs})/0.37$.
9	Box–Cox with calibrated lambda	The Box–Cox parameter λ is treated as a spatially correlated calibration parameter and fitted during calibration rather than being estimated from the time series.

a. Data transformation

The methodologies for postprocessing and regionalization mainly assume normally distributed data, which is usually not the case for runoff. Hemri et al. (2013) used a Box–Cox transformation in their application. We have tested a range of transformation methods here, listed in Table 2. Some of the transformation parameters are fitted locally and might also have to be interpolated for uncalibrated locations.

b. Postprocessing with ensemble model output statistics

Ensemble forecasts typically show a positive spread–error correlation, but the ensembles might still be uncalibrated and wrongly dispersed, most commonly underdispersed (Gneiting et al. 2005). A commonly used method is the EMOS method (Gneiting et al. 2005), which addresses both issues through minimization of the continuous ranked probability score (CRPS).

The description here is a short version of Gneiting et al. (2005). For the forecast issued at time t , we define the postprocessed mean forecast as a weighted average of the ensemble members for each calibration location i and lead time l :

$$Y_{il} = a_{il} + b_{il1}X_{il1} + b_{il2}X_{il2} + \dots + b_{ilN}X_{ilN} + e_{il}. \quad (1)$$

Here a_{il} and $b_{il1}, \dots, b_{ilN}$ are coefficients (i.e., weights), e_{il} is an error term with an average of zero, and $X_{il1}, \dots, X_{ilN}$ are the individual ensemble forecast members. This model is only valid for nonexchangeable ensemble members. In our case, we have a combination of deterministic and ensemble forecasts. We use the deterministic forecasts and the means of the ensemble forecasts as the individual ensemble members (X_{ilj}) in Eq. (1), so that N will be four for 1–5 days, three for 6–7 days, and two for 8–10 days lead time. As the individual members for

each ensemble are interchangeable (they come from the same distribution), we cannot estimate separate weights for each of them.

Optimizing Eq. (1) will give an unbiased mean forecast, at least for the in-sample data. To optimize the variance of the postprocessed forecast e_{il} , we assume, as Gneiting et al. (2005), that the variance can be modeled as a linear function of the raw ensemble variance:

$$\text{Var}(e_{il}) = \sigma_{il}^2 = c_{il} + d_{il}S_{il}^2, \quad (2)$$

where S_{il}^2 is the raw ensemble variance, c_{il} and d_{il} are non-negative coefficients. The variance of the raw ensemble is the variance of a super ensemble consisting of all ensembles and deterministic forecasts. We then assume that the postprocessed forecast has a Gaussian predictive distribution F :

$$F(Y_{il}) = N(a_{il} + b_{il1}X_{il1} + b_{il2}X_{il2} + \dots + b_{ilN}X_{ilN}, c_{il} + d_{il}S_{il}^2). \quad (3)$$

Gneiting et al. (2005) suggested to use the continuous ranked probability score (CRPS) as scoring rule. The CRPS is equal to the mean absolute error (MAE) for a deterministic forecast (Hersbach 2000) and can be seen as a generalization of the MAE. If F_{il} is the predictive distribution function and $y_{\text{obs},t+l,i}$ is the observation, the CRPS for a single observation is defined as

$$\text{crps}(F_{il}, y_{\text{obs},t+l,i}) = \int_{-\infty}^{\infty} [F_{il}(w) - H(w - y_{\text{obs},t+l,i})]^2 dw, \quad (4)$$

where $H(w - y_{\text{obs},t+l,i})$ is the Heaviside function, which is 0 for $w < y_{\text{obs},t+l,i}$ and 1 for $w \geq y_{\text{obs},t+l,i}$. If F_{il} is the CDF of a

normally distributed variable with mean μ_{iil} and variance σ_{iil}^2 , the integral can be replaced with

$$\text{crps}[N(Y_{iil}, \sigma_{iil}^2), y_{\text{obs}_t+l,i}] = \sigma_{iil} \left\{ z[2\Phi(z) - 1] + 2\phi(z) - \frac{1}{\sqrt{\pi}} \right\}, \quad (5)$$

where $z = (y_{\text{obs}_t+l,i} - y_{iil})/\sigma_{iil}$ is the normalized prediction error and $\Phi(z)$ and $\phi(z)$ represents the CDF and the PDF of a $N(0, 1)$ distribution. The CRPS is negatively oriented, i.e., low for precise forecasts with a small variance, and high for less precise forecasts with a higher variance. For each calibration day, it is possible to show through numerical analyses that the function will give a minimum error when $\sigma_{iil} \approx 1.2|y_{\text{obs}_t+l,i} - y_{iil}|$.

If the data are transformed with one of the methods from section 3a, Eq. (5) is not applicable. There are some other cases where analytical distributions can be found (Jordan et al. 2019), but we have treated all of the cases equal and introduced the transformation in Eq. (4):

$$\text{crps}(F_{iil}, y_{\text{obs}_t+l,i}) = \int_{-\infty}^{\infty} [F_{iil}[T(w)] - H(w - y_{\text{obs}_t+l,i})]^2 dw, \quad (6)$$

where T is the transformation function. The numerical integration of Eq. (6) is computationally much slower than application of Eq. (5). Numerical errors might also occur, as the integration has to be replaced by a summation with automatically identified upper and lower ranges, and increments of w .

Some cases where Eq. (6) is not well defined can be identified. In particular, this can happen for the normal quantile transformation (Krzysztofowicz 1997), where forecast values of $w > \max(y_{\text{obs}})$ are all transformed to $w_t = T[\max(y_{\text{obs}})] = \Phi^{-1}[1 - 1/(n + 1)]$ where n is the number of observations used for the normal score transform, and Φ^{-1} is the quantile function for a $N(0, 1)$ distribution (Bogner et al. 2012). Furthermore, if the forecasted value is close to or above $\max(y_{\text{obs}})$, and $\text{Var}(e_{iil}) > 0$, $F_{iil}[T(w)]$ will not asymptotically reach 1, a requirement for this integral. The importance of this issue will decrease with longer time series, when it becomes more likely that the forecasted value is within the previously observed values. For some of the other transformations, certain parameter combinations can give a transformed predictive distribution that is partly outside the range of permissible values.

Equations (4)–(6) can be seen as objective functions when summed over all time steps t used in the calibration, for a certain lead time l . The optimal set of EMOS parameters can be found by minimizing this sum.

c. Top-kriging for spatial correlation and interpolation

Top-kriging (Skøien et al. 2006, 2014) is a geostatistical method for interpolation when the observations have varying spatial support. The method is applied to variables that can be seen as spatial averages, such as specific runoff with the catchment as spatial support. The interpolation method considers both spatial proximity and network topology. Catchments which are stream connected will have higher correlations than non-connected, for the same distance between the centers of gravity.

The parameters are interpolated for each location in a leave-one-out cross-validation procedure to assess the postprocessing skill for noncalibrated locations. To assess the interpolation skill itself the coefficient of determination (R^2) between each calibrated parameter and interpolated parameter is computed:

$$R^2 = 1 - \frac{\sum_i (p_{oi} - p_{mi})^2}{\sum_i (p_{oi} - \bar{p}_o)^2}, \quad (7)$$

where p_{oi} is the “observation” (the calibrated EMOS parameter) for location i , p_{mi} is the interpolated parameter, and $\bar{p}_o$ is the mean of the calibrated parameter from all calibration locations. Here, $R^2 = 1$ when all parameters are perfectly interpolated, $R^2 < 0$ means that the mean of the particular parameter would be a better estimate than the interpolated parameters.

The top-kriging method can be used for interpolation of postprocessing parameters to uncalibrated locations. Each parameter is interpolated separately and cross-correlation between the parameters is not taken into account. However, here the sum of weights is interpolated, and the weights are scaled in the validation to match the interpolated sum. In addition, the lambda parameter from the Box–Cox transformation and the shape and scale from the GEV transformation are interpolated to uncalibrated locations in the same manner. The idea of interpolating parameters has some similarities with the procedure described by Scheuerer and Büermann (2014), but we use spatial support instead of point support, and no trend functions.

d. Calibration of EMOS parameters and including a spatial penalty

The traditional approach is to calibrate one set of EMOS parameters for the entire region, denoted here as global calibration. Different forecasting centers might have different skills for different regions, and thus a method for local calibration of the parameters is presented and analyzed in addition, where the EMOS parameters are calibrated separately for each calibration location. Such a local calibration should give equal or better results than the global calibration. A simplified global calibration is tested where only the variance inflation factor [d_{iil} in Eq. (2)] is fitted, below referred to as VIF calibration.

Whereas the global calibration is done by minimizing Eq. (4) for all catchments jointly, the local calibration is more complex. Contrary to the methods of Hamill et al. (2008) and Lerch and Baran (2017) the data from observation locations are not pooled. Instead, the local variability of the calibrated parameters is constrained. A first fitting of the EMOS parameters for all catchments is done by minimizing Eq. (4) above for each of the catchments. A traditional optimization problem is that different combinations of parameters can often reach similar values in the objective function, giving a spatially inconsistent set of parameters. In a second step, a spatial penalty to the objective function is added that is increasing with the difference between the parameters of the calibration catchment and the parameters of the surrounding 10 catchments.

TABLE 3. Overview of the different validation cases.

Case	Description	Calibration year	Validation year	Spatially interpolated postprocessing parameters
1	Calibration	2017	2017	No
2	Spatial cross validation	2017	2017	Yes
3	Temporal validation	2016	2017	No
4	Spatiotemporal cross validation	2016	2017	Yes

As parameters of neighboring catchments will change, recalibration is performed in an iterative procedure.

The spatial penalty is based on a covariance matrix for all observation catchments, created with the top-kriging method and based on a sample variogram of daily runoff. For a calibration catchment (index 0 below) the $n = 10$ neighbors with the lowest semivariance from the covariance matrix are found. The spatial penalty between the calibration catchment and the neighboring catchments is computed for all parameters $j = 1, \dots, m$, as a weighted norm of their differences:

$$S_{\text{pen},il} = C_{il} \sum_{j=1}^n w_{ij} \sum_{k=1}^m |p_{0kl} - p_{jkl}|, \quad (8)$$

where p_{0kl} is parameter k of the calibration catchment (lead time l) and p_{jkl} is parameter k of neighboring catchment j . The term w_{ij} is a weight, $w_{ij} = 1/\gamma_{0j}$, where γ_{0j} is the semivariance between catchments 0 and j . The term C_{il} is a coefficient that is used to set $S_{\text{pen},il}$ equal to the CRPS after the first iteration, to let the two values be of the same magnitude. This means $C_{il} = S_{\text{pen},il}/S_{\text{crps},il}$, where $S_{\text{crps},il} = \sum_{t=1}^T \text{crps}(F_{il}, y_{\text{obs},l+i,i})$. The spatial penalty is then added to the CRPS in the objective function. We use the absolute differences as all parameters are of comparable magnitude. C_{il} is kept constant during the iterations.

In addition, a small penalty for negative weights [b_{il} in Eq. (3)] is included, equal to the sum of the norm of the negative weights multiplied with the CRPS. Allowing negative weights means that predictions are not limited to the range of the observations. Our procedure allows small negative weights if they are really improving the forecasts, while avoiding large negative weights compensating for large positive weights. This could be the case for two forecasts that are highly correlated, and where the sum of their weights is more important than their individual weights. Such weights are likely to give very poor results when forecasting events where the forecasts are diverging.

Hemri et al. (2013) showed the usefulness of calibrating all lead times together. This would have been computationally demanding for the cases with transformations, whereas it is easier for the other cases. For consistency, we only present results where lead times were calibrated independently. However, analyses not presented here show that the effect of joint calibration has a relatively small effect on the comparisons presented in this study. It is more important for consistency between lead times, something not analyzed here.

The parameters were refitted in an iterative procedure until the sum of the incremental change of parameter values and the sum of the objective functions for all catchments go below a

threshold. We used parameters from the same iteration if these have already been estimated for a catchment, and parameters from the previous iteration for the remaining catchments. The order of the catchments was chosen randomly for each iteration.

It is common to use calibration periods of 30–60 days for estimation of EMOS parameters, but we chose one year in our case, to make it likely that some high values were included in the calibration period. This is consistent with results from Lang et al. (2020), who analyzed the effect of different training periods on the forecasts of temperature at measurement sites located across Austria, Germany, and Switzerland. They showed that long training periods yields an improved performance compared with the traditional shorter sliding window approaches. Similarly, Hemri (2014) found that calibration periods of 1–5 years were optimal for different meteorological variables.

e. Analyses of calibration results and validation

The dataset was split in three parts, where each part has the simulation and forecast data from one calendar year (2015–17). The parameters were calibrated for each year separately and the third year (2017) was used for verification. We compared the CRPS for four different cases, shown in Table 3. Case 1 gives the calibration result, to see differences in the calibration between methods. Case 2 is similar to the leave-one-out cross validation as traditionally done in geostatistics. This case helps us to identify the sources of errors for case 4. Case 3 is the temporal verification, traditionally the most common method for verification of postprocessing methods. Case 4 shows the effect of using spatially interpolated parameters from a different time period, a spatiotemporal cross validation. This gives an indication of the quality of the new method presented in this study, where postprocessing parameters can be interpolated to locations where we have not calibrated.

1) GOODNESS-OF-FIT FOR CALIBRATION AND VALIDATION

As the main indicator of goodness-of-fit for a method, we used the continuous ranked probability skill score (CRPSS), calculated for each individual catchment and lead time:

$$\text{CRPSS}_{it} = 1 - \frac{\text{crps}_{it}}{\text{crps}_{ib}}, \quad (9)$$

where the subscript t refers to a particular transformation and b to the benchmark. This is an indicator of the quality of the postprocessed forecast compared with the benchmark method. It gives a value in the range 0–1 when the method is better than the

benchmark, where 1 would indicate a perfect forecast. For CRPSS < 0, the benchmark method performs better than the method.

Benchmarking, i.e., what to compare with, is an important choice in evaluating forecasts (Pappenberger et al. 2015). Whereas many postprocessing studies have focused on the forecasting skill itself, compared to the climatological mean (Hemri et al. 2015; Engeland and Steinsland 2014), we instead compared with the raw ensemble, i.e., the mean of the forecasts (with the ensemble forecasts averaged) and the variance of a super ensemble with all ensemble members and the deterministic forecasts.

Additionally, we used the probability integral transformation (PIT) as another measure for the quality of the postprocessing (Laio and Tamea 2007). Here we use it as a visual indicator to which degree the distribution of the observations is reflected in the raw forecast or the postprocessed forecast. For all observations, we can define a new variable w_{til} for the time t , location i , and lead time l as

$$w_{til} = F_{til}(y_{\text{obs},ti+l})$$

based on the value of the cumulative distribution function for the observation value. This variable should ideally be uniformly distributed $U(0, 1)$ when we include a large number of predictions and observations. Although this is rarely achievable, we would expect the distribution to be closer to uniform after postprocessing.

f. Implementation of analyses

Some preprocessing of maps was done with PCraster (Karssen et al. 2010), otherwise all analyses were carried out in the statistical environment R (R Core Team 2019). The transformations used standard functionality from base packages or from the package extRemes (Gilleland and Katz 2016). The calibration of EMOS parameters combined the global optimization method Shuffle Complex Evolution method (SCEUA; Duan et al. 1992) implemented in the sceua function in rtop (Skøien et al. 2014) and the local optimization method by Nelder and Mead (1965) implemented in the optim function from base R. SCEUA was used to reduce the risk of being stuck in local optima, for iterations 1–5, 8, 11, and 14. Parallel computation was used to reduce computation time.

The time to calibrate a single catchment depends on how many iterations are needed inside the optimization function to reach an optimum, which will be different for different catchments and transformations. Some transformations also take more time than others. The global SCEUA routine requires more function evaluations than the local optimization, which is why the use of SCEUA is restricted to some of the iterations. It was typically necessary with 20–25 iterations for the process to converge. Using on average 15 nodes on a high-performance computing cluster, the entire calibration process took around one week for all lead times and transformations. Calibration time for untransformed data [using Eqs. (5) instead of (6)] was on the order of 2–3 days on a single processor, which could be reduced considerably by multiprocessing.

4. Results

In this section we will first present the results from local calibration validation of the EMOS parameters with different

transformations. Some of these results will then be compared with the global calibration before some further analyses to explain the observed results.

a. Fitting of EMOS parameters and effect of spatial penalty

Figure 2 shows how the spatial penalty affects the CRPS while the spatial correlation of parameters between the neighboring catchments is increased, comparing with the results from the first iteration (untransformed data with lead time 1 day). The results are similar for all transformations and lead times. The CRPS (central panel) generally increases when the spatial penalty is included in the calibration, but for some catchments it even declines, indicating that the first iteration found a local optimum. The maximum increase of the CRPS when calibrating with the spatial penalty is 36%, but only for 9% of the catchments the CRPS increases more than 10%. The CRPS is reduced for 16% of the catchments. These numbers are of similar magnitude for most other lead times, and for most transformations. There are some exceptions, where a much larger increase in CRPS is found for 1–2 catchments.

The spatial penalty is strongly reduced during calibration as seen in the left panel. It is reduced to less than half from the first to the last iteration for 98% of the catchments, and to less than a quarter for 90% of the catchments. These numbers are also of the same magnitude for all lead times and transformations.

The last panel shows that the CRPS is dominating the value of the objective function for all catchments after the end of the calibration. There are very few catchments for other lead times and transformations where the spatial penalty is slightly higher than the CRPS. We can also notice the large range in the values. Both, CRPS and spatial penalties are mostly in the range from 0.001 to 0.1 mm day⁻¹, with some smaller spatial penalties and some larger CRPS.

Figure 3 shows the calibrated weights [$b_{til1}-b_{tilN}$ in Eq. (1)] from the first (left) and last (right) iterations. We can see how the weights have a high spatial variability after the first iteration, whereas the spatial continuity is higher after the last iteration.

The sum of the weights is close to 1 for the majority of the catchments, but can be considerably higher or lower in some cases. These are catchments where the forecasts are either highly biased (particularly where the sum of the weights is high), or where the timing of flood peaks is wrong (often the case for small weights). In some cases, catchments with these types of weights will have a large negative or positive constant value [a_{til} in Eq. (1)] that compensates.

We can see that the weights for the mean of the ECMWF ensemble are very low, except for the Iberian Peninsula. The weights for the deterministic ECMWF are higher, mainly in Scandinavia and in Italy and the Balkans. The weights are close to zero in France. The weights for DWD are high in Scandinavia, and mostly around 0.2–0.4 for the rest of Europe. The weights for COSMO-LEPS are relatively high for most of Europe, except for the Balkans.

b. Interpolation of parameters

Each variable for each lead time and each transformation was interpolated separately and Table 4 shows the average

FIG. 2. Comparison of (left) spatial penalty and (center) CRPS from first to last iteration for untransformed data and lead time of 1 day (2017). (right) Comparison of CRPS and spatial penalty after last iteration. Solid line represents the 1:1 ratio, whereas dashed lines represent 2:1 and 1:2 and dotted lines 5:1 and 1:5. Units are mm day^{-1} .

coefficient of determination (R^2) for all lead times between the calibrated EMOS parameters and the interpolated parameters for different transformations for 2017. Columns *c* and *d* are the variance inflation factor and variance bias of Eq. (2), whereas the next five columns are the intercept and the weights for each forecast in the regression in Eq. (1). Each row represents a different transformation.

The weight parameters are relatively easy to interpolate, with R^2 mostly above 0.9. This is also the case for the bias parameter for the untransformed data, whereas R^2 is lower for this parameter for the other transformations. For the variance parameters, R^2 is high for *d*, whereas it is slightly lower for *c*, except for the untransformed data.

The interpolation results for the spatially fitted lambda (transformation 9 in Table 2) is just marginally better than for the locally estimated lambda (transformation 3). The scale parameter for the GEV with fixed shape is the only transformation parameter that could be interpolated with R^2 above 0.6.

The average R^2 was highest for untransformed data with 0.94, versus 0.92 for log-transformed data.

c. CRPS results for local calibration

Figure 4 shows the scatterplot of the CRPS from the local calibration versus the CRPS from calibration with spatial penalty (same as the central panel of Fig. 2), cross validation, verification, and temporal cross validation, for 1-day lead time and untransformed data, as described in section 5. The CRPS values increase but are mostly below the dashed line that represents two times the CRPS of the calibration for the cross-validation case (top right). The CRPS increases with less than 7% for half of the catchments. The increase is less than 64% for 97% of the catchments. The verification case (bottom right), which represents the use of previously calibrated parameters, shows that the CRPS increase less than 6% for half of the

catchments. For 99% of the catchments, the CRPS increases less than 45%.

The temporal cross-validation case (bottom right) indicates the ability of the method to forecast runoff at uncalibrated locations. Although the results are slightly poorer than for the two other cases, the difference to the temporal verification case is not large. For almost 20% of the catchments, the results were actually better than using the temporal verification parameters. For 80% of the catchments, the CRPS increased less than 20% compared with the calibration case, and less than 90% for 97% of the catchments.

The CRPS generally increases more for transformed data than for untransformed data. For longer lead times, the spatial cross validation on average gives smaller increase in CRPS than using the parameters from the year before. The spatiotemporal cross validation exhibits marginally larger increases than temporal verification, except for a few catchments.

The results for other transformations show a similar pattern, although more extreme increases of CRPS are seen for some catchments and transformations. Poorer results are particularly seen for transformations with parameters to be interpolated.

d. Comparison of transformations

Figure 5 shows the median CRPSS (based on the individual CRPSS values for each catchment) for the different transformations as a function of lead time for the different cases described in section 3e. The numbers in the figures refer to the transformations in Table 2. We use the median because very low values for some catchments could strongly impact the mean.

The median CRPSS for the calibration case is positive for all transformations and lead times, although it decreases with increasing lead time, particularly from 1 to 4 days. There are a

FIG. 3. The weights of the different forecasts after first and last iteration: b_1 (DWD), b_2 (ECMWF deterministic), b_3 (ECMWF mean of ensembles), and b_4 (COSMO-LEPS) for 1-day lead time and untransformed data from 2017. Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL.

few individual catchments that have negative CRPSS also for the calibration case for some transformations and lead times, from the use of the spatial penalty in the calibration procedure. The transformation with fitted lambda is best for the first 4 days

lead time, but worse for 8 days and longer lead times. The ordinary GEV transform, the normal score transform, and the untransformed data generally give the lowest CRPSS during calibration.

TABLE 4. The coefficient of determination (R^2) between the locally estimated EMOS parameters and the interpolated parameters for 1-day lead time for 2017. The terms c and d are the parameters of the variance equation Eq. (2), whereas the remaining columns are the intercept and the weights for the regression in Eq. (1). Italic numbers represent R^2 values below 0.5, and bold numbers represent R^2 values above 0.8.

Transformation	c	d	A	Deterministic	Deterministic	Ensemble mean ECMWF	Ensemble mean COSMO	Lambda	Scale	Shape	Average
				DWD	ECMWF						
1	0.85	0.96	0.92	0.97	0.98	0.97	0.97				0.94
2	0.79	0.91	0.85	0.95	0.97	0.98	0.93				0.89
3	0.53	0.80	<i>0.15</i>	0.95	0.95	0.93	0.83	<i>0.49</i>			0.71
4	<i>0.41</i>	0.89	<i>0.42</i>	0.97	0.94	0.97	0.95				0.79
5	0.82	0.97	0.83	0.97	0.98	0.97	0.97				0.92
6	0.79	0.97	0.83	0.97	0.98	0.97	0.96	<i>0.49</i>			0.87
7	0.72	0.91	0.82	0.95	0.98	0.97	0.93		0.59	<i>0.34</i>	0.80
8	0.79	0.96	0.89	0.97	0.98	0.97	0.97		0.77		0.91
9	0.88	0.94	0.87	0.94	0.96	0.91	0.78	0.50			0.85

The results from the spatial cross validation are quite similar to the calibration case, only that the CRPSS values are lower. The two GEV transformations give the poorest results, with negative median CRPSS for some of the lead times.

The lower-left panel shows that CRPSS deteriorates more for the temporal verification case. Both GEV and normal score transforms give negative median CRPSS for more than 4 days lead time. Also, for the other methods this is the case for some of the catchments.

The spatiotemporal verification gives slightly poorer results than the temporal verification. The ordinary Box–Cox transformation performs best. Most of the transformations have median CRPSS > 0 for most lead times; the exception is the N score and GEV transformations, which have negative median CRPSS for most lead times. The best transformations are the other Box–Cox transformations. The median CRPSS for the untransformed data is slightly lower than for the transformed data, but the difference is small.

e. Global calibration

Figure 6 shows three quantiles (10th, 50th, and 90th) of CRPSS values for all the catchments, for calibration (left) and validation (right) from fitting the parameters globally, compared with the results from the local calibration of untransformed data (case 3 in Table 3). These show similar patterns for the median CRPSS as the untransformed cases in Fig. 5. Both for calibration and validation the CRPSS is highest for 1-day lead time, lowest for 5–6 days, and then slightly increasing for the longest lead times, except for when only calibrating the variance inflation factor (VIF calibration) globally. The median CRPSS is above zero for most lead times for all methods.

We can see that the method with locally calibrated parameters perform substantially better for all percentiles for the calibration. We can also notice that the 10th percentile is considerably below zero for the globally calibrated parameters.

For the validation case, the median CRPSS of the local calibration is better than or equal to the other methods for all lead times. The VIF calibration is slightly better than global calibration of all parameters for short lead times, and slightly worse for longer lead times. However, VIF calibration has fewer cases with really poor results, i.e., the 10th percentile is higher than for the other methods (except for local calibration

for the first 2 days lead time). The 10th percentile for global calibration of all parameters is poorer than the others for the first 2 days lead time, and similar to local calibration for the remaining lead times. For the 90th percentile, the locally calibrated parameters clearly give the best results, whereas the poorest result is from VIF calibration. From Fig. 5 we recall that the CRPSS for the validation case of the locally fitted parameters was slightly higher than the CRPSS from the spatiotemporal validation for all quantiles.

f. PIT analysis

Figure 7 shows a histogram plot of all the probability integral transformed values for the calibration and validation case for the three methods for lead times 1 (left) and 7 (right) days. The bottom panel shows the PIT values for the raw ensembles for the different lead times. The high frequencies for the smallest and highest quantiles mean that the variance of raw ensemble variance is underestimating the real uncertainty of the forecast, it is underdispersed. After calibration, the PIT values are much closer to uniform distribution for 1 day lead time, indicating a more correctly dispersed forecast. This is not the case for 7 days lead time, where the raw ensemble also has an additional peak for the quantile around 0.4. The PIT values for the calibration and validation cases are only marginally more uniform than for the raw ensemble, which then partly explains the low CRPSS values for the longer lead times.

g. Analysis of calibrated EMOS parameters

Figure 8 shows an overview and comparison of the different parameters for the three methods as a function of lead time to investigate the behavior of the CRPSS and PIT plots above. The locally calibrated parameters are shown as black boxplots. The globally calibrated EMOS parameters are shown either as red lines or red boxplots (slightly shifted from the boxplots of the locally calibrated parameters), whereas the globally calibrated inflation factor is shown in blue.

The top-left panel shows the variance inflation factor, d_{ii} in Eq. (2); $d_{ii} = 1$ gives the variance of the raw ensemble (if the variance bias correction is zero). We can see that the calibration substantially increases the variance for short lead times for

FIG. 4. Comparison of CRPS from calibration and validation cases for untransformed data with 1-day lead time. The calibration CRPS is on the x axis of all panels. (top left) Calibration with spatial penalty. (top right) Spatial cross validation (calibration year and verification year the same). (bottom left) Temporal validation. (bottom right) Spatiotemporal cross validation. Solid line represents 1:1, whereas dashed lines represent 2:1 and 1:2 and dotted lines 5:1 and 1:5.

all methods, d_{ii} > 2 for 1-day lead time, decreasing toward 1 for a lead time of 5 days. The calibrated variance inflation parameters are similar for the different methods, with the globally calibrated parameters close to the median of the locally calibrated parameters. For longer lead times, the variance inflation parameters are even a bit below one.

The top-right panel shows the constant that is added to the variance, the variance bias correction, normalized with the 90th percentile of the daily standard deviation of the raw ensembles for each station. This parameter is shown as a boxplot also for the globally calibrated variance bias correction, as the relative impact on each catchment depends on the variability of the raw ensembles. This will then give an indication of the magnitude of the “fixed” variance added to the ensembles. We see that this is very small for most catchments. And as the values for standard

deviation are rather skewed, the fixed variance will usually be even lower relative to the typical variance during a flood event. The relative variance bias correction parameter has a median between 0.03 (lead time 1 day) and 0.07 (lead time 6 days) for the locally calibrated parameters. The median for the globally calibrated parameters is slightly higher for short lead times, and slightly lower for longer lead times.

The panel on the bottom left shows the bias correction factor (a_{ii}) in Eq. (1), normalized with the mean flow of the calibration period. We see that also this correction is relatively small for most catchments. The median for the locally calibrated parameters is slightly negative for the shorter lead times (1–4 days) and slightly above zero for 8–10 days. The globally calibrated parameters are negative for all lead times. The biases are in general small for all catchments, with a small effect for larger floods.

FIG. 5. Lines of median CRPSS for the different cases and for the different transformations as a function of lead time for 2017. The numbers refer to the different transformations in Table 2. Further details about the different cases are in Table 3. The raw forecast was used as reference forecast.

The last panel shows the sum of the weights for the forecasts in Eq. (1). These are very close to one for most catchments and lead times. However, the median sum is slightly above one for the lead times with negative bias correction, and slightly below zero for the lead times with positive bias correction for the locally calibrated parameters. For the globally calibrated parameters, the sum is constantly above 1, corresponding to the negative bias correction in the bottom-left panel.

These patterns are quite consistent also for the other years (not shown). Figure 9 shows the calibrated weights for each forecast, as a function of lead time for the locally calibrated EMOS parameters (solid lines) and the globally calibrated parameters (dashed line) for each year. There are some differences between the globally and locally calibrated parameters, but overall they are rather similar. For short lead times, the mean of the ECMWF ensembles has very low weights, whereas its influence increases with lead time, and has higher

weights than the deterministic ECMWF forecasts for long lead times. For the first 5–7 days of the forecast, the deterministic ECMWF forecast and the mean of the COSMO-LEPS ensemble gets the highest weights for 2015, the forecast from DWD gets the highest weights for 2016, whereas the forecasts from COSMO-LEPS gets the highest weights for 2017. This is another indication that the weights from the calibration period might not be so good for the validation period, there is a higher correlation between 2015 and 2017 than between 2016 and 2017.

5. Discussion

The necessity for postprocessing ensemble forecasts is often taken for granted. Here we have developed a new method for local calibration and regionalization of postprocessing parameters, and compared this with the traditional methods for postprocessing. We have also analyzed why the methods in some cases do not give the expected improvements.

FIG. 6. Quantiles of CRPSS as a function of lead time for local calibration, global calibration and from calibration of the variance inflation factor. The raw forecast was used as reference forecast. Local calibration refers to case 3 in Table 3. Calibration year is 2016 and validation year 2017.

a. Local calibration of parameters

The local calibration of postprocessing parameters and the spatial regionalization of these works well for most cases. A few catchments see an increase in CRPS after introducing spatial penalty, mostly because they have different characteristics from the surrounding catchments. Locations where regulation is dominant has been excluded, so this should not be the cause. There is a small decrease in skill between the calibration case and the cross-validation case, but the number of catchments with negative skill is low. Although this result is not relevant for the forecast case, it shows that the parameters can be spatially interpolated with the top-kriging method.

The use of parameters from a previous calibration period is more challenging. The temporal verification gives a skill that is closer to zero for more of the transformations and lead times, and declines more than for the spatial regionalization. It should be mentioned that the temporal verification has fewer catchments with strongly negative CRPSS (< -0.5) than the spatial interpolation.

We assumed that different forecasts have different skills in different regions, and that this advantage is preserved from the calibration to the verification period. The calibration with a spatial penalty did improve the spatial continuity of parameters, but the comparison between the years showed that the temporal continuity was low. This was also confirmed by the results of the temporal verification. We have tested different calibration periods for

untransformed data (not shown), and the results were similar.

The results of the spatiotemporal cross-validation have a decline in skill compared to the calibration that is similar to the sum of the declines of the two methods above. There is a skill for the first lead times for all transformations, except for GEV and N score (only 1 day). Some of the transformations also have a small positive median skill for 8–10 days, whereas none of the transformations have a positive skill for 4–7 days lead time. This means that the benchmark is equally good for these lead times.

b. Transformations

Whereas most of the transformation methods gave better results for the calibration case, the differences were smaller for the verification cases, and several transformations performed worse than using the untransformed data.

The N -score transformation, the normal Box-Cox transformation, and the GEV transformations generally performed rather poorly. The results were better for other types of Box-Cox transformations, so it seems better to limit the use of locally estimated transformation parameters. The Box-Cox with spatially calibrated lambda performs well for lead times up to 6–7 days, but is among the poorest for 8–10 days lead times. This is consistent for calibration, verification, and spatiotemporal cross validation, and also for different combinations of calibration and verification years. We have not been able to find the reason for this performance decay with lead time. We

FIG. 7. Histogram of PIT values for the calibration and the validation case for the three methods for lead times (left) 1 and (right) 7 days.

did not analyze how each transformation worked for each catchment, and whereas the transformation could lead to possibilities for unrealistic high runoff or runoff below zero. We could maybe have constricted the transformation parameters further by introducing more limiting assumptions, but this would make the presentation much more complex, and we do not think this would have much impact on the results.

The parameters from the untransformed data had poor results for the calibration case. But the results are better for the verification and the cross-validation cases and among the best ones for the forecasting case. Additionally, there are fewer catchments with poor results when using untransformed data.

This raises the question if a transformation should be based on the distribution of the entire time series of the runoff. Normality is an assumption for postprocessing, but this is related to the prediction errors for a particular prediction, not the time series. One of them could be normal and the other not, although the most common case is probably that both time series and the

error distributions are nonnormal, but with different and maybe varying distributions. Finding a transformation for the time series is easier than finding a transformation for the errors. This might be a reason for the lacking improvement from transforming the data.

Assumptions about transformation parameters could have an impact as well, although we have tested several different restrictions on these. [McInerney et al. \(2017\)](#) suggested that Box-Cox parameters should be in the range 0–0.3. They achieved best results for a fixed value of 0.2 ([Berthet et al. 2020](#)). We used a lower fixed value, as this was what the data indicated, but included one option where lambda had to be larger than 0.

[Hemri et al. \(2013\)](#) based their approach on BMA, using Box-Cox transformed data. However, the results are not completely comparable. First of all, they used the historical data for estimating the lambda for the transformation. In addition, their benchmark in the CRPSS estimates was the long-term climatology, computing the CRPSS both for the raw

FIG. 8. Boxplots of (normalized) calibrated parameter values as function of lead time for 2017. (top) The variance inflation factor (d_{fit}) and the variance bias correction (normalized with 90th percentile of daily standard deviations of the raw ensembles) (c_{fit}) in Eq. (1), respectively. (bottom) The bias correction divided by (left) mean flow and (right) the sum of weights.

ensemble and their BMA model with this benchmark. Their results indicated that the forecast quality declined strongly after 5 days, when COSMO-LEPS is no longer available. Comparing the CRPSS of the raw ensembles and the BMA varieties visually, their results show that a univariate method (looking at different lead times separately) is better than the raw ensemble until 5 days lead time, whereas their multivariate method (mixing the lead times) is better for one more day. However, this was for a small region. Engeland and Steinsland (2014) analyzed their method differently, making a comparison with their results more challenging.

c. Comparison with global calibration methods

We looked at three different methods in this study, the “ordinary” EMOS method, a simplified version and to calibrate

parameters locally. They all have different advantages and disadvantages.

The local calibration gives best results for many catchments, but there are also many cases where the CRPSS values of the postprocessed ensembles will be poorer than the raw ensemble itself. The global calibration of all EMOS parameters gives better scores for the 10th percentile for longer lead times than the local calibration, but also worse scores for the median and the 90th percentile. The VIF calibration rarely gives very bad results, but also gives a limited improvement for the 90th percentile.

d. Indications of when it is not beneficial to postprocess

The postprocessing methods had skill for short lead times, but less for the longer lead times. We analyzed the properties of the ensembles to improve our understanding of this.

FIG. 9. Weights for the different forecasts as a function of lead time for the three forecast years. Median of locally calibrated weights as solid lines and globally calibrated weights as dashed lines.

First, we realize that the PIT values in Fig. 7 give a good indication of when we can improve the variability of the forecasts. A variance transformation as in the EMOS method seems efficient when the PIT values of the raw ensemble have a symmetric distribution, which is concave or convex. The postprocessed distribution will also depend on the bias correction, but if the distribution is trimodal, as for the longer lead times, it seems challenging to transform the variance into a more uniform distribution with the EMOS method we have used. The BMA method (Raftery et al. 2005) could potentially do better, as it can be modeled as a combination of the different distributions of the input ensembles. However, this would not be possible if the same double modality is also present in all the different ensemble forecasts. This is something that could be studied in the future. We have not analyzed the reasons for the shapes of the PIT diagrams.

A second explanation relates to biases and to regional differences in forecast quality for the local calibration method. Both our calibration data and our forecast data are simulated output from a model, which eliminates model biases. However, we would still expect to see some biases due to the different forcing of the hydrological model for calibration and forecasting. The forecasted precipitation might need bias correction at some locations, and the interpolation of rainfall does not have orographic correction. We would therefore expect errors in regions with high precipitation gradients and where the station density is low and maybe not representative, such as in mountainous regions. The very low bias corrections in Fig. 8 indicate that this is not the case. On the other hand, using observed runoff would most likely increase the need to correct (increase) the variance further, unless observation data are assimilated in the forecasts.

Traditionally, the EMOS method has been calibrated with shorter calibration periods, which in theory should be better

able to capture possible changes in the forecasting models. We also tested shorter calibration periods, but the results were similar, it only increased the variability of the results. It did not increase the median quality of the postprocessed forecasts, and did not seem to give a better identification of “the best forecast.” This is in accordance with the study of Lang et al. (2020), which suggested that longer calibration periods are recommended. One difference with their study is that we combine several different forecasts, whereas they only looked at ensembles from ECMWF. Additionally, a hydrological model will be less affected by small differences in the meteorological forecasts during periods of low flow. A longer calibration period (at least a year) could be better and more likely to capture the behavior during some larger events. The results at the end of section 4g could indicate that even longer calibration periods could be beneficial, as there were larger similarities between the parameters from 2015 and 2017 than between the parameters from 2016 and the two other years.

Figures 8 and 9 indicate that we cannot expect large improvements compared with the raw ensemble for longer lead times. The bias corrections (variance and mean) are both close to zero, and the variance inflation factor is close to one. When the distribution of weights also appears to be inconsistent between years, the postprocessing will not be able to improve the forecasts.

The hydrological model can be seen as a space–time filter of the meteorological input (Skøien and Blöschl 2006). Differences in quality between the forecasts can therefore be smoothed reducing the differences between them.

e. Final remarks

We are particularly interested in flooding, and it would have been tempting to restrict the calibration only to observations above a certain threshold. However, Lerch et al. (2017) show

that this will bias the calibration toward extreme forecasts. Unbiased calibration can be achieved by using forecasts above a certain threshold for calibration, but then we would not calibrate with floods that were missed by the forecasters. However, we did a validation only against the extremes for comparison, when the runoff is above a certain percentile in the range from 95 to 99.5. The results were not substantially different from the uncalibrated case in Fig. 5, using all data, but with a much higher uncertainty, most likely a result of the lower number of days used in the validation for each catchment (19, 11, 8, 4, and 2, respectively).

We did not look into hydrologic model uncertainty, as EFAS is currently only run with a single model and a single parameter set, neither did we apply a postprocessor such as the hydrologic uncertainty processor (Krzysztofowicz and Kelly 2000). However, except for biases due to erroneous catchment borders, we assume that hydrological model uncertainties are typically of higher importance for low flows and in some cases for short lead times, whereas meteorological uncertainties and biases are of higher importance for floods and lead times beyond the catchment response time (Benninga et al. 2017; based on Demirel et al. 2013; Olsson and Lindström 2008).

We have used the specific runoff in the analyses in this study. Whereas the long-term average specific runoff is area independent, this is usually not the case for specific flood magnitude (Merz and Blöschl 2005), where smaller catchments tend to have larger specific floods. This is something that could also affect ensemble forecasts of discharge. However, we have not tried to correct this, as the optimum correction will again depend on flood magnitude and location and would further complicate this study.

We have looked at a rather simple method for postprocessing, for several reasons, including computational simplicity, possibility to regionalize and possibility to analyze the calibrated parameters. It might be possible that more complex methods that have been developed during recent years [see, e.g., Vannitsem et al. (2020) for an overview] would have achieved better postprocessing results also for longer lead times.

6. Conclusions

Postprocessing is often regarded as necessary before interpretation of ensemble forecasts (Gneiting et al. 2005). It is also often advised to transform hydrological data before statistical processing, as done by, e.g., Hemri et al. (2015). Here we have compared different postprocessing techniques and transformations, and analyzed the effect of these on forecast runoff data for entire Europe, both for the calibration case and for a set of different verification cases. Whereas postprocessing with the EMOS method always improved the forecasts in the calibration case, and some of the transformations appear to give better results than untransformed data, this is less clear for the forecasting cases. Here we have both looked at forecasts at the calibration locations, and spatiotemporal cross validation, where we have interpolated the postprocessing parameters to uncalibrated locations. Our first objective was to develop and test postprocessing with interpolated parameters. Although

these parameters will lead to slightly poorer postprocessed forecasts than the calibrated parameters, the difference is relatively small and the spatial interpolation of parameters could be useful if there was a stronger need for postprocessing. Our second objective was to analyze the effect of transformations of the input data, and our results indicated that none of the transformations give substantially better results than for untransformed data for the forecast case.

The choice of the best method will depend on the usage. If any improvement is seen as positive, the locally calibrated parameters give the best results for the validation case. However, in an operational service it would also be a target to avoid the risk of deteriorating the ensemble forecasts through the postprocessing. Simplicity is also a factor, as more complex methodologies also increase the risk of errors and bugs in the system, therefore deteriorating the robustness of the service. For the dataset we have studied here, the best solution for an operational service might therefore be to only calibrate the variance inflation factor. In this way, we can see improvements in the postprocessed forecasts for up to 5 days lead time, but at the same time we limit the number of cases where the postprocessing will deteriorate the raw forecasts.

Acknowledgments. The European Flood Awareness System, as part of the Copernicus Emergency Management Service, is financed through the Copernicus program. We thank the editor and three reviewers for their valuable and constructive comments to earlier versions of the manuscript.

Data availability statement. Some of the key data in this study are subject to confidentiality agreements. The supporting data can therefore only be made available to bona fide researchers subject to a nondisclosure agreement. Details of the data and how to request access are available from peter.salamon@ec.europa.eu at European Commissions Joint Research Centre.

REFERENCES

- Alfieri, L., F. Pappenberger, F. Wetterhall, T. Haiden, D. Richardson, and P. Salamon, 2014: Evaluation of ensemble streamflow predictions in Europe. *J. Hydrol.*, **517**, 913–922, <https://doi.org/10.1016/j.jhydrol.2014.06.035>.
- Bartholmes, J. C., J. Thielen, M. H. Ramos, and S. Gentilini, 2009: The European Flood Alert System EFAS – Part 2: Statistical skill assessment of probabilistic and deterministic operational forecasts. *Hydrol. Earth Syst. Sci.*, **13**, 141–153, <https://doi.org/10.5194/hess-13-141-2009>.
- Benninga, H.-J. F., M. J. Booij, R. J. Romanowicz, and T. H. M. Rientjes, 2017: Performance of ensemble streamflow forecasts under varied hydrometeorological conditions. *Hydrol. Earth Syst. Sci.*, **21**, 5273–5291, <https://doi.org/10.5194/hess-21-5273-2017>.
- Berrocal, V. J., A. E. Raftery, and T. Gneiting, 2007: Combining spatial statistical and ensemble information in probabilistic weather forecasts. *Mon. Wea. Rev.*, **135**, 1386–1402, <https://doi.org/10.1175/MWR3341.1>.
- Berthet, L., F. Bourgin, C. Perrin, J. Viatgé, R. Marty, and O. Piotte, 2020: A crash-testing framework for predictive uncertainty assessment when forecasting high flows in an extrapolation context. *Hydrol. Earth Syst. Sci.*, **24**, 2017–2041, <https://doi.org/10.5194/hess-24-2017-2020>.

- Bogner, K., and F. Pappenberger, 2011: Multiscale error analysis, correction, and predictive uncertainty estimation in a flood forecasting system. *Water Resour. Res.*, **47**, W07524, <https://doi.org/10.1029/2010WR009137>.
- , —, and H. L. Cloke, 2012: Technical note: The normal quantile transformation and its application in a flood forecasting system. *Hydrol. Earth Syst. Sci.*, **16**, 1085–1094, <https://doi.org/10.5194/hess-16-1085-2012>.
- Cloke, H. L., and F. Pappenberger, 2009: Ensemble flood forecasting: A review. *J. Hydrol.*, **375**, 613–626, <https://doi.org/10.1016/j.jhydrol.2009.06.005>.
- De Roo, A. P. J., C. G. Wesseling, and W. P. A. Van Deursen, 2000: Physically based river basin modelling within a GIS: The LISFLOOD model. *Hydrol. Processes*, **14**, 1981–1992, [https://doi.org/10.1002/1099-1085\(20000815/30\)14:11/12<1981::AID-HYP49>3.0.CO;2-F](https://doi.org/10.1002/1099-1085(20000815/30)14:11/12<1981::AID-HYP49>3.0.CO;2-F).
- Demirel, M. C., M. J. Booij, and A. Y. Hoekstra, 2013: Effect of different uncertainty sources on the skill of 10 day ensemble low flow forecasts for two hydrological models. *Water Resour. Res.*, **49**, 4035–4053, <https://doi.org/10.1002/wrcr.20294>.
- Duan, Q., S. Sorooshian, and V. K. Gupta, 1992: Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resour. Res.*, **28**, 1015–1031, <https://doi.org/10.1029/91WR02985>.
- DWD, 2017: NWP forecast data. DWS, accessed 14 June 2021, https://www.dwd.de/EN/ourservices/nwp_forecast_data/nwp_forecast_data.html.
- Engeland, K., and I. Steinsland, 2014: Probabilistic postprocessing models for flow forecasts for a system of catchments and several lead times. *Water Resour. Res.*, **50**, 182–197, <https://doi.org/10.1002/2012WR012757>.
- Gilleland, E., and R. W. Katz, 2016: ExtRemes 2.0: An extreme value analysis package in R. *J. Stat. Software*, **72**, 1–39, <https://doi.org/10.18637/jss.v072.i08>.
- Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, <https://doi.org/10.1175/MWR2904.1>.
- Hamill, T. M., R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Wea. Rev.*, **136**, 2620–2632, <https://doi.org/10.1175/2007MWR2411.1>.
- Hemri, S., 2014: Trends in the predictive performance of raw ensemble weather forecasts. *Geophys. Res. Lett.*, **41**, 9197–9205, <https://doi.org/10.1002/2014GL062472>.
- , 2018: Applications of postprocessing for hydrological forecasts. *Statistical Postprocessing of Ensemble Forecasts*, S. Vannitsem, D. S. Wilks, and J. W. Messner, Eds., Elsevier, 219–240, <https://doi.org/10.1016/B978-0-12-812372-0.00008-X>.
- , F. Fundel, and M. Zappa, 2013: Simultaneous calibration of ensemble river flow predictions over an entire range of lead times. *Water Resour. Res.*, **49**, 6744–6755, <https://doi.org/10.1002/wrcr.20542>.
- , D. Lisniak, and B. Klein, 2015: Multivariate post-processing techniques for probabilistic hydrological forecasting. *Water Resour. Res.*, **51**, 7436–7451, <https://doi.org/10.1002/2014WR016473>.
- Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570, [https://doi.org/10.1175/1520-0434\(2000\)015<0559:DOTCRP>2.0.CO;2](https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2).
- Holm, E., R. Forbes, S. Lang, L. Magnusson, and S. Malardel, 2016: New model cycle brings higher resolution. *ECMWF Newsletter*, No. 147, ECMWF, Reading, United Kingdom, 14–19, <https://www.ecmwf.int/en/newsletter/147/meteorology/new-model-cycle-brings-higher-resolution>.
- Jordan, A., F. Krüger, and S. Lerch, 2019: Evaluating probabilistic forecasts with scoring rules. *J. Stat. Software*, **90**, 1–37, <https://doi.org/10.18637/jss.v090.i12>.
- Karssen, D., O. Schmitz, P. Salamon, K. de Jong, and M. F. P. Bierkens, 2010: A software framework for construction of process-based stochastic spatio-temporal models and data assimilation. *Environ. Modell. Software*, **25**, 489–502, <https://doi.org/10.1016/j.envsoft.2009.10.004>.
- Krzysztofowicz, R., 1997: Transformation and normalization of variates with specified distributions. *J. Hydrol.*, **197**, 286–292, [https://doi.org/10.1016/S0022-1694\(96\)03276-3](https://doi.org/10.1016/S0022-1694(96)03276-3).
- , and K. S. Kelly, 2000: Hydrologic uncertainty processor for probabilistic river stage forecasting. *Water Resour. Res.*, **36**, 3265–3277, <https://doi.org/10.1029/2000WR900108>.
- Laio, F., and S. Tamea, 2007: Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrol. Earth Syst. Sci.*, **11**, 1267–1277, <https://doi.org/10.5194/hess-11-1267-2007>.
- Lang, M. N., S. Lerch, G. J. Mayr, T. Simon, R. Stauffer, and A. Zeileis, 2020: Remember the past: A comparison of time-adaptive training schemes for non-homogeneous regression. *Nonlinear Processes Geophys.*, **27**, 23–34, <https://doi.org/10.5194/npg-27-23-2020>.
- Lerch, S., and S. Baran, 2017: Similarity-based semilocal estimation of post-processing models. *J. Roy. Stat. Soc.*, **66**, 29–51, <https://doi.org/10.1111/rssc.12153>.
- , T. L. Thorarindottir, F. Ravazzolo, and T. Gneiting, 2017: Forecaster's dilemma: Extreme events and forecast evaluation. *Stat. Sci.*, **32**, 106–127, <https://doi.org/10.1214/16-STS588>.
- Marsigli, C., F. Boccanera, A. Montani, and T. Paccagnella, 2005: The COSMO-LEPS mesoscale ensemble system: Validation of the methodology and verification. *Nonlinear Processes Geophys.*, **12**, 527–536, <https://doi.org/10.5194/npg-12-527-2005>.
- McInerney, D., M. Thyer, D. Kavetski, J. Lerat, and G. Kuczera, 2017: Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors. *Water Resour. Res.*, **53**, 2199–2239, <https://doi.org/10.1002/2016WR019168>.
- Merz, R., and G. Blöschl, 2005: Flood frequency regionalisation – Spatial proximity vs. catchment attributes. *J. Hydrol.*, **302**, 283–306, <https://doi.org/10.1016/j.jhydrol.2004.07.018>.
- Nelder, J. A., and R. Mead, 1965: A simplex method for function minimization. *Comput. J.*, **7**, 308–313, <https://doi.org/10.1093/comjnl/7.4.308>.
- Olsson, J., and G. Lindström, 2008: Evaluation and calibration of operational hydrological ensemble forecasts in Sweden. *J. Hydrol.*, **350**, 14–24, <https://doi.org/10.1016/j.jhydrol.2007.11.010>.
- Pappenberger, F., M. H. Ramos, H. L. Cloke, F. Wetterhall, L. Alfieri, K. Bogner, A. Mueller, and P. Salamon, 2015: How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. *J. Hydrol.*, **522**, 697–713, <https://doi.org/10.1016/j.jhydrol.2015.01.024>.
- R Core Team, 2019: R: A language and environment for statistical computing. R Foundation for Statistical Computing, <http://www.r-project.org>.
- Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, <https://doi.org/10.1175/MWR2906.1>.

- Scheuerer, M., and L. Büermann, 2014: Spatially adaptive post-processing of ensemble forecasts for temperature. *J. Roy. Stat. Soc.*, **63**, 405–422, <https://doi.org/10.1111/rssc.12040>.
- Skøien, J. O., and G. Blöschl, 2006: Catchments as space-time filters – A joint spatio-temporal geostatistical analysis of runoff and precipitation. *Hydrol. Earth Syst. Sci.*, **10**, 645–662, <https://doi.org/10.5194/hess-10-645-2006>.
- , R. Merz, and G. Blöschl, 2006: Top-kriging - Geostatistics on stream networks. *Hydrol. Earth Syst. Sci.*, **10**, 277–287, <https://doi.org/10.5194/hess-10-277-2006>.
- , G. Blöschl, G. Laaha, E. Pebesma, J. Parajka, and A. Viglione, 2014: Rtop: An R package for interpolation of data with a variable spatial support, with an example from river networks. *Comput. Geosci.*, **67**, 180–190, <https://doi.org/10.1016/j.cageo.2014.02.009>.
- Smith, P. J., and Coauthors, 2016: On the operational implementation of the European Flood Awareness System (EFAS). *Flood Forecasting*, T. E. Adams III and T. C. Pagano, Eds., Academic Press, 313–348, <https://doi.org/10.1016/B978-0-12-801884-2.00011-6>.
- Thielen, J., J. Bartholmes, M.-H. Ramos, and A. de Roo, 2009: The European Flood Alert System – Part 1: Concept and development. *Hydrol. Earth Syst. Sci.*, **13**, 125–140, <https://doi.org/10.5194/hess-13-125-2009>.
- Van Der Knijff, J. M., J. Younis, and A. P. J. De Roo, 2010: LISFLOOD: A GIS-based distributed model for river Basin scale water balance and flood simulation. *Int. J. Geogr. Inf. Sci.*, **24**, 189–212, <https://doi.org/10.1080/13658810802549154>.
- Vannitsem, S., and Coauthors, 2020: Statistical postprocessing for weather forecasts – Review, challenges and avenues in a big data world. *Bull. Amer. Meteor. Soc.*, **102**, E681–E699, <https://doi.org/10.1175/BAMS-D-19-0308.1>.
- Wetterhall, F., and Coauthors, 2013: HESS opinions “Forecaster priorities for improving probabilistic flood forecasts.” *Hydrol. Earth Syst. Sci.*, **17**, 4389–4399, <https://doi.org/10.5194/hess-17-4389-2013>.
- Wilks, D. S., 2017: On assessing calibration of multivariate ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **143**, 164–172, <https://doi.org/10.1002/qj.2906>.